Counting and sampling gene family evolutionary histories in the duplication-loss and duplication-loss-transfer models

Cedric Chauve, Yann Ponty, Michael Wallner

Noname manuscript No.
(will be inserted by the editor)

Counting and sampling gene family evolutionary histories in the duplication-loss and duplication-loss-transfer models

Cedric Chauve · Yann Ponty · Michael Wallner

Abstract Given a set of species whose evolution is represented by a species tree, a gene family is a group of genes having evolved from a single ancestral gene. A gene family evolves along the branches of a species tree through various mechanisms, including – but not limited to – speciation ($\mathbb{S}$), gene duplication ($\mathbb{D}$), gene loss ($\mathbb{L}$), horizontal gene transfer ($\mathbb{T}$). The reconstruction of a gene tree representing the evolution of a gene family constrained by a species tree is an important problem in phylogenomics. However, unlike in the multispecies coalescent evolutionary model that considers only speciation and incomplete lineage sorting events, very little is known about the search space for gene family histories accounting for gene duplication, gene loss and horizontal gene transfer (the $\mathbb{DLT}$-model).

In this work, we introduce the notion of evolutionary histories defined as a binary ordered rooted tree describing the evolution of a gene family, constrained by a species tree in the $\mathbb{DLT}$-model. We provide formal grammars describing the set of all evolutionary histories that are compatible with a given species tree, whether it is ranked or unranked. These grammars allow us, using either analytic combinatorics or dynamic programming, to efficiently compute the number of histories of a given size, and also to generate random histories of a given size under the uniform distribution. We apply these tools to obtain exact asymptotics for the number of gene family histories for two species trees, the rooted caterpillar and the complete binary tree, as well as estimates of the range of the exponential growth factor of the number of histories for random species trees of size up to 25. Our results show that including horizontal gene transfer induce

Cedric Chauve
Department of Mathematics, Simon Fraser University, Burnaby (BC), Canada
LaBRI, Université de Bordeaux, Talence, France
LIX, Ecole Polytechnique, Palaiseau, France
0000-0001-9837-1878

Yann Ponty
CNRS and LIX, Ecole Polytechnique, Palaiseau, France
0000-0002-7615-3930

Michael Wallner
LaBRI, Université de Bordeaux, Talence, France
Institut für Diskrete Mathematik und Geometrie, TU Wien, Vienna, Austria
0000-0001-8581-449X

a dramatic increase of the number of evolutionary histories. We also show that, within ranked species trees, the number of evolutionary histories in the $\mathbb{DLT}$-model is almost independent of the species tree topology. These results establish firm foundations for the development of ensemble methods for the prediction of reconciliations.

Keywords Phylogenetics, Enumerative Combinatorics, Asymptotics, Sampling Algorithms
Mathematics Subject Classification (2010) 92B99,05A15,05A16

1 Introduction

A gene tree represents the evolution of a gene family, a group of genes assumed to descend from a single ancestral gene. The reconstruction of gene trees from molecular sequence data is a central but difficult problem in computational biology. Indeed, while species are mostly expected to evolve through *speciation*, gene families evolve through a wider variety of mechanisms including gene duplication, gene loss, horizontal gene transfer (HGT) and incomplete lineage sorting (ILS). As a result, it is common to observe an incongruence between gene trees and species trees [32]. This discrepancy has motivated an intense research activity on the problem of reconstructing the gene tree of a gene family, conditional to a given species tree for the considered species. We refer to [43, 45] for extensive reviews discussing how gene trees evolve within a species tree, describe existing models and methods for reconstructing gene trees within species trees.

In the case where a gene family contains a single gene per species, observed incongruences between a gene tree and a species tree can be analyzed through the prism of ILS in the *multispecies coalescent model* [11]. The natural question is then to compute the probability of *coalescent histories* conditional to the given species tree [12,35,49,50]. For gene families that might contain duplicate copies (or no copy) of a gene in a given species, the multispecies coalescent model is not appropriate, and gene trees need to be inferred in a model including gene duplication, gene loss and, ideally, transfers. Most methods developed to understand the evolution of gene families in this context rely on the concept of *gene tree-species tree reconciliation*, illustrated in Fig. 1. In this framework, given a gene tree G and a species tree S, one aims to embed G within S, often optimizing a parsimony or probabilistic criterion with regard to the considered evolutionary model.

Early reconciliation methods were developed for an evolutionary model considering only gene duplications and gene losses (the $\mathbb{DL}$-model), and considered a parsimony criterion. This problem, introduced by Goodman *et al.* [26], is computationally tractable through dynamic programming. Extending the model to include HGT, while ensuring that HGT events are time-consistent, makes the problem of predicting of the most parsimonious reconciliation intractable in general [34, 47]. However, if the provided species tree is *ranked*, i.e. is provided with a total ordering of its internal nodes describing the order of speciation events, the reconciliation problem becomes tractable (see the discussion in [19]). Over the last 20 years, various efficient dynamic programming algorithms were designed to compute a parsimonious reconciliation, implemented in widely used phylogenomics packages [6, 22, 31, 39]. Similar to parsimony-based methods, probabilistic reconciliation methods were first developed in a model considering only gene duplication and gene loss [1, 2, 27, 29], before being extended to include HGTs [40, 44].

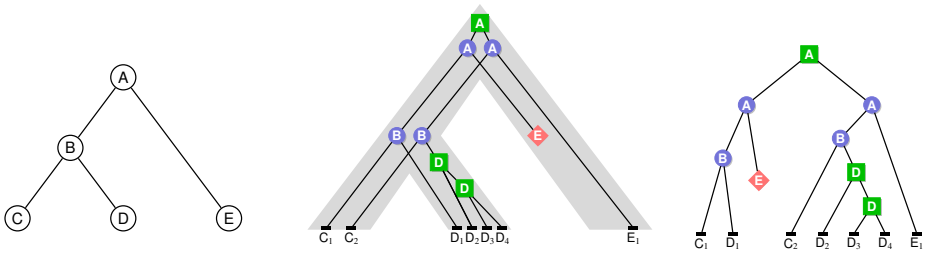

Fig. 1: A species tree $\mathbf{S}$ (left), a $\mathbb{DL}$-history for $\mathbf{S}$ (center) and its associated gene tree (right). Green squares (resp. blue circles, red diamonds, black rectangles) correspond to nodes x such that $e(x) = \mathbb{D}$ (resp. $e(x) = \mathbb{S}, e(x) = \mathbb{L}, e(x) = Extant$). The mapping s is represented by the location of the internal nodes of the history within the species tree in the center tree and by the species names in the nodes in the right tree.

Most methods that reconstruct a gene tree, conditional to a species tree, rely on the exploration of the space of possible evolutionary histories. It is then important to develop conceptual tools that can describe this combinatorial space and further enable its efficient exploration. This naturally raises the questions to compute the size of the space of evolutionary histories for a given gene family and a given species tree, and to be able to sample such histories. Both questions are naturally related, as precise counting results often translate into efficient sampling algorithms [24, 48]. The former (counting) question has been studied by Rosenberg *et al.* in the case of the multi-species coalescent model [13–17, 38]. However similar questions have not been explored as thoroughly for evolutionary models including gene duplication, gene loss and HGT. In this framework, dynamic programming equations aimed at computing a parsimonious reconciled gene tree can be turned into a specification of the corresponding search space [28, 36]. This then leads to efficient algorithms for counting or sampling parsimonious reconciliations [5, 18] or sampling reconciled gene trees under the Boltzmann probability distribution [31]. However, to the best of our knowledge, such questions have not been considered in the case where a gene tree is not specified at first, i.e. we are only given a species tree and gene family.

This paper provides analytic and algorithmic answers to those questions. We show that, for a given species tree, whether ranked or unranked, the space of all possible evolutionary histories of a fixed size in the $\mathbb{DLT}$-model can be described using a formal grammar. This allows us to compute, in polynomial time and space, for given species tree and gene family size, the number of evolutionary histories of this size conditional to the given species tree, as well as to sample among these histories under the uniform probability. Using these algorithms, we can provide estimates of the exponential growth factor of the number of histories in the $\mathbb{DL}$-model and $\mathbb{DLT}$-model. We show that, as expected, including HGT in a model results in an exponential increase of the number of histories. We also notice that with a ranked species tree, the exponential growth factor of the number of histories in the $\mathbb{DLT}$-model seems to be almost independent of the chosen species tree. Finally, using enumerative and analytic combinatorics, we provide exact values for the asymptotic number of histories for two specific species tree: the rooted caterpillar tree and the rooted complete binary tree.

2 Model: gene families evolutionary histories

In this section, we introduce the combinatorial objects modeling the evolution of a gene family within a given species tree, that we call *histories*.

Preliminaries on trees. For a given rooted tree[1] $\mathbf{T}$, we say it is *uniquely labeled* if every node has a label, and no two nodes have the same label. For a node x in $\mathbf{T}$, we denote by $\mathbf{T}_x$ the subtree of $\mathbf{T}$ rooted at x. In this work, we consider only *binary* and *unary-binary trees*: in a binary tree, every internal node has exactly two children, while in a unary-binary tree, an internal node can have either one child or two children. If a uniquely labeled tree $\mathbf{T}$ is unordered we take advantage of the nodes labeling to see it as an ordered tree, with the two children of an internal node x being ordered from left to right in increasing order of their labels; so from now on all trees we consider are ordered. If an internal node x of a tree $\mathbf{T}$ is binary, we denote by x_ℓ the left child of x and by x_r its right child; if x is unary, i.e. has a single child, we denote it by x_c. We denote by $r(\mathbf{T})$ the root of $\mathbf{T}$. For a node x of $\mathbf{T}$, we denote by $p(x)$ its parent in $\mathbf{T}$. The *size* of a tree $\mathbf{T}$ is the number of its leaves.

A rooted tree describes a partial order on the set of its nodes, and two nodes are said to be *comparable* if one is an ancestor of the other one and *incomparable* otherwise. For a node u, we denote by $\overline{C}(u)$ the set of nodes that are incomparable with u.

Ranked trees. A *ranking* of a tree T of size n is a mapping π from the nodes of T to $\{1, \ldots, n\}$ such that (1) $\pi(x) = n$ if x is a leaf, (2) $\pi(x) \neq \pi(y)$ if x and y are internal nodes, and (3) $\pi(x) < \pi(y)$ if x is an ancestor of y. A tree augmented with a ranking is called a *ranked tree*; in our context it models the evolution of a set of species, the ranking providing the relative order of speciation events, under the assumption that no two speciations can occur at the same time.

Given a binary tree $\mathbf{T}$ and a ranking π, we define an unranked unary-binary tree $\mathbf{T}_\pi$ that encodes the ranking information as follows: for each internal node u, considered iteratively in increasing ranking order, and for every edge $(p(v), v)$ such that $\pi(p(v)) < \pi(u) < \pi(v)$, we subdivide the edge $(p(v), v)$ into two edges $(p(v), v_u)$ and (v_u, v), so adding a unary node v_u on this edge. We denote by $t(u)$ the set of all unary nodes created in this way and we call this set of nodes together with u a *time slice*. Additionally, we also define the set of all leaves as a time slice (see Figure 2). Note that in this way we create n different time slices which correspond to the n different values of the ranking. We modify the notion of incomparability for such unary-binary trees as follows: for a node u, $\overline{C}(u) = t(u) \setminus \{u\}$.

Gene Families Evolutionary Histories. The objects we study in this work model the evolution of a gene family within a species tree. A species tree, which will be denoted by $\mathbf{S}$ from now on, is a uniquely labeled rooted binary tree that represents the evolution of a set of species through speciation events; $\mathbf{S}$ can be either unranked or ranked. A gene family evolves within $\mathbf{S}$ from a single ancestral gene, present in the species $r(\mathbf{S})$, through four possible kinds of *evolutionary events*:

– *Speciation* $\mathbb{S}$: a gene x present in species u has two descendant genes x_ℓ present in species u_ℓ and x_r present in species u_r.

[1] In the present work we consider only rooted trees.

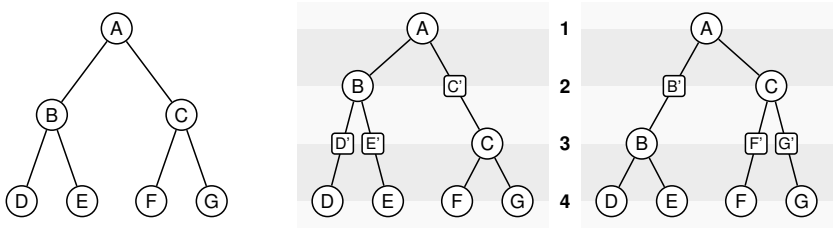

Fig. 2: An example of a ranked tree with time slices. (Left) The complete binary tree $\mathbf{T}$ of size 4. (Center) The unary-binary tree $\mathbf{T}_\pi$ for the ranking π defined by $\pi(\mathsf{A}) = 1$, $\pi(\mathsf{B}) = 2$, $\pi(\mathsf{C}) = 3$ and $\pi(\mathsf{D}) = \pi(\mathsf{E}) = \pi(\mathsf{F}) = \pi(\mathsf{G}) = 4$; the time slices in $\mathbf{T}_\pi$ are the following sets of nodes: $\{\mathsf{A}\}$, $\{\mathsf{B}, \mathsf{C}'\}$, $\{\mathsf{C}, \mathsf{D}', \mathsf{E}'\}$,$\{\mathsf{D}, \mathsf{E}, \mathsf{F}, \mathsf{G}\}$; (Right) Alternative unary-binary tree $\mathbf{T}_{\pi'}$, induced by exchanging the rankings of B and C.

- *Duplication* $\mathbb{D}$: a gene x present in species u is duplicated, with a new copy x_d of x appearing in species u; x is said to be the *original gene* while x_d is the *novel gene*.
- *Loss* $\mathbb{L}$: a gene x present in species u has exactly one descendant either in x_ℓ or in x_r, implying that after a speciation at species u, exactly one of the two resulting genes is lost along the branch toward either u_ℓ or u_r.
- *Horizontal Gene Transfer* $\mathbb{T}$ *(HGT)*: this is similar to a duplication but the novel copy, denoted x_t here, appears in a species v different from u and incomparable with u, called the *receiver* of the HGT, while u is called the *donor* of the HGT. If $\mathbf{S}$ is ranked, with ranking π, the receiver species v is required to exist at the same time as u, i.e. to satisfy two ranking constraints, $\pi(p(v)) < \pi(u) < \pi(v)$.

Definition 2.1 An evolutionary history for a gene family within a species tree $\mathbf{S}$ is a unary-binary ordered rooted tree $\mathbf{T}$ together with two mappings $s : V(\mathbf{T}) \to V(\mathbf{S})$ and $e : V(\mathbf{T}) \to \{\mathbb{S}, \mathbb{D}, \mathbb{L}, \mathbb{T}, \mathit{Extant}\}$ satisfying the following constraints:

- if x is a leaf, $e(x) \in \{\mathit{Extant}, \mathbb{L}\}$;
- if x is internal and binary, $e(x) \in \{\mathbb{S}, \mathbb{D}, \mathbb{T}\}$;
- if x is internal and unary then $e(x) = \mathbb{S}$[2];
- if $e(x) = \mathbb{S}$ and $s(x) = u$ is binary then $s(x_\ell) = u_\ell$ and $s(x_r) = u_r$;
- if $e(x) = \mathbb{S}$ and $s(x) = u$ is unary then $s(x_c) = u_c$;
- if $e(x) = \mathbb{D}$ then $s(x_\ell) = s(x_r) = s(x)$;
- if $e(x) = \mathbb{T}$ then $s(x_\ell) = s(x)$ and $s(x_r) \in \overline{C}(s(x))$.

The size of a history is the number of leaves x such that $e(x) = \mathit{Extant}$.

Intuitively, this definition states that a history is represented by a tree where each node corresponds to a gene present in a species, either extant or ancestral (the mapping s), and each ancestral gene either was lost ($e(x) = \mathbb{L}$) or evolved toward extant genes through a duplication ($e(x) = \mathbb{D}$), an HGT to an incomparable receiver species ($e(x) = \mathbb{T}$) or a speciation ($e(x) = \mathbb{S}$), while extant genes belong to extant species; the constraints on the species mapping s ensure that this history can be embedded within $\mathbf{S}$ as illustrated in Figure 1.

[2] Note that technically the event associated to a unary node in the species tree is not speciation in the biological meaning, but we chose to label it as such for expository reasons.

By convention, for duplications, we consider that the novel copy of a gene x is its right child x_r, x_ℓ representing the original copy. Histories considered by the $\mathbb{DL}$-model, which allows both duplications and losses (resp. duplications, losses and HGTs), are called $\mathbb{DL}$-histories (resp. $\mathbb{DLT}$-histories).

Remark 2.1 By modeling the evolution of a gene family with ordered trees we differ from the classical notion of *reconciliation*, that also models the evolution of a gene family but considers that when a gene duplication occurs, the original gene and the novel gene are indistinguishable. As a result, the children of a duplication are ordered within a history, whereas they are not in a reconciliation.

Remark 2.2 Gene losses are modeled as speciation events with one disappearing gene. As a consequence, we can not have a duplication or a HGT that results in one of the resulting two gene copies being lost. This is necessary to avoid creating an infinite number of histories of a given size, due to an arbitrary number of duplications within a species, each followed by a loss, or an arbitrary long sequence of HGT, again each followed by a loss, leading to at most one extant gene.

Time Consistency of $\mathbb{DLT}$-histories. Given an unranked species tree $\mathbf{S}$, a $\mathbb{DLT}$-history as defined above is *time inconsistent* if there exists a gene x belonging to a species u such that one of its ancestors belongs to a species v and one of its descendants belongs to a species v' ancestral to v. This pattern can be observed due to the fact that, in the definition of a $\mathbb{DLT}$-history, the choice of the receiver species v of an HGT of gene x belonging to species u is not restricted to the set of species that are also incomparable with all species containing genes that are ancestral to x; see Figure 3 for an illustration.

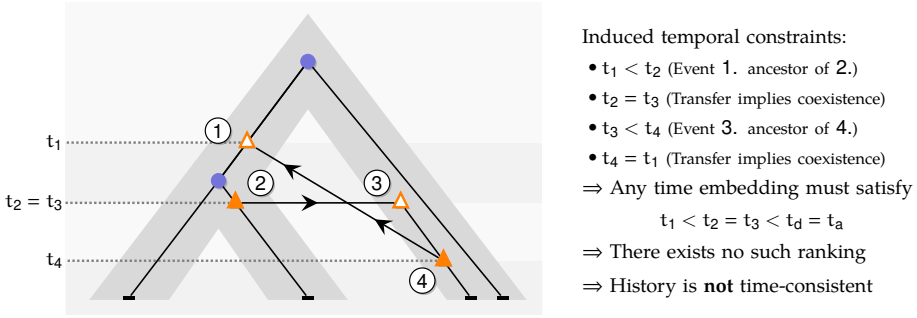

Fig. 3: An example of time-inconsistent $\mathbb{DLT}$-history

The problem of computing gene family evolutionary scenarios that are both parsimonious and time-consistent has been shown to be intractable when such scenarios are modeled by reconciliations with an unranked species tree [34,47], while, when the provided species tree $\mathbf{S}$ is ranked, the problem becomes tractable (see [19] and references therein). Similarly, when $\mathbf{S}$ is ranked, we can ensure time-consistency of evolutionary histories, by requiring that the donor and receiver of any HGT belong to the same time slice in $\mathbf{S}_\pi$, *i.e.* the receiver of an HGT of a gene belonging to a species u belongs to $\overline{C}(u) = t(u) - \{u\}$.

3 Methods

Our results (counting and sampling algorithms) are based on the design of formal grammars specifying, for a given species tree $\mathbf{S}$, the combinatorial families of $\mathbb{DL}$-histories and $\mathbb{DLT}$-histories constrained by $\mathbf{S}$. These grammars are then used as templates to design dynamic programming algorithms for counting and sampling (under the uniform distribution) the number of histories of a fixed size. Moreover, these grammars are amenable to techniques of analytic combinatorics that allow us to compute the asymptotic growth constant for the number of histories. We first describe our grammars, then the counting and sampling algorithms, and finally the asymptotic analysis of these grammars.

3.1 General grammars specifying $\mathbb{DL}$-histories and $\mathbb{DLT}$-histories

In this section we describe grammars specifying histories evolving within a species tree using the formalism developed in [23]. We describe grammars for $\mathbb{DLT}$-histories, for both an unranked and a ranked species tree; these grammars can then be specialized into grammars for $\mathbb{DL}$-histories by omitting the rules related to HGT.

Let $\mathbf{S}$ be a species tree. If $\mathbf{S}$ is unranked, it is a binary tree, otherwise, if it comes with a ranking π, we consider the unary-binary species tree $\mathbf{S}_\pi$. So in the statements below, when mentioning a ranked species tree we mean the unary-binary tree $\mathbf{S}_\pi$ defined by the ranking.

We denote by H_u the set of $\mathbb{DLT}$-histories for the tree $\mathbf{S}_u$. In the most general setting, following [23], these grammars contain both terminal symbols, corresponding to atomic elements of the histories (nodes) and non-terminal symbols, corresponding to combinatorial operators applied to sets of histories. We use the non-terminal $\mathcal{Z}_u$ to encode a gene present in extant species u; moreover, we use $\mathcal{X}_u$ for a gene lost at species u, $\mathcal{Y}_u$ for a duplication at species u and $\mathcal{W}_u$ for a HGT with donor species u. We consider two combinatorial operators, $\cup$ the disjoint union and $\times$ the Cartesian product.

Theorem 3.1 *The set $H_{r(\mathbf{S})}$ defined by the grammar below specifies the set of all $\mathbb{DLT}$-histories for a species tree $\mathbf{S}$.*

$$H_u = \mathcal{S}_u \cup D_u \cup T_u \qquad \text{if } u \text{ is internal} \qquad (1)$$

$$H_u = \mathcal{Z}_u \cup D_u \cup T_u \qquad \text{if } u \text{ is a leaf} \qquad (2)$$

$$\mathcal{S}_u = H_{u_\ell} \times H_{u_r} \cup H_{u_\ell} \times \mathcal{X}_{u_r} \cup \mathcal{X}_{u_\ell} \times H_{u_r} \qquad \text{if } u \text{ is internal and binary} \qquad (3)$$

$$\mathcal{S}_u = H_{u_c} \qquad \text{if } u \text{ is internal and unary} \qquad (4)$$

$$D_u = H_u \times H_u \times \mathcal{Y}_u \qquad (5)$$

$$T_u = \bigcup_{v \in \overline{C}(u)} H_u \times H_v \times \mathcal{W}_u \qquad (6)$$

where $\overline{C}(u)$ is the set of nodes that are incomparable with u in $\mathbf{S}$. The set of $\mathbb{DL}$-histories is specified by the same grammar where rule (6) is removed and the terms T_u are removed from rules (1) and (2).

Proof The grammar follows the definition of histories, Definition 2.1. Rule (1) simply states that the root (*i.e.* the first evolutionary event of the history) of a $\mathbb{DLT}$-history

within the subtree $\mathbf{S}_u$, assuming it is not reduced to a leaf, is either a speciation, a duplication or a transfer of the ancestral gene present in species u: non-terminal $\mathcal{S}_u$, D_u and T_u represent respectively these three subsets of H_u. Rule (2) addresses the case where $\mathbf{S}_u$ is composed of a single leaf, in which case there can not be a speciation event, but a history reduced to a single gene in species u.

Rule (3) describes a speciation event at species u. The ancestral gene can either evolve into a gene in each of the two children of u (first term of the union) or into a gene in a single child of u due to a gene loss in the other child of u. In the case where u is unary (due to being a node created by the time slicing in a ranked $\mathbf{S}$), the ancestral gene evolves into a copy in the unique child u_c of u.

Rule (5) addresses the case of a duplication. It results in two ordered independent histories starting at species u: the first one being the history of the original copy of the starting ancestral gene and the second one the history rooted at the novel gene created by the duplication.

Last, Rule (6) addresses the case of histories starting by a HGT. Generally, a HGT has a structure similar to a duplication but for the fact that the novel gene appears in a species that is incomparable with u.

These various rules cover all cases for describing the possible first event of a history and are mutually exclusive, thus providing a complete recursive specification of $\mathbb{DLT}$-histories for a given species tree $\mathbf{S}$. It follows immediately that removing the rule and non-terminals associated to HGT gives a grammar specifying $\mathbb{DL}$-histories for $\mathbf{S}$. □

Remark 3.1 The above grammar can be greatly simplified if one is interested only in the number of histories of a given size, as opposed to the specific species where gene duplication, gene loss and HGT events occur and the precise gene content of extant species. In this case, one simply identifies all non-terminals $\mathcal{Z}_u$ (resp. $\mathcal{X}_u$, $\mathcal{Y}_u$, $\mathcal{W}_u$) to a single variable $\mathcal{Z}$ (resp. $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{W}$). From now, we follow this approach.

3.2 Counting and sampling algorithms

The grammar defined above can naturally be turned into a dynamic programming algorithm computing the number of histories of a given size. This algorithm computes tables H, D, S, T where, for a given node u of $\mathbf{S}$ and a given history size n, $H[u,n]$ (respectively, $D[u,n]$, $S[u,n]$, $T[u,n]$) is the number of $\mathbb{DLT}$-histories of size n evolving within $\mathbf{S}_u$ (respectively, starting with a duplication, a speciation, and an HGT). We illustrate this in the case of $\mathbb{DLT}$-histories with an unranked species tree $\mathbf{S}$.

$$H[u,n] = S[u,n] + D[u,n] + T[u,n] \qquad \text{if } u \text{ is internal} \quad (7)$$

$$H[u,n] = \mathbb{1}_{n=1} + D[u,n] + T[u,n] \qquad \text{if } u \text{ is a leaf} \quad (8)$$

$$S[u,n] = \sum_{m=1}^{n-1} \left(H[u_\ell, m]H[u_r, n-m]\right) + H[u_\ell, n] + H[u_r, n] \qquad \text{if } u \text{ is internal} \quad (9)$$

$$D[u,n] = \sum_{m=1}^{n-1} \left(H[u,m]H[u,n-m]\right) \quad (10)$$

$$T[u,n] = \sum_{m=1}^{n-1} \left(\sum_{v \in \overline{C}(u)} H[u,m]H[v,n-m] \right) \quad (11)$$

Counting Time $\Phi(n,k)$	$\mathbb{DL}$	$\mathbb{DLT}$
Unranked	$k\,n^2$	$k^2\,n^2$
Ranked	$k^2\,n^2$	$k^3\,n^2$

Counting Space $\Psi(n,k)$	$\mathbb{DL}$	$\mathbb{DLT}$
Unranked	$k\,n^2$	$k^2\,n^2$
Ranked	$k^2\,n^2$	$k^3\,n^2$

Generation Time $\Psi(n,k)$	$\mathbb{DL}$	$\mathbb{DLT}$
Unranked	$n\log n$	$k\,n\log n$
Ranked	$n\log n$	$k\,n\log n$

Table 1: Leading terms for the time ($\Phi(n,k)$) and space ($\Phi(n,k)$) complexities incurred by the evaluation of the counting recurrences for histories consisting of n genes in a species tree of size k.

A random generation algorithm can then be adapted from the counting recurrences, resulting in an instance of the so-called recursive method [48]. Right-hand sides of the counting equation are split into sums of multiplicative terms. Starting from the initial state $H[r(\mathbf{S}),n]$, the algorithm randomly chooses a term from the right-hand side of the current state, with probability proportional to its contribution to the counting. When the selected term is a multiplication of two terms, the length n needs to be distributed across the two terms, and a pair of lengths $(m,n-m)$, is chosen with probability proportional to the associated count. For the sake of performances, the various alternatives can be explored in Boustrophedon order, ensuring an overall $\mathcal{O}(n\log(n))$ worst-case complexity [24]. Recursive calls are then performed over the states associated with the chosen term, until a leaf is chosen (term $\mathbb{1}$). This leads to the following result.

Theorem 3.2 *The number of histories of size n constrained by a species tree of size k can be computed in polynomial time $\mathcal{O}(\Phi(n,k))$ and space $\mathcal{O}(\Psi(n,k))$, where $\Phi(n,k)$ and $\Psi(n,k)$ both depend on the model ($\mathbb{DL}$ or $\mathbb{DLT}$) and the ranked/unranked nature of the species tree, as summarized in Table 1.*

The uniform random generation of h histories of size n can be performed in time $\mathcal{O}(\Phi(n,k)+h\cdot\Upsilon(n,k))$.

3.3 Asymptotic number of histories in the $\mathbb{DL}$-model

The grammar given in Theorem 3.1 defines a combinatorial specification of the set of histories for a given species tree in a given evolutionary model. In this section, we derive the asymptotic number of histories in the $\mathbb{DL}$-model and use it later on two specific species trees: the caterpillar and complete binary trees. The following theorem is the main result of this section and describes their asymptotic growth for n tending to infinity.

Theorem 3.3 *For any given species tree $\mathbf{S}$, the number of histories in the unranked $\mathbb{DL}$-model given by Equations (1)-(5) is, for large n, equal to*

$$\gamma_{\mathbf{S}}\frac{\rho_{\mathbf{S}}^{-n}}{n^{3/2}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right), \tag{12}$$

for explicitly computable constants $\gamma_{\mathbf{S}}>0$ and $\rho_{\mathbf{S}}\in(0,1/4]$.

In the remainder of this section we prove this theorem. The grammars are amenable to enumerative and analytic combinatorics techniques. We follow the general approach presented in Flajolet and Sedgewick [23] and Drmota [20]. It consists mainly in translating the combinatorial specification of a combinatorial family into equations defining its counting generating function. Then, its analytic properties lead to precise asymptotic formulas for its coefficients. We provide an overview of this approach in Example 3.1.

Example 3.1 Consider the class of rooted binary trees B. Such a tree is either a leaf, or it consists of a root with two children which are also each roots of binary trees. Let us mark each leaf with the variable $\mathcal{Z}$. Then, the grammar is given by

$$B = \mathcal{Z} \cup B^2.$$

Let b_n be the number of binary trees with n leaves and let $B(z) = \sum_{n\geq 1} b_n z^n$ be the counting generating function of binary trees. The symbolic method [23, Part A] translates this grammar directly into an equation for the generating function:

$$B(z) = z + B(z)^2. \tag{13}$$

Its generating function is thus given by $B(z) = \frac{1-\sqrt{1-4z}}{2}$.

The general method of singularity analysis from analytic combinatorics [23, Chapter VI] allows us to directly get the asymptotics of the coefficients. First, by the Cauchy–Hadamard theorem, the asymptotic growth is directly connected with the dominant singularities (and the radius of convergence) of the counting generating function. Here, the generating function $B(z)$ becomes singular at $z = 1/4$, which is also the unique singular point. Hence, the coefficients b_n grow like 4^n. Second, using transfer theorems of analytic combinatorics [23, Theorem VI.1 and Theorem VI.3] we also get the subexponential terms and recover the well-known result for Catalan numbers $b_{n+1} = \frac{1}{n+1}\binom{2n}{n}$ (see OEIS A000108 [41]):

$$b_n = \frac{4^{n-1}}{\sqrt{\pi n^3}}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

for $n \to \infty$. ■

We will now describe this approach applied to the grammar specifying the $\mathbb{DL}$-histories with an unranked species tree $\mathbf{S}$. Let $h_{u,n}$ be the number of $\mathbb{DL}$-histories of $\mathbf{S}_u$ consisting of n genes represented in the generating function by the formal variable z. We define the counting generating functions

$$H_u(z) = \sum_{n\geq 0} h_{u,n} z^n.$$

The coefficients $h_{u,n}$ represent the number of histories of size n associated with the species tree $\mathbf{S}_u$ independent on the number of losses or duplications. These generating functions (one per species u of $\mathbf{S}$) are strongly related to the generating function of binary trees $B(z)$ introduced in Example 3.1.

Lemma 3.1 *For a given species tree* $\mathbf{S}$ *the counting generating function* $H_{r(\mathbf{S})}(z)$ *for histories in the unranked* $\mathbb{DL}$*-model is defined by the system of functional equations*

$$\begin{aligned} H_u(z) &= B\left(H_{u_\ell}(z)H_{u_r}(z) + H_{u_\ell}(z) + H_{u_r}(z)\right) && \text{if } u \text{ is internal,} \\ H_u(z) &= B\left(z\right) && \text{if } u \text{ is a leaf,} \end{aligned} \tag{14}$$

over all nodes u of $\mathbf{S}$, where

$$B(z) = \frac{1 - \sqrt{1-4z}}{2}.$$

Proof The symbolic method [23, Part A] translates the unranked $\mathbb{DL}$-grammar of Equations (1)-(5) directly into a system of equations for the generating functions. We get

$$\begin{aligned} H_u(z) &= H_u(z)^2 + H_{u_\ell}(z)H_{u_r}(z) + H_{u_\ell}(z) + H_{u_r}(z) && \text{if } u \text{ is internal,} \\ H_u(z) &= H_u(z)^2 + z && \text{if } u \text{ is a leaf.} \end{aligned} \tag{15}$$

Comparing these equations with the one for binary trees from Equation (13) the claim follows. □

The advantage of a generating function approach is that we are able to identify the subexponential growth as $n^{-3/2}$, and that we are able to explicitly compute exponential growth $\rho_{\mathbf{S}}^{-1}$ and the constant $\gamma_{\mathbf{S}}$ for a fixed species tree $\mathbf{S}$. We will compute the involved constants explicitly for the caterpillar tree in Section 4.1.1 and for the complete binary tree in Section 4.1.2.

By basic principles of analytic combinatorics, the asymptotic growth of a counting sequence is directly related to the radius of convergence of the corresponding generating function. In particular, its dominant singularity (i.e. the one closest to the origin) defines its asymptotic growth. By the construction in terms of nested radicals, the generating function $H_u(z)$ is singular if and only if at least one of its radicals becomes zero. Therefore, we make the structure of nested radicals visible. Writing the explicit form of the outermost $B(z)$ in (14) gives

$$H_u(z) = \frac{1 - \sqrt{R_u(u)}}{2}. \tag{16}$$

Then, the radicands satisfy the following recurrence

$$\begin{aligned} R_u(z) &= -4 + 3\sqrt{R_{u_\ell}(z)} + 3\sqrt{R_{u_r}(z)} - \sqrt{R_{u_\ell}(z)R_{u_r}(z)} && \text{if } u \text{ is internal,} \\ R_u(z) &= 1 - 4z && \text{if } u \text{ is a leaf.} \end{aligned} \tag{17}$$

The recurrence can be used to determine the nature of the radii of convergence. For a node u we define ρ_u as the radius of convergence of $H_u(z)$.

Lemma 3.2 *Let u be the parent of v in $\mathbf{S}$. Then, $\rho_u < \rho_v$ and $\rho_u \in (0, 1/4]$ with $\rho_u = 1/4$ if u is a leaf. Furthermore, $R_u(z)$ is the only radicand that vanishes at $z = \rho_u$ and ρ_u is a simple root.*

Proof By combinatorial construction $H_u(z)$ is built of nested radicals and does not include any poles. Therefore, its dominant singularity must be at a point where (at least) one of its radicands vanishes.

We continue by induction on the depth of the subtree with root u given by $\mathbf{S}_u$. The depth is the longest path from the root to any leaf. As a first step, we prove that $R_u(0) = 1$ and that $\rho_u \leq 1/4$. For a leaf u it is clear from Relation (17) that $R_u(0) = 1$ and that $\rho_u = 1/4$.

Next, let v and w be the children of u such that $\rho_v \leq \rho_w$. By the induction hypothesis we directly get

$$R_u(0) = -4 + 3\sqrt{R_v(0)} + 3\sqrt{R_w(0)} - \sqrt{R_v(0)R_w(0)} = 1.$$

In order to continue, note that $R_u(z)$ is monotonically decreasing on $[0, +\infty]$, because from the decomposition in (16) and (15) we see that

$$R_u(z) = 1 - \sum_{n \geq 1} a_n z^n, \tag{18}$$

for certain non-negative numbers a_n.

By the induction hypothesis and Relation (17), $R_u(z)$ is a continuous function on $(0, \rho_v)$. Hence, we get

$$R_u(\rho_v) = -4 + 3\sqrt{R_w(\rho_v)} < 0.$$

Thus, on the one hand, by the intermediate value theorem $R_u(z)$ must have at least one zero in the interval $(0, \rho_v)$. On the other hand, as $R_u(z)$ is monotonically decreasing it has at most one zero in $(0, \rho_v)$. Hence, this zero is equal to ρ_u.

Finally, the above reasoning implies that among the nested radicals of $H_u(z)$ the outermost one is the first one that vanishes, and no other radical vanishes at the same time. Thus, ρ_u is the radius of convergence of $H_u(z)$. Moreover, by (18) we see that the derivative $R'_u(z)$ has non-positive coefficients. Hence, ρ_u is a simple root. □

Let us shortly digress and discuss in a more general context how to numerically compute the exponential growth for the coefficients of the generating function with the fastest exponential growth that is defined by a system of functional equations involving generating functions $B_1, \ldots, B_k$ of the form

$$B_i = \Phi_i\left(z, B_1, \ldots, B_k\right),$$

where the Φ_i are polynomials with non-negative integer coefficients in $k+1$ variables. Note that the grammar given in Theorem 3.1 is of this shape. In order to decide which of the B_i's has this specific exponential growth, further information on the problem, like in our case given by Lemma 3.2, is needed. By Banach's fixed point theorem, these equations admit a unique solution vector $(B_1, \ldots, B_k) \in (\mathbb{C}[[z]])^k$ with respect to the formal topology [23, Section A.5]. Furthermore, each $B_i(z)$ has non-negative coefficients in its expansion around 0 (which is already clear from the combinatorial nature of the problem). Then, the multivariate version of the implicit function theorem implies that each of them has a non-zero radius of convergence which we call ρ_i. By Pringsheim's Theorem [23, Theorem IV.6], $\rho_i \in [0, +\infty]$ is a singularity of $B_i(z)$. Moreover, as $B_i(z)$ is an ordinary generating function of an infinite combinatorial class, we must have $\rho_i \in [0, 1]$. Finally, in order to compute the radius of convergence, we find the minimal point $z \in [0, 1]$ where the implicit function theorem fails. To be more precise, we numerically compute solutions $\rho \in [0, 1]$ and $b_1, \ldots, b_k \in [0, +\infty)$ of the following system

$$\begin{cases} b_1 = \Phi_1(\rho, b_1, \ldots, b_k) \\ \quad \vdots \\ b_k = \Phi_k(\rho, b_1, \ldots, b_k) \\ 0 = \det\left(\delta_{i,j} - \frac{\partial}{\partial b_j}\Phi_i(\rho, b_1, \ldots, b_k)\right), \end{cases}$$

where $\delta_{i,j}$ is the Kronecker symbol: $\delta_{i,i} = 1$, and $\delta_{i,j} = 0$ for $i \neq j$.

Remark 3.2 The unranked $\mathbb{DL}$-grammars lead to the following specific shape

$$\begin{cases} B_1 = \Phi_1(z, B_1) \\ B_2 = \Phi_2(z, B_1, B_2) \\ \quad\vdots \\ B_k = \Phi_k(z, B_1, \ldots, B_k) \end{cases}$$

Hence, we get $\det\left(\delta_{i,j} - \frac{\partial}{\partial b_j}\Phi_i(\rho, b_1, \ldots, b_k)\right) = \prod_{i=1}^{k}(1-2b_i)$. We actually know by Lemma 3.2 that the outermost square-root vanishes, which gives $b_k = B_k(\rho) = 1/2$. Additionally, we can also directly deduce from this system that $\rho_k \leq \rho_{k-1}$.

In the unranked $\mathbb{DLT}$-model the system looks like

$$\begin{cases} B_1 = \Phi_1(z, B_1, B_2, \ldots, B_{k-1}) \\ \quad\vdots \\ B_{k-1} = \Phi_{k-1}(z, B_1, \ldots, B_{k-1}) \\ B_k = \Phi_k(z, B_1, \ldots, B_k) \end{cases}$$

where the last equation is the only one involving B_k, as the root can not be a receiver of an HGT. Note that the subsystem of the first $k-1$ equations is strongly connected and but still not satisfies the a-properness condition (i.e. it is no contraction in the formal topology) of the Drmota–Lalley–Woods Theorem [23, Theorem VII.6] which would directly imply a square root singularity. Thus, we conjecture that the dominant singularity still comes solely from the outermost square root of B_k implying $b_k = 1/2$.

In the ranked $\mathbb{DLT}$-model we are dealing with blocks of strongly connected components that correspond to the time slices. Note that the root is contained in a singleton time slice. Experiments suggest the same behavior as in the previous cases.

However, one thing is for sure in all models: we always have $\rho_{r(\mathbf{S})} \leq \rho_u$ for all other subtrees with root u of the species tree. Hence, there will be always a dominant minimal singularity in $[0, 1]$ that can be (numerically) computed. Note however, that the determinant computation soon becomes extremely heavy.

After determining the radius of convergence, we must determine the number of singularities on it. As shown in the case of λ-terms in [8, Lemma 8] there can only be one dominant singularity ρ_u. Let us quickly repeat this argument here. Assume that there exists a root $z_0 = \rho_u e^{i\theta}$ of the same modules. Substituting this value into $R_u(z)$ from (18) gives

$$1 = \sum_{n\geq 1} a_n \rho_u^n = \Big|\sum_{n\geq 1} a_n z_0^n\Big|,$$

which can only hold if $e^{in\theta} = 1$ whenever $a_n \neq 0$. Now, due to $a_1 \neq 0$ we have $z_0 = \rho_u$. Hence, ρ_u is the unique dominant real singularity of $H_u(z)$.

Combining the previous results, we have shown for a family of constants $\gamma_{u,i}$ the following local singular expansion

$$H_u(z) = \frac{1}{2} - \sum_{i\geq 0} \gamma_{u,i}\,(1 - z/\rho_u)^{i+1/2}\,.$$

The fact that $R_u(z)$ has a simple root at $z = \rho_u$ shows that $\gamma_{u,0} > 0$. Then, by transfer theorems of analytic combinatorics [23, Theorem VI.1 and Theorem VI.3], we get the

claimed asymptotic expansion of Equation (12), where $\gamma_T = \frac{\gamma_{u,0}}{2\sqrt{\pi}} > 0$ and this ends the proof of Theorem 3.3.

Remark 3.3 There are several possible extensions of the previous approach. First of all, it is straightforward to extend it to the ranked $\mathbb{DL}$-model. In that case one only needs to incorporate unary nodes arising from the time slices. Second, an extension to the $\mathbb{DLT}$-model is also possible, yet the computations are more involved as the binary tree structure leading to Lemma 3.1 does not hold anymore. However, it can still be modeled with colored binary trees, where the number of colors depends on the size of the set of incomparable nodes (in the the current time slice). Third, it is also possible to consider the distribution of certain parameters, such as the number of gene losses, or the number of gene duplications, see e.g. for related results in lattice paths and trees [4, 10, 25]. Using multivariate generating functions and marking each such event by an additional variable like in the general grammar of Theorem 3.1, the above results for the $\mathbb{DL}$-model directly generalize to the respective ones on multivariate generating functions. All these generalizations are interesting future research directions.

The counting and sampling algorithms described above have been implemented in `Python`, and are available at `https://github.com/cchauve/DLTcount`.

4 Results

Over the next two sections, we will apply Theorem 3.3 to the special cases of the caterpillar and complete species tree in the unranked $\mathbb{DL}$-model, and explicitly determine the constants involved in the asymptotic expansion. Then, we apply our dynamic programming counting and sampling algorithms to study properties of random evolutionary histories.

4.1 Asymptotic expansion for extremal species trees in the $\mathbb{DL}$-model

Our experimental results (Section 4.2) suggest that for a given k, the species trees having the largest (resp. smallest) number of $\mathbb{DL}$-histories are respectively the caterpillar tree and the balanced binary tree (Conjecture 4.1), defined below. In the present section, our main results are the explicit computation of the asymptotic growth and the leading constant of Theorem 3.3 for the caterpillar species tree (Propositions 4.1 and 4.2) and for the complete binary species tree, the special case of balanced trees when k is a power of 2 (Propositions 4.3 and 4.4, see also Table 2).

The rooted caterpillar tree $\mathbf{CT}_k$ can be defined as follows: $\mathbf{CT}_1$ is the tree reduced to a single leaf, while $\mathbf{CT}_k$ ($k > 1$) is the tree formed by a left subtree equal to $\mathbf{CT}_{k-1}$ and a right subtree equal to $\mathbf{CT}_1$. Observe that every subtree of a caterpillar tree is itself a caterpillar tree, see Figure 4.

The complete binary tree $\mathbf{CB}_h$ with $k = 2^h$ leaves can be defined as follows: $\mathbf{CB}_0$ is the tree reduced to a single leaf, while $\mathbf{CB}_h$ ($h \geq 1$) is the tree formed by a left and a right subtree both equal to $\mathbf{CB}_{h-1}$. Observe again that every subtree is itself a complete binary tree, see Figure 4. The complete binary tree is a special case of the class of *balanced* trees, defined as trees where, for each node, the number of leaves in the left subtree differs from the number of leaves in the right subtree by at most one.

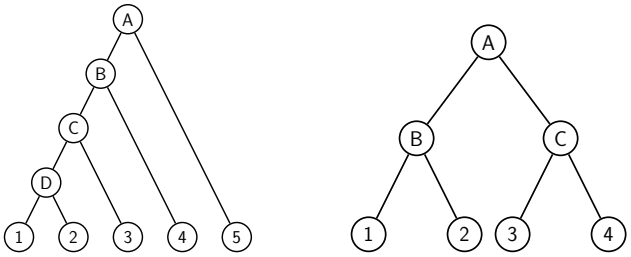

Fig. 4: (Left) The caterpillar species tree $\mathbf{CT}_5$. (Right) The complete binary tree $\mathbf{CB}_2$.

Complete binary trees are the only balanced trees in which the number of leaves is a power of two.

We can observe that the number of $\mathbb{DL}$-histories grows much faster for the caterpillar tree than for the complete binary tree. This is actually unsurprising given that the number of $\mathbb{DL}$-histories can be linked to the size of the grammar, which itself depends on the structure of the species tree. More precisely, the size of the grammar depends on the number of unique subtrees of the considered species tree S. Each such subtree may be identified by its root u and corresponds to one set of rules (1)-(6), while subtrees having the same topology lead to isomorphic subgrammars with the same counting generating functions. The caterpillar (resp. complete binary) tree has the largest (resp. smallest) number of unique subtrees within the set of species trees of the same size (when k is a power of 2 for the complete binary tree), compare also Table 2.

#Species	Caterpillar tree $\mathbf{CT}_k$		Complete binary tree $\mathbf{CB}_k$	
k	α_k	Exp. Growth λ_k^{-1}	β_h	Exp. Growth μ_h^{-1}
1	0.1410	4.00	0.1410	4.00
2	0.1557	9.61	0.1557	9.61
3	0.1647	15.72	–	–
4	0.1742	22.69	0.1620	20.75
5	0.1835	30.53	–	–
6	0.1927	39.25	–	–
7	0.2015	48.84	–	–
8	0.2101	59.31	0.1650	43.02
9	0.2184	70.65	–	–
10	0.2265	82.86	–	–
11	0.2342	95.93	–	–
12	0.2418	109.85	–	–
13	0.2491	124.64	–	–
14	0.2563	140.28	–	–
15	0.2632	156.77	–	–
16	0.2700	174.11	0.1664	87.56

Table 2: Leading constants and exponential growth factors for the number of $\mathbb{DL}$-histories consistent with the unranked caterpillar and complete species tree. Their closed forms are given in Propositions 4.1–4.4.

k	Sequence	OEIS
1	$1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, \ldots$	A000108
2	$2, 7, 34, 200, 1318, 9354, 69864, 541323, 4310950, 35066384, \ldots$	A307696
3	$3, 19, 159, 1565, 17022, 197928, 2413494, 30490089, 395828145, \ldots$	A307697
4	$4, 39, 495, 7235, 115303, 1948791, 34379505, 626684162, \ldots$	A307698
5	$5, 69, 1230, 24843, 541315, 12426996, 296546600, 7292489761, \ldots$	A307700

Table 3: $\mathbb{DL}$-history counting sequences of the caterpillar species trees $\mathbf{CT}_k$.

4.1.1 Counting $\mathbb{DL}$-histories associated with the caterpillar species tree

Denote by $H_k^{\mathbf{CT}}$ the set of $\mathbb{DL}$-histories over the caterpillar $\mathbf{CT}_k$, then the general grammar of $\mathbb{DL}$-histories, where extant genes are marked by a single terminal $\mathcal{Z}$, is the following:

$$H_k^{\mathbf{CT}} = D_k^{\mathbf{CT}} + S_k^{\mathbf{CT}} \qquad \text{if } k > 1 \tag{19}$$

$$H_1^{\mathbf{CT}} = \mathcal{Z} + D_1^{\mathbf{CT}} \tag{20}$$

$$S_k^{\mathbf{CT}} = H_{k-1}^{\mathbf{CT}} \times H_0^{\mathbf{CT}} + H_{k-1}^{\mathbf{CT}} + H_0^{\mathbf{CT}} \qquad \text{if } k > 1 \tag{21}$$

$$D_k^{\mathbf{CT}} = H_k^{\mathbf{CT}} \times H_k^{\mathbf{CT}} \tag{22}$$

Let $f_{k,n}$ be the number of $\mathbb{DL}$-histories of the caterpillar $\mathbf{CT}_k$ consisting of n genes. The corresponding counting generating function is given by

$$F_k(z) = \sum_{n \geq 0} f_{k,n} z^n,$$

and, by Lemma 3.1, it is defined by the functional equation

$$F_k(z) = B\left(F_{k-1}^2(z) + F_{k-1}(z) + B(z)\right).$$

In Table 3 we computed the first few initial terms for $k = 1, ..., 5$. Note that none but the first one was found in the OEIS [41] before we added them. Applying Theorem 3.3, the asymptotic expansion of the coefficients for $n \to \infty$ is

$$f_{k,n} = \alpha_k \frac{\lambda_k^{-n}}{n^{3/2}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right). \tag{23}$$

for some constants $\alpha_k > 0$ and $\lambda_k > 0$ that are made explicit below.

Proposition 4.1 *Let $a(X) = 3X - 4$ and $b(X) = X - 3$. We define the following sequence of rational functions in X*

$$\begin{cases} s_1(X) = 0, \\ s_k(X) = \frac{a(X) - s_{k-1}(X)^2}{b(X)} \quad \text{for } k > 1. \end{cases}$$

Let X_k be the minimal positive real solution of the fixed point equation

$$s_k(X) = X.$$

Then, the dominant singularity of $F_k(z)$ can be found at $\lambda_k = \frac{1 - X_k^2}{4}$.

Proof We need to analyze the nested radicals of $F_k(z)$ in more detail. Therefore, as done in Equation (16) for the general case, we define the decomposition

$$F_k(z) = \frac{1 - \sqrt{P_k(z)}}{2}.$$

Thus, we directly get the specialized version of the reucurrence for the radicands from Equation (17) by

$$\begin{cases} P_1(z) = 1 - 4z, \\ P_k(z) = -4 + 3\sqrt{1-4z} + (3 - \sqrt{1-4z})\sqrt{P_{k-1}(z)}, & \text{for } k > 1. \end{cases} \tag{24}$$

The dominant singularity λ_k is given by the minimal positive root of $P_k(z)$. This already proves the case $k = 1$. We introduce the shorthand $X = \sqrt{1-4z}$ and use it from now on as our new variable. This directly gives

$$P_k(X) = a(X) - b(X)\sqrt{P_{k-1}(X)}. \tag{25}$$

Hence, this equation is zero if and only if

$$\sqrt{P_{k-1}(X)} = \frac{a(X)}{b(X)} =: s_2(X).$$

For $k = 2$ this proves the claim as $\sqrt{P_1(X)} = X$. Now we proceed by induction. Squaring this equation and substituting the known expression for $P_{k-1}(X)$ gives

$$\sqrt{P_{k-2}(X)} = \frac{a(X) - s_2(X)^2}{b(X)} =: s_3(X).$$

Repeating this process proves the claim. □

Proposition 4.2 *Using the notation of Proposition 4.1, the constant α_k is equal to*

$$\alpha_k = \sqrt{\frac{\lambda_k}{8\pi X_k} \sum_{i=2}^{k+1} \sigma_{i,k}(X_k) \left(\frac{3 - X_k}{2}\right)^{i-2} \prod_{j=2}^{i-1} \frac{1}{s_j(X_k)}},$$

$$\sigma_{i,k}(X) = \begin{cases} 3 - s_i(X) & \text{if } i \leq k, \\ 2X & \text{if } i = k+1. \end{cases}$$

In particular, $\alpha_k > 0$.

Proof We will prove that $P_k(X)$ admits the following extension in a neighborhood of X_k:

$$P_k(X) = P_k'(X_k)(X - X_k) + \mathcal{O}((X - X_k)^2),$$

where the derivative is with respect to X. Note that this derivative exists, as $P_k(X)$ is analytic on $(0, (1 - X_{k-1}^2)/4)$ and we know from Lemma 3.2 that $X_{k-1} < X_k$.

Next, recall the shorthand $X = \sqrt{1-4z}$ and that by the chain rule $\partial_z P_k(X) = \partial_X P_k(X) \partial_z X$. Then, the transfer theorems of analytic combinatorics [23] directly show that the n-th coefficient of $F_k(z)$ satisfies the form (12) with $\alpha_k = \sqrt{\lambda_k P_k'(X_k)/(8\pi X_k)}$. Therefore, it remains to find an expression for $P_k'(X_k)$.

h	k	Sequence	OEIS
0	1	$1, 1, 2, 5, 14, 42, 132, 429, 1430, 4862, 16796, 58786, 208012, 742900, \ldots$	A000108
1	2	$2, 7, 34, 200, 1318, 9354, 69864, 541323, 4310950, 35066384, \ldots$	A307696
2	4	$4, 34, 368, 4685, 66416, 1013268, 16279788, 271594611, 4660794200, \ldots$	A307941
3	8	$8, 148, 3376, 89390, 2624872, 82866636, 2755019736, 95135709027, \ldots$	A307942
4	16	$16, 616, 28832, 1556780, 93017264, 5971377672, 403667945712, \ldots$	A307943

Table 4: $\mathbb{DL}$-history counting sequences of the complete species trees $\mathbf{CB}_h$ with $k = 2^h$ leaves.

Let us take the derivative of Equation (25). We get

$$P'_k(X) = 3 - \sqrt{P_{k-1}(X)} + \frac{3 - X}{2\sqrt{P_{k-1}(X)}} P'_{k-1}(X).$$

In the proof of Proposition 4.1 we have seen that $\sqrt{P_i(X_k)} = s_{k-i+1}(X_k)$. Iterating this equation until $P'_1(X) = 2X$ shows the claim. Finally, the positivity of the constant holds as all terms are positive. □

With these formulas it is easy to compute explicit values for the constant α_k and the asymptotic growth factor λ_k^{-1}. We show the first few values in Table 2.

4.1.2 Counting $\mathbb{DL}$-histories associated with the complete species tree

Let $H_h^{\mathbf{CB}}$ be the set of $\mathbb{DL}$-histories associated with the complete binary tree $\mathbf{CB}_h$. Then, the respective grammar, considering again only terminals $\mathcal{Z}$ marking extant genes, is the following:

$$H_h^{\mathbf{CB}} = D_h^{\mathbf{CB}} + S_h^{\mathbf{CB}} \qquad \text{if } h \geq 1 \tag{26}$$

$$H_0^{\mathbf{CB}} = \mathcal{Z} + D_0^{\mathbf{CB}} \tag{27}$$

$$S_h^{\mathbf{CB}} = H_{h-1}^{\mathbf{CB}} \times H_{h-1}^{\mathbf{CB}} + H_{h-1}^{\mathbf{CB}} + H_{h-1}^{\mathbf{CB}} \qquad \text{if } h \geq 1 \tag{28}$$

$$D_h^{\mathbf{CB}} = H_h^{\mathbf{CB}} \times H_h^{\mathbf{CB}} \tag{29}$$

Let $g_{h,n}$ be the number of histories over the complete binary tree $\mathbf{CB}_h$ consisting of n genes represented by z. As before, we analyze the counting generating function which is given by

$$G_h(z) = \sum_{n \geq 0} g_{h,n} z^n,$$

and, by Lemma 3.1, it is defined by the functional equation

$$G_h(z) = B\left(G_{h-1}^2(z) + 2G_{h-1}(z)\right).$$

As before, we computed the first few initial terms in Table 4. Again, none but the first one was found in the OEIS [41] before we added them.

Applying Theorem 3.3 gives the asymptotic expansion of the coefficients for $n \to \infty$ as

$$g_{h,n} = \beta_h \frac{\mu_h^{-n}}{n^{3/2}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right),$$

where $\beta_h > 0$ and $\mu_h > 0$ are nonnegative constants computed as follows.

Proposition 4.3 *The dominant singularity of* $G_h(z)$ *is* $\mu_h = \frac{1-q_h}{4}$*, where*

$$\begin{cases} q_0 = 0, \\ q_{h+1} = (3 - \sqrt{5 - q_h})^2 \quad \text{for } h \geq 0. \end{cases}$$

Furthermore, q_h *and* μ_h *are algebraic numbers of degree* 2^h*.*

Proof As for the caterpillar tree, we need to analyze the nested radicals. To make this structure visible, we again define

$$G_h(z) = \frac{1 - \sqrt{Q_h(z)}}{2}. \tag{30}$$

Then, the radicands satisfy the following recurrence

$$\begin{cases} Q_0(z) = 1 - 4z, \\ Q_{h+1}(z) = -4 + 6\sqrt{Q_h(z)} - Q_h(z), \quad \text{for } h \geq 0. \end{cases} \tag{31}$$

When comparing it with the recurrence of radicands for the caterpillar grammar in (24) we notice a major difference: the coefficients are independent of z.

Then, the reasoning follows the same lines as the proof of Proposition 4.1. Yet, due to the independence of the coefficients of z, the induction yields an explicit expression. Note that $Q_{h-i}(\mu_h) = q_i$. □

In a similar way we are also able to compute the constant β_h explicitly.

Proposition 4.4 *Using the notation of Proposition 4.3, the constant* β_h *is equal to*

$$\beta_h = \sqrt{\frac{\mu_h}{16\pi} \prod_{i=1}^{h-1} \left(\frac{3}{q_i^2} - 1 \right)}.$$

Proof By Equation (30) the singularity of $G_h(z)$ is determined by the smallest root μ_h of $Q_h(z)$. The constant is determined by the expansion for $z \to \mu_h$:

$$Q_h(z) = b_h(z - \mu_h) + \mathcal{O}\left((z - \mu_h)^2\right).$$

By the recursive definition, $Q_h(z)$ is differentiable in $(0, \mu_{h-1})$ due to $\mu_h < \mu_{h-1}$. Thus, $b_h = Q_h'(\mu_h)$ is well-defined. Differentiating the recurrence of $Q_h(z)$ we get

$$Q_h'(z) = \left(\frac{3}{\sqrt{Q_{h-1}(z)}} - 1 \right) Q_{h-1}'(z).$$

Iterating this relation and applying $Q_{h-i}(\mu_h) = q_i$ proves the claim. □

As before, we computed the first few explicit values for the constant β_h and the asymptotic growth factor μ_h^{-1}, where h is a power of 2, and show them in Table 2.

4.2 Empirical investigations and open questions

In this section we present empirical results and observations derived using the counting and sampling algorithms described in Section 3.2. These results provide the first detailed view, especially in the $\mathbb{DL}$-model, of the general question: in how many ways can n genes have evolved from a single ancestral gene, for a given species tree?

4.2.1 Counting histories for random species trees

We are first interested in computing the number of histories in a given evolutionary model. We considered the following models: $\mathbb{DL}$-histories with an unranked or ranked species tree (called respectively models $\mathsf{u}\mathbb{DL}$ and $\mathsf{r}\mathbb{DL}$ from now), $\mathbb{DLT}$-histories with an unranked species tree or a ranked species tree (called respectively models $\mathsf{u}\mathbb{DLT}$ and $\mathsf{r}\mathbb{DLT}$ from now).

For a given evolutionary model and species tree S of size k, let $h_S(n)$ be the number of histories of size n. As shown in Equation (12) for the $\mathsf{u}\mathbb{DL}$-model, this number grows asymptotically with n as follows

$$h_S(n) \simeq \gamma_S \frac{\rho_S^{-n}}{n^{3/2}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

where γ_S and ρ_S, both depend only on S. From now, we denote $\alpha_S = \rho_S^{-1}$ the *exponential growth factor* for the number $h_S(n)$. In the $\mathsf{u}\mathbb{DL}$-model, as discussed in Section 3.3, we can compute precisely the growth factor from the grammar specifying the $\mathbb{DL}$-histories for the given species tree S. For other models, we can estimate λ_S from the number $h_S(n)$ of histories of size n as follows:

$$\alpha_S \simeq \frac{h_S(n)}{h_S(n-1)}, \tag{32}$$

this estimate precision increasing naturally with n.

$\mathbb{DL}$-models. We considered species trees of size ranging from $k = 3$ to $k = 25$ and for each species tree size k, we generated 98 random species tree of size k under the uniform distribution, using the RANRUT algorithm described in [33], and we completed this set of species tree by adding the caterpillar species tree with k leaves and the balanced tree with k leaves[3]; so for small values of k, the same species tree can occur several times in the sample of 100 trees. When working in the $\mathsf{r}\mathbb{DLT}$-model, we generated, for each species tree 10 random rankings under the uniform distribution, using the algorithm described in [8]. Then, for each instance, we computed the number of histories of size $n = 50$ in the models $\mathsf{u}\mathbb{DL}$, $\mathsf{u}\mathbb{DLT}$ and $\mathsf{r}\mathbb{DLT}$[4] and used these numbers to estimate the growth factor using (32).

Figure 5 shows the exponential growth factor in the $\mathsf{u}\mathbb{DL}$-model obtained using the exact approach described in Section 3.3 and the ratio between this exact growth factor and the growth factor estimated using the experimental approach described above. A

[3] Note that for a given k, any two balanced ordered binary trees with k leaves differ only by swapping the left and right children of some internal nodes, so for our purpose there is essentially a unique balanced species tree for every value of k.

[4] We omit here the results for the $\mathsf{r}\mathbb{DL}$-model as they are very similar to the results for the $\mathsf{u}\mathbb{DL}$-model, with a lower dispersion.

first observation from Figure 5 is that estimating the growth factor from the number of histories of size $n = 50$ approximates well the exact growth factor in the $\mathsf{u}\mathbb{DL}$-model; we believe it is also the case in the other models (data not shown).

Moreover, following up on the results shown in Table 2, our experiments lead to the following conjecture, characterizing the species trees leading to extreme growth factors for a given value of k.

Conjecture 4.1 For a given k, and n large enough, the unranked species tree of size k having the largest number of $\mathbb{DL}$-histories of size n is the caterpillar tree; moreover the exponential growth factor of the number of histories for a caterpillar of size k grows superlinearly as a function of k. Species trees having the smallest number of $\mathbb{DL}$-histories are balanced species trees of size k and the exponential growth factor of the number of histories for a balanced tree of size k grows linearly as a function of k.

We verified that the conjecture is true for all values of k in our experiments. We investigated several proof ideas, in particular linking the exponential growth factor to the number of unique subtrees in a species tree. Indeed this is a feature for which caterpillar and balanced trees reach extreme values for a given value of k; actually the caterpillar is the unique tree with the maximum number of subtrees, while balanced trees have the minimum number of subtrees, although if k is not a power of 2, some unbalanced trees can have the same number of subtrees than balanced ones. We did find examples of pairs of species trees for which the one with the larger (resp. smaller) number of unique subtrees has a smaller (resp. larger) exponential growth factor. There are also species trees with the same number of unique subtrees than balanced trees of the same size and showing a larger exponential growth rate. So the number of unique subtrees is not the determinant leading to an extreme growth factor. We observed similar examples when considering the height of the species tree, another feature for which caterpillar and balanced trees attain extreme values. Generally the question of understanding which features of species trees of the same size that makes one having more $\mathbb{DL}$-histories than the other one is open.

$\mathbb{DLT}$-models. Next, we consider models including HGT; in Figure 6 we show the estimated growth constants in the $\mathsf{u}\mathbb{DLT}$- and $\mathsf{r}\mathbb{DLT}$-models.

An observation that addresses one of the main questions motivating our work, is that the number of histories in models involving HGT grows much faster than in models excluding HGT; this is apparent by comparing the growth factors in the $\mathsf{u}\mathbb{DL}$ and $\mathsf{u}\mathbb{DLT}$ models, but even more through Figure 7 that shows the ratio of the number of $\mathbb{DLT}$-histories over the number of $\mathbb{DL}$-histories for selected pairs (k, n), considered over all randomly chosen ranked or unranked species trees. We can observe that the ratios grow as large as 10^{40} in the unranked model and 10^{29} in the ranked model for histories of size 50 over a species tree of size 25, that correspond to parameters of realistic phylogenomics datasets. It is nevertheless interesting to observe that considering ranked species trees tames significantly the magnitude of the search space explosion when introducing HGT in a model.

Finally, we can observe that in the $\mathsf{r}\mathbb{DLT}$-model, the growth factor seems to be almost independent of the topology of the chosen species tree and ranking (Figure 6 (Bottom)). Intuitively, this can be explained by the fact that a ranked species tree can almost be seen as a sequence of time slices, each composed of a set of branches (from 1 branch for the time slice containing the root of S to k branches for the time slice

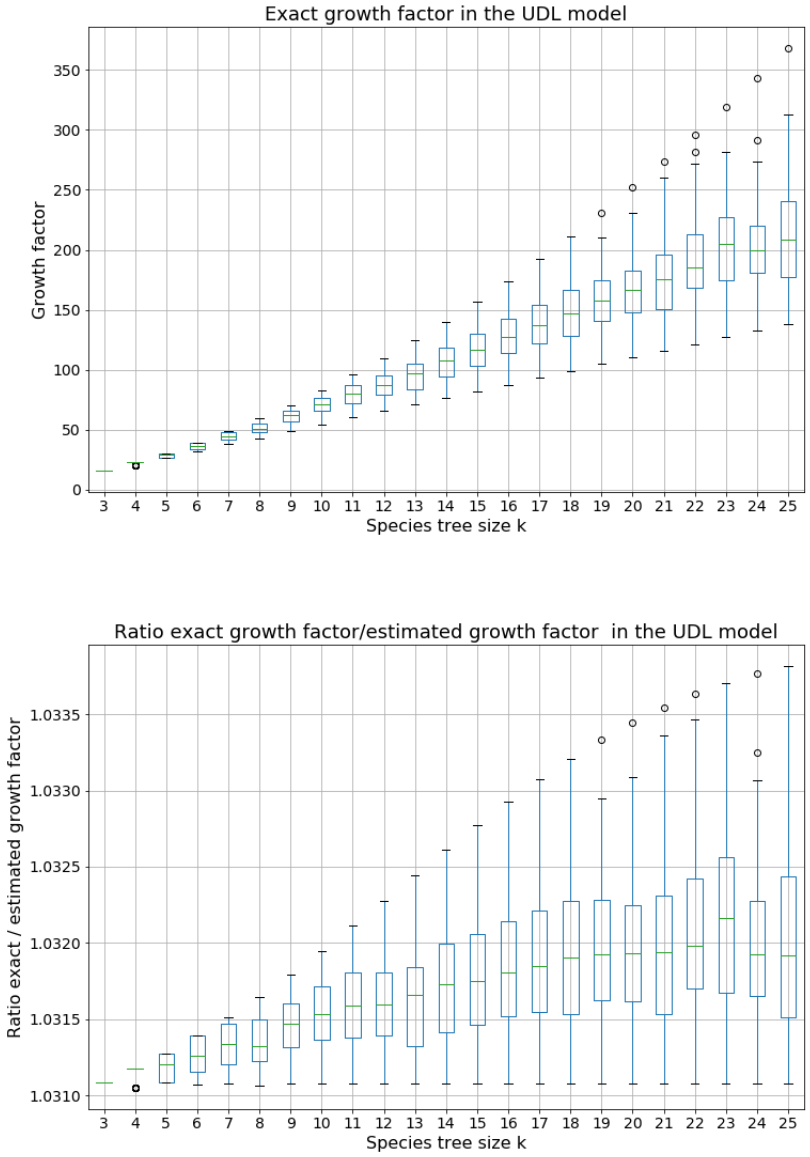

Fig. 5: Box-plot of the distribution of the growth factor for each 100 random species tree per size k in the u$\mathbb{DL}$-model. (Top) Exact growth factor; (Bottom) Box-plot of the distribution, for each species tree, of the ratio between the exact growth factor and the estimated growth factor.

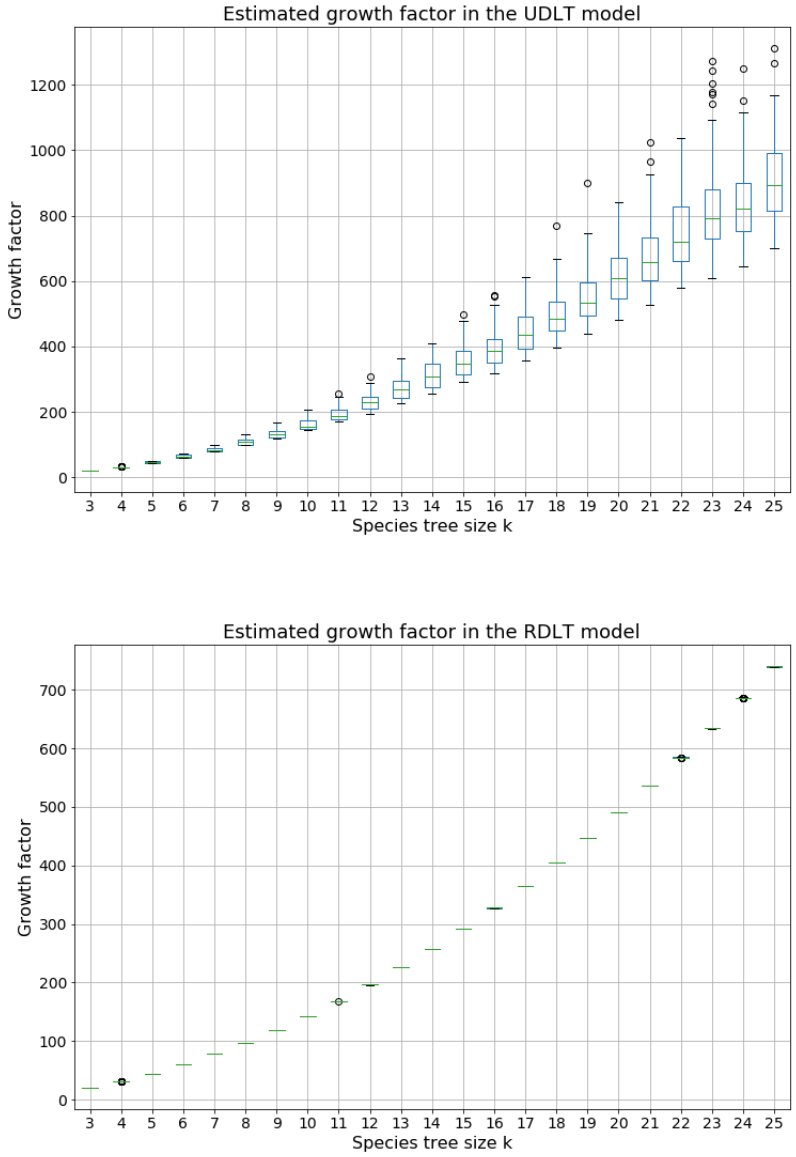

Fig. 6: Box-plot of the distribution of the growth factor for each 100 random species tree per size k in the $\mathsf{u}\mathbb{DLT}$ (Top) and $\mathsf{r}\mathbb{DLT}$ (Bottom) models. The growth factor is estimated from the number of $\mathbb{DLT}$-histories of size $n = 50$ using formula (32).

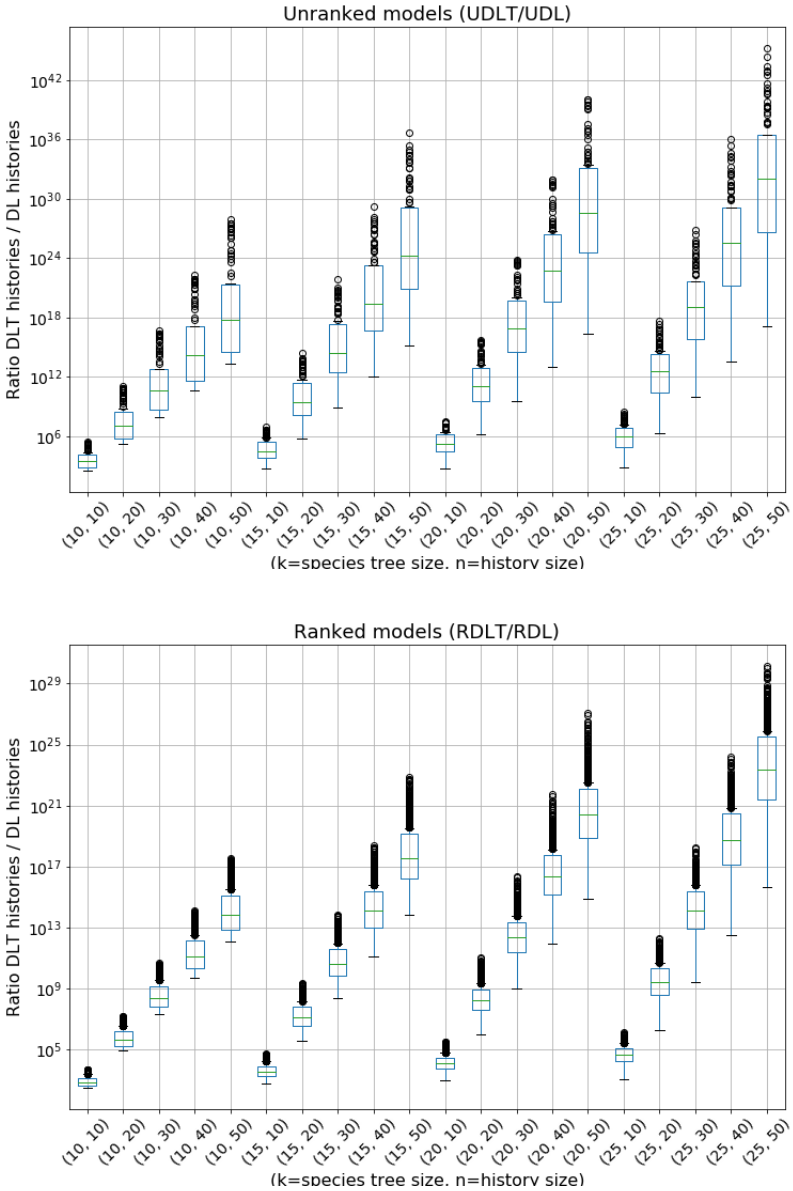

Fig. 7: Box-plots of the distribution of the ratio of the number of DLT-histories over the number of $\mathbb{DL}$-histories over all species trees size k and histories size n for selected pairs (k, n). The distributions are obtained, for each (k, n), over 100 randomly chosen (resp. 1000) unranked (resp. ranked) species trees.

containing all leaves), with exactly one ending with a speciation node while all other end by a unary node. Within each time slice, the genes can evolve freely by duplication and HGT, where a duplication can be seen as equivalent to a HGT within the same branch. Thus, the number of histories is dominated by the number of evolutionary events taking place in each time slice, with some variability being introduced by the number of genes leaving a time slice right after the only speciation node it contains, that can create extra gene copies entering the next time slice.

In order to understand this phenomenon, we investigated a reduced evolutionary model, in which every speciation is followed by a random loss, i.e. does not create an extra gene copy entering the next time slice; we name this model the r$\mathbb{DT}$-$\mathbb{SL}$-model, where $\mathbb{SL}$ stands for *Speciation-Loss*. In this model, we are able to prove the independence of the chosen species trees.

Theorem 4.1 *In the r$\mathbb{DT}$-$\mathbb{SL}$-model, the number of histories of size n is the same for every ranked species tree of size k.*

Proof Let a ranked species tree of size k be given, and consider the unary-binary tree induced by its time slices. We then transform this tree into a directed graph called the *events graph* describing the possible events of duplication, HGT, and speciation in the following way:

1. Label the leaves from 1 to k.
2. Label each internal node with a set containing the labels of the leaves of its induced subtree. These labels are the possible leaves reachable by speciation;
3. Encode speciation events by super edges called *speciation edges* which consist of the one (unary) or two (binary) edges leading to the children of a node. By doing so, the two edges are treated as a single edge;
4. Encode duplication events by adding loops called *duplication edges* to each node;
5. Encode HGT events by adding edges called *transfer edges* from each node to each other node within the same time slice;

An example of this transformation is shown in Figure 8.

Let us briefly state some properties of the events graph. The labels of the nodes of each time slice form a set partition of $\{1, \ldots, k\}$ by construction. Due to the rankings, each time slice contains one node more than the previous one and every path from the root to the previous leaves contains $k - 1$ speciation edges.

The main idea of the proof is that we can encode an history H for a species tree S of size k by an ordered unary-binary tree H_e whose nodes are labeled by nodes of the events graph, that encodes unambiguously H, and then show that in the r$\mathbb{DT}$-$\mathbb{SL}$-model, given the events graph E' of another ranked species tree S' of the same size, we can transform H_e into an ordered unary-binary tree H'_e whose nodes are labeled by nodes of E' that encodes a unique history for S'. This establishes a one-to-one correspondence between the sets of histories for two arbitrary ranked species trees of size k, S and S', and thus proves the stated result.

The principle of the encoding is to associate each internal node of a history with a (deterministic) label which is a node of the events graph. Let E be the events graph of S. The encoding works as follows: for a node x of a history H for species tree S, if t is the time slice it belongs to and i its left-most leaf (defined in a depth-first traversal of the ordered tree representing the history), then we label x by the unique node of E in the time slice t that contains i. Extant leaves stay labeled by their extant

species. After deleting leaves corresponding to gene losses from the history, speciation-loss nodes become unary, while duplication and HGT nodes stay binary. Call H_e the ordered unary-binary tree for history H. The original history H can be unambiguously recovered from H_e and E, by reinserting these losses and removing the labels, as any edge of H_e corresponds to an edge of E, so defines an evolutionary event.

Next, let S' be another ranked species tree of the same size k as S and E' its events graph. We transform H_e into H'_e as follows: for every node x, whose left-most leaf is u and that belongs to time slice t, replace its label by the unique node of time slice t of E' that contains the u. This is always possible, as, by construction of the events graph in models with HGT, any leaf is reachable from any node. We claim that H'_e defines unambiguously a history for S'. The key argument to prove this claim is that, by the way we constructed E' and H'_e, for any edge in H'_e the labels of its two nodes, that are either in the same time slice or in consecutive time slices, are incident in E': if both nodes are in the same time slice, then by construction of E' they are either the same node (so linked by a duplication edge) or are incident by a transfer edge, while if they are in consecutive time slices, they contain a common species and so are incident by a speciation edge. It follows that H'_e encodes a history H' for S'. The construction from H to H' is deterministic and reversible, which provides a one-to-one correspondence between the histories of S and the histories of S' in the r$\mathbb{DT}$-$\mathbb{SL}$-model.

Note that this construction does not work in models with no duplication, HGT or unrestricted speciation as the key argument that any edge in H'_e can be found in E' does not hold anymore, thus preventing to be able to transform H'_e into a history for S'.

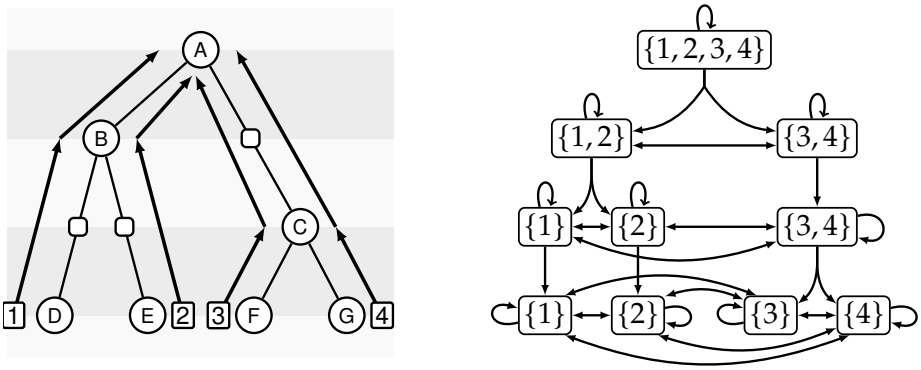

Fig. 8: Transformation of a ranked species tree (left) ($\pi(A) = 1, \pi(B) = 2, \pi(C) = 3, \pi(D) = \pi(E) = \pi(F) = \pi(G) = 4$) into an events graph (right) in the r$\mathbb{DT}$-$\mathbb{SL}$-model used in the proof of Theorem 4.1.

Remark 4.1 From the previous proof we can also deduce an iterative tree growing algorithm for the histories offering an alternative explanation for Theorem 4.1. Every internal node gets a label that is a pair consisting of its time slice and the number of its left-most leaf. Note that this uniquely identifies a node in the species tree.

We start with a root node labeled by the first time slice and an arbitrary number from $\{1, \ldots, k\}$. At every step, choose a leaf of the current history and consider the corresponding node in the events graph. Then traverse one of its edges and perform the action of this edge: If it is a speciation edge then add a new node with a label consisting of the successive time slice and the same number as only child. If it is a duplication or transfer edge then add a left child with the same label as the root and a right child labeled with the current time slice and an arbitrary number from the set the edge is pointing to. Once all leaves correspond to extant nodes the tree is a valid history.

Remark 4.2 The construction of the events graph in Theorem 4.1 can be adapted to all models. If there are no duplication events, the duplication edges are removed; if there are no HGT events, the transfer edges are removed. The characteristics of the $\mathbb{SL}$ dynamics are not encoded in the events graph but in the bijection or the history growing algorithm.

4.2.2 On the parsimony and profile of random histories.

We also considered at the distribution of the evolutionary score for randomly sampled histories, where the score of a history is the sum of the number of duplications, losses and HGT, for $k = 16$ and $n = 30$, over 50 random unranked species trees, sampling $10,000$ random histories for each species tree.

Figure 9 below suggests that the space of histories for a given species tree is dominated by histories with a relatively high score and that, as expected, for a given species tree including HGT in the evolutionary model leads to a significant decrease of the evolutionary score of histories.

In fact, when looking at the distribution of the number of duplications in the $\mathsf{u}\mathbb{DLT}$-model (results not shown), we observed that the duplication number drops significantly in the $\mathsf{u}\mathbb{DLT}$-model compared to the $\mathsf{u}\mathbb{DL}$-model. We can also note that, when comparing the score of histories in the $\mathsf{u}\mathbb{DL}$-model and the number of duplications, most of the score is due to gene losses (Figure 10), a characteristic we also see in the $\mathsf{u}\mathbb{DLT}$-model where the number of duplications (resp. HGT) exceeds rarely 5 (resp. 25) in the sampled histories.

5 Conclusion and perspectives

Our work introduces the first results on counting and sampling evolutionary scenarios in models accounting for gene duplication, gene loss and HGT. The originality of our work, compared to previous work in the reconciliation framework, is that we only consider the species tree to be given, and thus consider all possible evolutionary histories of a given size, i.e. leading to a given number of genes. Our results include formal grammars describing this combinatorial space, together with counting and sampling algorithms, obtained using either dynamic programming or enumerative and analytic combinatorics methods. These results complement a growing body of work developed over the last few years in the case of matching gene and species trees.

Using our method, we were able to obtain precise asymptotics on the number of histories for the two specific species trees, the rooted caterpillar and the complete binary tree in the unranked $\mathbb{DL}$-model, although our method also applies to any given species

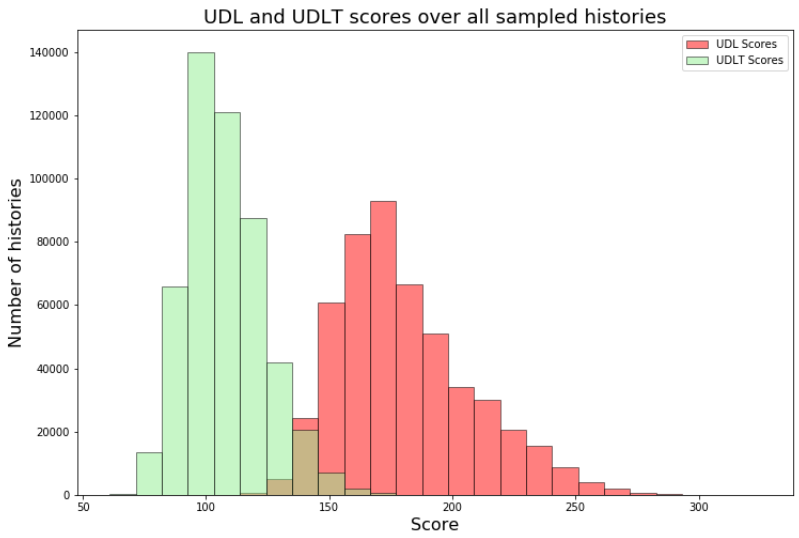

Fig. 9: Distribution of the score (number of duplications plus losses plus HGT) over 50 random species trees of size 16 and 10,000 random histories of size 30 per tree in the u$\mathbb{DL}$- and u$\mathbb{DLT}$-models.

tree in this model. Our counting and sampling algorithms allowed us to complement these results for other models, especially models accounting for HGT. Our experimental results provide a first global view of the space of potential evolutionary histories for a given species tree. They confirm the expected fact that introducing HGT in a model result in a dramatic increase of the space of possible histories; they also lead to the interesting observation that in the ranked $\mathbb{DLT}$-model, the total number of histories is asymptotically almost independent of the given species tree.

Our work suggests several avenues for further research. First, our notion of evolutionary history assumes that gene trees are ordered, i.e. that gene copies created by a gene duplication are distinguishable; this differs from the notion of reconciled gene trees, where duplicated copies are not distinguishable. While our assumption follows naturally from an evolutionary biology point of view, it would be interesting to see if our approach could be applied to count and sample reconciliations instead of histories. Next, the last few years have seen the development of more comprehensive models of gene family evolution, accounting for example for genes appearing at a given species by an HGT from an unsampled or extinct species [46], incomplete lineage sorting (ILS) [3, 21, 37, 42, 51, 52], or gene conversion [30]. In these models, reconciled gene trees can be computed using dynamic programming algorithms and it is natural to ask if such algorithms could be turned into grammars for the corresponding space of evolutionary scenarios. Last, from an applied point of view, a limitation of our work lies in the fact that histories are parameterized by their size, i.e. the number of extant genes, while in applications, the genes of a gene family are assigned to specific extant species. Ideally, in order to explore (through counting or sampling) the space of all possible

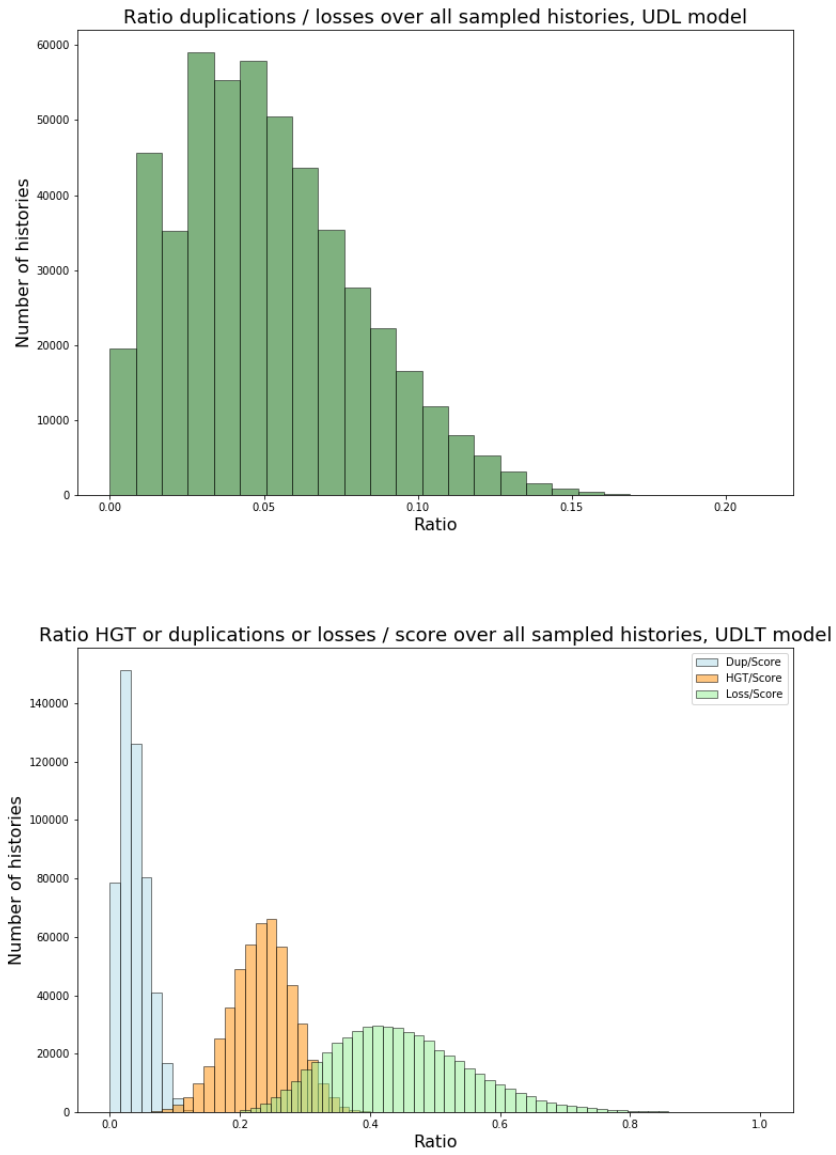

Fig. 10: Distribution of the ratio Duplications / Losses in the $\mathsf{u}\mathbb{DL}$ (Top) and of the ratios HGT / score, Duplications/score and Losses/score in the $\mathsf{u}\mathbb{DLT}$-model (Bottom). For both figures the distribution is over 50 random species trees of size 16 and $10,000$ random histories of size 30 per tree.

evolutionary scenarios for a gene families whose distribution of genes in extant species is given, we would need to parameterize our algorithms by this distribution, which leads to dynamic programming algorithms with a much higher time and space complexity, dependent on the number of extant species. However, we believe that advanced combinatorial sampling, especially multiparametric combinatorial samplers [7, 9], can be used within the framework we developed in the present work to provide efficient counting and sampling algorithms.

Funding: The first author is supported by a Discovery Grant of the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-03986). This research was enabled in part by support provided by Westgrid (https://www.westgrid.ca/) and Compute Canada (https://www.computecanada.ca) through a Resource Allocation (ID 838) to the first author. The third author was supported by the Exzellenzstipendium of the Austrian Federal Ministry of Education, Science and Research and the Erwin Schrödinger Fellowship of the Austrian Science Fund (FWF): J 4162-N35.

Conflict of interest: The authors declare that they have no conflict of interest.

References

1. Ö. Åkerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–5719, 2009.
2. L. Arvestad, J. Lagergren, and B. Sennblad. The gene evolution model and computing its associated probabilities. *Journal of the ACM*, 56(2):7:1–7:44, 2009.
3. Y. ban Chan, V. Ranwez, and C. Scornavacca. Inferring incomplete lineage sorting, duplications, transfers and losses with reconciliations. *Journal of Theoretical Biology*, 432:1–13, 2017.
4. C. Banderier and M. Wallner. Lattice paths with catastrophes. *Discrete Mathematics & Theoretical Computer Science*, Vol 19 no. 1, Sept. 2017. Full version of extended abstract with the same title appeared in the Proceedings of conference on Random Generation of Combinatorial Structures – {GASCom} 2016.
5. M. S. Bansal, E. J. Alm, and M. Kellis. Reconciliation revisited: Handling multiple optima when reconciling with duplication, transfer, and loss. *Journal of Computational Biology*, 20(10):738–754, 2013.
6. M. S. Bansal, M. Kellis, M. Kordi, and S. Kundu. Ranger-dtl 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, 34(18):3214–3216, 2018.
7. M. Bendkowski, O. Bodini, and S. Dovgal. Polynomial tuning of multiparametric combinatorial samplers. In *Proceedings of the Fifteenth Workshop on Analytic Algorithmics and Combinatorics, ANALCO 2018, New Orleans, LA, USA, January 8-9, 2018.*, pages 92–106. SIAM, 2018.
8. O. Bodini, D. Gardy, B. Gittenberger, and Z. Gołębiewski. On the number of unary-binary tree-like structures with restrictions on the unary height. *Annals of Combinatorics*, 22(1):45–91, 2018.
9. O. Bodini and Y. Ponty. Multi-dimensional Boltzmann Sampling of Languages. In M. Drmota and B. Gittenberger, editors, *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, volume DMTCS Proceedings vol. AM, 21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10) of *DMTCS Proceedings*, pages 49–64, Vienna, Austria, June 2010. Discrete Mathematics and Theoretical Computer Science.
10. M. Bóna and P. Flajolet. Isomorphism and symmetries in random phylogenetic trees. *Journal of Applied Probability*, 46(4):1005–1019, 2009.

11. J. Degnan and N. Rosenberg. Gene tree discordance, phylogenetic and the multispecies coalescent. *Trends in Ecology & Evolution*, 24:332–340, 2009.
12. J. H. Degnan and L. A. Salter. Gene tree distribution under the coalescent process. *Evolution*, 59(1):24–37, 2005.
13. F. Disanto and N. A. Rosenberg. Coalescent histories for lodgepole species trees. *Journal of Computational Biology*, 22(10):918–929, 2015.
14. F. Disanto and N. A. Rosenberg. Asymptotic properties of the number of matching coalescent histories for caterpillar-like families of species trees. *IEEE/ACM Transactions on Compututational Biology and Bioinformatics*, 13(5):913–925, 2016.
15. F. Disanto and N. A. Rosenberg. Enumeration of ancestral configurations for matching gene trees and species trees. *Journal of Computational Biology*, 24(9):831–850, 2017.
16. F. Disanto and N. A. Rosenberg. On the number of non-equivalent ancestral configurations for matching gene trees and species trees. *Bulletin of Mathematical Biology*, 2017. in press.
17. F. Disanto and N. A. Rosenberg. Enumeration of compact coalescent histories for matching gene trees and species trees. *Journal of Mathematical Biology*, 2018. in press.
18. J.-P. Doyon, C. Chauve, and S. Hamel. Space of gene/species trees reconciliations and parsimonious models. *Journal of Computational Biology*, 16(10):1399–1418, 2009.
19. J.-P. Doyon, V. Ranwez, V. Daubin, and V. Berry. Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics*, 12(5):392–400, 2011.
20. M. Drmota. Systems of functional equations. *Random Structures & Algorithms*, 10(1-2):103–124, 1997.
21. P. Du and L. Nakhleh. Species tree and reconciliation estimation under a duplication-loss-coalescence model. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '18, pages 376–385. ACM, 2018.
22. D. Durand, B. V. Halldórsson, and B. Vernot. A hybrid micro–macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13(2):320–335, 2006.
23. P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, 2009.
24. P. Flajolet, P. Zimmermann, and B. Van Cutsem. A calculus for the random generation of labelled combinatorial structures. *Theoretical Computer Science*, 132(1-2):1–35, 1994.
25. B. Gittenberger, E. Y. Jin, and M. Wallner. On the shape of random Pólya structures. *Discrete Mathematics*, 341(4):896 – 911, 2018.
26. M. Goodman, J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28(2):132–163, 1979.
27. P. Górecki, G. J. Burleigh, and O. Eulenstein. Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics*, 12(1):S15, 2011.
28. P. Górecki and J. Tiuryn. DLS-trees: A model of evolutionary scenarios. *Theoretical Computer Science*, 359(1):378–399, 2006.
29. P. Górecki and O. Eulenstein. Drml: Probabilistic modeling of gene duplications. *Journal of Computational Biology*, 21(1):89–98, 2014.
30. D. Hasić and E. Tannier. Gene tree species tree reconciliation with gene conversion. *Journal of Mathematical Biology*, 78(6):1981–2014, 2019.
31. E. Jacox, C. Chauve, G. J. Szöllősi, Y. Ponty, and C. Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 32(13):2056–2058, 2016.
32. W. P. Maddison. Gene trees in species trees. *Systematic Biology*, 46(3):523–536, 1997.
33. A. Nijenhuis and H. S. Wilf. *Combinatorial Algorithms*. Academic Press, 1978.
34. Y. Ovadia, D. Fielder, C. Conow, and R. Libeskind-Hadas. The cophylogeny reconstruction problem is NP-complete. *Journal of Computational Biology*, 18(1):59–65, 2011.
35. J. Pei and Y. Wu. STELLS2: fast and accurate coalescent-based maximum likelihood inference of species trees from gene tree topologies. *Bioinformatics*, 33(12):1789–1797, 2017.
36. V. Ranwez, C. Scornavacca, J.-P. Doyon, and V. Berry. Inferring gene duplications, transfers and losses can be done in a discrete framework. *Journal of Mathematical Biology*, 72(7):1811–1844, 2016.
37. M. D. Rasmussen and M. Kellis. Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Research*, 22(4):755–765, 2012.
38. N. A. Rosenberg. Counting coalescent histories. *Journal of Computational Biology*, 14(3):360–377, 2007.

39. C. Scornavacca, E. Jacox, and G. J. Szöllősi. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, 31(6):841–848, 2015.
40. J. Sjöstrand, A. Tofigh, V. Daubin, L. Arvestad, B. Sennblad, and J. Lagergren. A bayesian method for analyzing lateral gene transfer. *Systematic Biology*, 63(3):409–420, 2014.
41. N. J. A. Sloane. The On-Line Encyclopedia of Integer Sequences (OEIS). http://oeis.org.
42. M. Stolzer, H. Lai, M. Xu, D. Sathaye, B. Vernot, and D. Durand. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18):i409–i415, 2012.
43. G. J. Szöllősi and V. Daubin. Modeling gene family evolution and reconciling phylogenetic discord. In M. Anisimova, editor, *Evolutionary Genomics: Statistical and Computational Methods, Volume 2*, volume 856 of *Methods in Molecular Biology*, pages 29–51. Humana Press, 2012.
44. G. J. Szöllősi, W. Rosikiewicz, B. Boussau, E. Tannier, and V. Daubin. Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6):901–912, 2013.
45. G. J. Szöllősi, E. Tannier, V. Daubin, and B. Boussau. The inference of gene trees with species trees. *Systematic Biology*, 64(1):e42–e62, 2015.
46. G. J. Szöllősi, E. Tannier, N. Lartillot, and V. Daubin. Lateral gene transfer from the dead. *Systematic Biology*, 62(3):386–397, 2013.
47. A. Tofigh, M. T. Hallett, and J. Lagergren. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):517–535, 2011.
48. H. S. Wilf. A unified setting for sequencing, ranking, and selection algorithms for combinatorial objects. *Advances in Mathematics*, 24(3):281–291, 1977.
49. Y. Wu. Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, 66(3):763–775, 2012.
50. Y. Wu. An algorithm for computing the gene tree probability under the multispecies coalescent and its application in the inference of population tree. *Bioinformatics*, 32(12):i225–i233, 2016.
51. Y.-C. Wu, M. D. Rasmussen, and M. Kellis. Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Research*, 24(3):475–486, 2014.
52. B. Zhang and Y.-C. Wu. Coestimation of gene trees and reconciliations under a duplication-loss-coalescence model. In Z. Cai, O. Daescu, and M. Li, editors, *Bioinformatics Research and Applications*, volume 10330 of *Lecture Notes in Computer Science*, pages 196–210. Springer, 2017.