Gain-loss-duplication models for copy number evolution on a phylogeny: Exact algorithms for computing the likelihood and its gradient

Miklós Csűrös

Department of Computer Science and Operations Research, Université de Montréal, C.P. 6128 succursale Centre-Ville, Montréal, Québec H3C 3J7, Canada

ARTICLE INFO

ABSTRACT

Gene gain-loss-duplication models are commonly based on continuous-time birth–death processes. Employed in a phylogenetic context, such models have been increasingly popular in studies of gene content evolution across multiple genomes. While the applications are becoming more varied and demanding, bioinformatics methods for probabilistic inference on copy numbers (or integer-valued evolutionary characters, in general) are scarce.

We describe a flexible probabilistic framework for phylogenetic gain-loss-duplication models. The framework is based on a novel elementary representation by dependent random variables with well-characterized conditional distributions: binomial, Pólya (negative binomial), and Poisson.

The corresponding graphical model yields exact numerical procedures for computing the likelihood and the posterior distribution of ancestral copy numbers. The resulting algorithms take quadratic time in the total number of copies. In addition, we show how the likelihood gradient can be computed by a linear-time algorithm.

© 2022 Published by Elsevier Inc.

1. Introduction

Homology between two genes is the equivalence relation of descent from the same ancestral gene (Fitch, 2000), defining the corresponding equivalence classes of gene *families*. During genome annotation, homologies are routinely recognized by sequence similarity, and annotated genes are assigned to families (Aziz et al., 2008). The *copy number* for a family is the number of family representatives in the genome, a non-negative integer. The *profile* of a family comprises the copy numbers across different genomes. Family profiles are used in evolutionary genomics to infer ancestral gene content (Csűrös and Miklós, 2009), and in functional genomics to recognize associations between families (Dey and Meyer, 2015).

Probabilistic approaches to copy number evolution are based on continuous-time birth–death processes (Novozhilov et al., 2006; Nye, 2009). Such processes are also fundamental in queuing theory (Takács, 1962), epidemiology and population growth models (Kendall, 1949). In the context of genome evolution, the process captures fixation events eliminating a gene ("death") by segmental loss and pseudogenization, or adding a gene ("birth") to the genome, either by duplication within the same genome, or by lateral transfer from another genome. Birth-death events

are used also to model speciation-extinction in phylogenies (Nee et al., 1994; Stadler, 2009) and infection-recovery rates in epidemics (Tanaka et al., 2006). The evolution of various integer-valued traits can be modeled by birth–death processes with additional events: polyploidization for chromosome number evolution (Mayrose et al., 2010), locus shifting for transposons (Rosenberg et al., 2003; Xu et al., 2015), or point mutations for microsatellite repeats (Kruglyak et al., 1998).

This work focuses on branching *linear* birth–death processes. In particular, we assume the classic birth, death, and immigration (BDI) process (Karlin and McGregor, 1958; Tavaré, 1989), where "immigration" and "birth" for a multigene family are lateral transfer and duplication events, respectively. The phylogeny encodes the branching pattern of the process: child lineages evolve independently by BDI processes from the same set of ancestor copies. The algorithmics of phylogenetic birth–death models is difficult mainly because of the intricate transition probabilities in birth–death processes without known closed expressions. Non-linear birth–death models can be analyzed for their asymptotic behavior (Karev et al., 2004). The transient probabilities can be approximated (Crawford et al., 2014) when the process is sampled at discrete time intervals. But even the simplest non-linear model, the M/M/1 queue (constant birth and death rates Iwasaki and Takagi, 2007) lacks a closed-form solution (Leguesdron et al., 1993).

E-mail address: csuros@iro.umontreal.ca.

Bioinformatics problems for copy numbers differ fundamentally from molecular sequence evolution problems with a finite character set, and porting standard methods to an unbounded character domain is generally not possible. A simple, but unsatisfactory fix is to impose a limit on maximum copy number, and usual techniques like Felsenstein's peeling method (Felsenstein, 1973) apply. Indeed, pioneer applications of birth–death processes for gene content evolution by Hahn et al. (2005) and by Iwasaki and Takagi (2007) employ the same workaround. Setting a maximum is useful for non-linear birth–death models, as transition probabilities can be computed by (finite) exponentiation from an arbitrary rate matrix (Iwasaki and Takagi, 2007; Spencer et al., 2006; Ames et al., 2012; Fukunaga and Iwasaki, 2021). The first algorithm for computing the profile likelihood for unbounded copy numbers with a gain-loss-duplication model (Csűrös and Miklós, 2006) is based on an inclusion-exclusion formula. A numerically more stable algorithm was developed subsequently (Csűrös and Miklós, 2009), and the corresponding methods are implemented in the software package Count (Csűrös, 2010), which has been used in hundreds of studies.

We introduce a fresh mathematical framework for linear gain-loss-duplication (and gain-loss without duplication, or duplication-loss without gain) models on a phylogeny. The novel formulation is based on Theorem 2 giving the transition probabilities in a closed form that involves only basic discrete distributions. The theorem suggests a fundamental dependency network of random variables along the phylogeny, representing ancestral copy numbers and inheritance (survival) in child lineages. The elementary decomposition yields relatively simple algorithms to compute the likelihood of a family profile (Theorem 5). The network factorization also leads to an algorithm for posterior probabilities of ancestral copy numbers (Theorem 7 and Corollary 8), and to our main result, a hitherto elusive algorithm for computing the gradient of the log-likelihood with respect to model parameters (Corollary 11 and Theorem 12). The gradient and likelihood calculations take quadratic time in the total number of observed copies (Theorem 15), making them pragmatic even for large phylogenies.

2. Theory

A *phylogeny* is a rooted binary tree with nodes numbered $u \in [R] = \{1, 2, \ldots, R\}$. Every node either has two non-null child nodes, or is a terminal node (a *leaf*) with two null children. For ease of notation, we assume that the nodes are indexed respecting postfix order, with every child's index being less than the parent's, so that the last one is the root. The tree is identified by its root R and its edges $T \subset [R] \times [R]$ directed from parent to child. The edges in the subtree rooted at a node u are denoted by T_u, including $T_R = T$. The set of leaves is denoted by $\mathcal{L}$, and the leaf set for T_u by $\mathcal{L}_u$; in particular, $\mathcal{L}_R = \mathcal{L}$. For simplicity, start the indices with the leaves respecting the postfix order, so that $\mathcal{L} = [L]$ and every subset $\mathcal{L}_u$ comprises consecutive integers.

Consider the problem of *copy number evolution*: each node u has an associated random variable ξ_u, called the copy number, taking non-negative integer values, and the joint distribution is determined by dependencies along the phylogeny:

$$\mathbb{P}\{\xi_1 = n_1, \ldots, \xi_R = n_t\} = \mathbb{P}\{\xi_R = n_R\} \prod_{uv \in T} \underbrace{\mathbb{P}\Big\{\xi_v = n_v \mid \xi_u = n_u\Big\}}_{\text{transition on edge } uv}.$$

$$(1)$$

The leaf variables are observable, corresponding to extant species, forming the *profile* $\Xi = \{\xi_v\}_{v \in \mathcal{L}}$. Non-leaf nodes are (hypothetical) ancestors with unobserved copy numbers. The *ancestral inference* problem is that of estimating $\{\xi_u\}_{u \notin \mathcal{L}}$ for ancestral
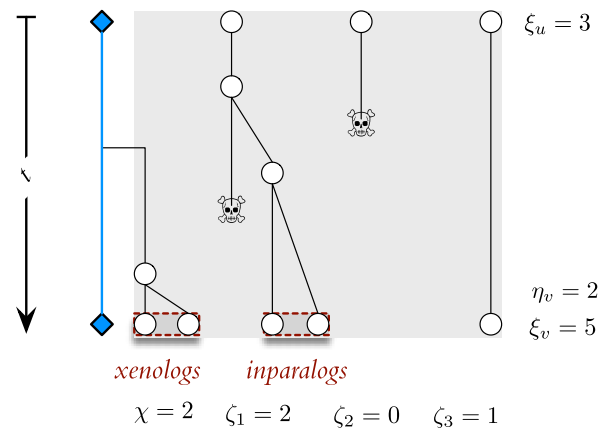
Fig. 1. Example of copy number evolution on an edge uv. Parental copies 1 and 3 *survive* in the child, with $\zeta_1 = 2$ and $\zeta_3 = 1$ inparalogs resulting from the histories of duplications and losses. In contrast, parental gene 2 is lost ($\zeta_2 = 0$), so the number of surviving copies is $\eta_v = 2$. A copy is gained from an external source (diamond) and subsequently duplicates to result in $\chi = 2$ xenolog copies.

nodes, knowing the distribution of Eq. (1) and Ξ. Suppose that we observe the leaf variables across sample profiles called *families* $f = 1, \ldots, F$, with independent and identically distributed (*iid*) copy number vectors $(\xi_{1,1}, \ldots, \xi_{1,R}), \ldots, (\xi_{F,1}, \ldots, \xi_{F,R})$. The *model inference* problem is that of deducing the distribution of (1) from an iid sample $(\Xi_1, \ldots, \Xi_F)$.

2.1. Copy number evolution with gain, loss and duplication

For every family, the copies evolve independently on each edge uv via gene duplications and losses, as well as non-duplication *gains* (e.g. lateral transfer from another species). In particular, the child's copy number ξ_v for $n = \xi_u$ parental copies is the sum of the random variables χ and $\zeta_1, \ldots, \zeta_n$, where every ζ_i is the number of *inparalog* child copies of the same parent gene i, and χ is the number of *xenolog* copies without an ancestor at the parent:

$$\xi_v = \chi + \sum_{i=1}^{\xi_u} \zeta_i.$$

$$(2)$$

See Fig. 1 for an illustration. Our terminology for xenologs and inparalogs follows classic definitions of different homology types (Fitch, 2000; Sonnhammer and Koonin, 2002). *Out*paralogs, as opposed to inparalogs, are child copies that descend from different parent copies. Inparalogs are indistinguishable in the model, and therefore *orthology* between gene copies at different nodes is not considered here. Xenolog histories involve an interspecies transfer of genetic material, modeled by imagining a virtual, external gene that produces copies within the genome. In the copy number evolution model, the source of the gain is immaterial, so innovation or *de novo* gene birth also creates xenologs.

We assume a linear model in Eq (2), where χ and all ζ_i are independent, and the inparalogs ζ_i are identically distributed. If $\zeta_i = 0$ for some i, the parental copy i is *lost*, otherwise it *survives* in v. A continuous-time model is obtained by imposing a birth–death process $\xi(t)$ on every edge $uv \in T$ during some time $t_{uv} \geq 0$ (the *edge length*), so that $\xi_u = \xi(0)$ and $\xi_v = \xi(t_{uv})$. In the linear setting of Eq. (2),

$$\xi(t) = \chi(t) + \sum_{i=1}^{\xi_u} \zeta_i(t),$$

with $\chi(t)$ counting the xenolog copies, and $\zeta_i(t)$ counting inparalogs descending from the same parent gene i. The xenolog and

inparalog processes create a set of Galton–Watson trees over the copies with time-annotated nodes (Tavaré, 1989). Loss events create terminal nodes and duplication events create bifurcations (see Fig. 1). Our interest lies not in inferring the genealogies (i.e., in *reconciling* the gene histories with the species phylogeny), but rather in designing a model for the copy numbers without explaining where each copy originates.

The inparalog process is characterized by the constant instantaneous rates for loss $\mu > 0$ (per copy) and duplication $\lambda > 0$ (per copy). The xenolog process has the same per-copy rates, but also includes birth by gain with some rate $\kappa\lambda > 0$. Accordingly, for the total copy number $\xi(t)$, $n \to (n-1)$ death events arrive with a rate of μn, and $n \to (n+1)$ birth events arrive with a rate of $\lambda(n+\kappa)$. In particular, for $P_n(t) = \mathbb{P}\{\xi(t) = n\}$, the Kolmogorov backward equations are

$$P_0'(t) = \mu P_1(t) - \lambda\kappa P_0(t)$$

and for $0 < n$,

$$P_n'(t) = \lambda(n-1+\kappa)P_{n-1}(t) + \mu(n+1)P_{n+1}(t) - (\lambda(n+\kappa) + \mu n)P_n(t)$$

with $P_n'(t) = \frac{\partial P_n(t)}{\partial t}$.

The *no-duplication model* is the limit for $\kappa\lambda \to \gamma\mu$ while the duplication rate $\lambda \to 0$:

$$P_0'(t) = \mu P_1(t) - \mu\gamma P_0(t)$$

and, for $0 < n$

$$P_n'(t) = \mu\gamma P_{n-1}(t) + \mu(n+1)P_{n+1}(t) - \mu(\gamma + n)P_n(t)$$

which is characterized by loss rate μ and the relative gain rate γ. In the no-duplication model, birth events arrive at the same rate, independently from the number of existing copies, while each copy is lost independently with the same rate.

The parametrization by *relative* gain rates κ or γ with respect to λ or μ is mathematically convenient because they show up directly in the formulas for transition probabilities. In a biological interpretation, the components μ and λ are the per-copy instantaneous rates of loss and duplication, which can be embedded in a population-genetic model of genome size evolution, so that they are determined by a selection coefficient, and the (constant) population size (Sela et al., 2016). If selection, and, consequently, fixation probabilities, for gained xenolog copies are identical to inparalog copies, then $0 < \kappa$ represents the relative rate of gene birth by non-duplication processes (including horizontal gene transfer and innovation), and gene birth by duplication (including segmental duplication and retrotranscription). The particular case $\kappa = 1$ is the process of gene length evolution in the Thorne-Kishino-Felsenstein model (Thorne et al., 1991) for molecular sequences, where insertions by the *immortal link* play the same role as xenolog copies here, and μ, λ are the deletion and insertion rates per residue.

2.2. Transient probabilities for linear birth–death processes

The basic transition probabilities for the xenolog and inparalog processes are well understood (Kendall, 1949; Karlin and McGregor, 1958). Denote the transient probabilities by

$$h_n(t) = \mathbb{P}\{\chi(t) = n\} \quad \text{and} \quad g_n(t) = \mathbb{P}\{\zeta(t) = n\}$$

with starting values $\chi(0) = 0$ and $\zeta(0) = 1$. Then $\chi(t)$ has either Pólya distribution (the generalized version of negative binomial, allowing for non-integer κ parameter) when $\lambda > 0$, or Poisson distribution when $\lambda = 0$. The inparalogs $\zeta(t)$ follow a shifted geometric distribution.

$$h_n(t) = \binom{\kappa + n - 1}{n}(1-q)^\kappa q^n \qquad \text{if } \lambda, \kappa > 0 \quad (3a)$$

$$= \begin{cases} (1-q)^\kappa & \{n = 0\} \\ \frac{\kappa(\kappa+1)\cdots(\kappa+n-1)}{n!}(1-q)^\kappa q^n & \{n > 0\} \end{cases}$$

$$h_n(t) = e^{-r}\frac{r^n}{n!} \qquad \text{if } \lambda = 0 \quad (3b)$$

$$g_n(t) = \begin{cases} p & \{n = 0\} \\ (1-p)(1-q)q^{n-1} & \{n > 0\} \end{cases} \qquad (3c)$$

with the parameters

$$p = \frac{\mu - \mu e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} \qquad (4a)$$

$$q = \frac{\lambda - \lambda e^{-(\mu-\lambda)t}}{\mu - \lambda e^{-(\mu-\lambda)t}} \quad \text{if } \lambda > 0 \qquad r = \gamma(1 - e^{-\mu t}) \quad \text{if } \lambda = 0$$

$$(4b)$$

assuming $\lambda \neq \mu$; or if $\lambda = \mu$,

$$p = q = \frac{\mu t}{1 + \mu t}. \qquad (4c)$$

Note that the formulas remain valid for all transient probabilities ($t < \infty$), including when $\lambda > \mu$, even if the birth–death process has a stationary distribution only when $\lambda \leq \mu$.

The rates and the edge length can be rescaled simultaneously without affecting the distributions. Dissecting into scale-independent parameters (assuming $q \neq p$):

$$p = \frac{1 - e^{-\delta(\mu t)}}{1 - (1-\delta)e^{-\delta(\mu t)}} \qquad 1 - p = \frac{\delta e^{-\delta(\mu t)}}{1 - (1-\delta)e^{-\delta(\mu t)}}$$

$$q = \frac{(1-\delta)\left(1 - e^{-\delta(\mu t)}\right)}{1 - (1-\delta)e^{-\delta(\mu t)}} \qquad 1 - q = \frac{\delta}{1 - (1-\delta)e^{-\delta(\mu t)}}$$

with $\delta = 1 - \lambda/\mu = 1 - q/p$. The formulas are invertible: for a given $0 < p, q < 1$ we can find δ and the scaled edge length (μt). Therefore, copy number evolution on every edge is fully determined by the *distribution parameters* (p, q, κ) or (p, r), because they define the birth–death process parameters.

Theorem 1 (*Unicity of Distribution Parameters*). *Let $0 < t$ be fixed. For any given $0 < p, q < 1$ and $0 < \kappa$, or with $q = 0$, for any given $0 < p < 1$ and $0 < r$, there exist rate settings $0 < \mu, 0 \leq \lambda$ that yield those distribution parameters as in Eq. (4).*

Proof. If $q = 0$, then $\lambda = 0$, and by $p = 1 - e^{-\mu t}$ and $r = \gamma p$, we can set $\mu t = -\ln(1-p)$ and $\gamma = r/p$ to match p and r. If $0 < q = p$, then set $(\mu t) = p/(1-p)$ and $\lambda = \mu$. Otherwise, since $q/p = 1 - \delta$ and $(1-q)/(1-p) = e^{\delta\mu t}$, set $\delta = 1 - \frac{q}{p}$, $(\mu t) = \ln\frac{1-q}{1-p}/(1 - q/p)$ and $\lambda = \mu(1 - \delta)$. $\square$

3. Results and discussion

Based on the transient probabilities for xenolog and inparalog copy numbers (starting with 0 or 1 copies), we infer the probabilities for copy number changes in the general case (starting with n copies) in Section 3.1. In Section 3.2, we discuss the likelihood model for copy numbers on a phylogeny using two random variables per node: the copy number ξ at the node and the number of non-empty inparalog groups η in the lineage leading to the node. The model is parametrized by a triple of (κ, p, q) gain-loss-duplication parameters on each edge; or by the gain-loss parameter couple (r, p) in the no-duplication model. The infinite summations in the likelihood formulas of (10) and (11) are exploited in Section 3.3 to infer the probability of unobserved gene families with 0 copies at all leaves. Theorem 4 not only gives the probability of the empty profile, but also

establishes the (per-copy) *survival parameters* $\tilde{p}, \tilde{q}$ and the (per-copy) *extinction probability* ϵ that are used in further calculations. A finite likelihood computation for arbitrary copy number profiles at the leaves is introduced in Section 3.4 that uses conditioning on *ancestral* copy numbers (random variables $\tilde{\xi}_u, \tilde{\eta}_u$ at every node u): genes with descendants in at least one terminal lineage. In particular, Theorem 5 gives the postorder recurrences for computing the likelihood.

In Section 3.5, the postorder likelihood computation is coupled by a preorder recurrence on the phylogeny for complementary *outside* likelihoods in Theorem 7, which are plugged into the formulas for posterior probabilities of ancestral copy numbers in Corollary 8. In order to lay the foundations for the optimization of model parameters by maximizing the likelihood, we describe a numerical procedure for computing the partial derivatives of the likelihood in Section 3.6. The gradient with respect to survival parameters $(\tilde{p}, \tilde{q})$ can be obtained directly from the posterior expected copy numbers (Theorem 10 and Corollary 11). Theorem 12 describes an algorithm for computing the gradient with respect to the original model parameters p, q in a preorder traversal of the phylogeny. The numerical algorithms for the no-duplication model are described in Section 3.7. Our concluding result in Section 3.8 shows that the algorithms take quadratic time for computing the likelihood, the posteriors and the gradient. Proofs for the major theorems are in the Appendix. Table 1 summarizes the important notations in our discussion.

3.1. Transient probabilities in the general case

First, suppose that duplications are allowed, and $\lambda_v > 0$ on all edges $uv \in T$. If there are $\xi_u = n$ copies at an ancestral node u, then they evolve independently along each child edge uv:

$$\xi_v = \chi + \zeta_1 + \cdots + \zeta_n \tag{5}$$

where χ denotes the xenolog copies, and ζ_i denote iid variables for the descendant inparalog copies from each parental instance $i = 1, \ldots, n$. The ζ_i variables follow the basic transition probabilities from Eq. (3)

$$\mathbb{P}\{\chi = k\} = h_k(t_{uv}) \quad \text{and} \quad \mathbb{P}\{\zeta_i = k\} = g_k(t_{uv}).$$

The key observation for calculating $\mathbb{P}\left\{\xi_v = m \mid \xi_u = n\right\} = \mathbb{P}\{\chi + \zeta_1 + \cdots + \zeta_n = m\}$ is that $\zeta_i | \zeta_i > 0$ has the same geometric tail as the Pólya distribution of χ. Since the distributions with the same tail parameter q can be summed directly, $\xi_v - s$ has a Pólya distribution with parameter $(\kappa + s)$, where $s = \sum_{i=1}^{n}\{\zeta_i > 0\}$ is the number of surviving copies. (The shorthand notation $\{\zeta_i > 0\}$ denotes indicator variable that takes the value 1 whenever ζ_i is positive, and the value 0 when $\zeta_i = 0$.)

Theorem 2 (*Transient Probabilities in the General Case*). *For a linear birth–death process with parameters $\kappa, \lambda, \mu > 0$,*

$$\mathbb{P}\left\{\xi(t) = m \mid \xi(0) = n\right\}$$
$$= \sum_{s=0}^{\min\{n,m\}} \binom{\kappa + m - 1}{m - s}(1 - q)^{\kappa + s} q^{m - s} \binom{n}{s} p^{n - s}(1 - p)^s \tag{6}$$

with the parameters p, q defined in Eqs. (4).

For a linear birth–death process with parameters $\lambda = 0$ and $\mu, \gamma > 0$,

$$\mathbb{P}\left\{\xi(t) = m \mid \xi(0) = n\right\} = \sum_{s=0}^{\min\{n,m\}} e^{-r}\frac{r^{m-s}}{(m - s)!}\binom{n}{s}(1 - p)^s p^{n - s} \tag{7}$$

with the parameters p, r defined in Eqs. (4).

The original recursive algorithm of Csűrös and Miklós (2009) for computing the profile likelihood uses the basic birth–death transitions from (3), and arrives at a set of recurrences by combinatorial principles. We can infer the same method algebraically in the present framework — by extracting the recurrences for transition probabilities from Theorem 2.

Corollary 3 (*Transition Probability Recurrences*). *Let $uv \in T$ be any edge and $w(m \mid n) = \mathbb{P}\left\{\xi_v = m \mid \xi_u = n\right\}$ denote the transition probabilities.*

For $\lambda_v > 0$, let $p = p_v, q = q_v, \kappa = \kappa_v$ denote the applicable distribution parameters from Eq. (4). Then

$$w(m \mid 0) = \binom{\kappa + m - 1}{m}(1 - q)^\kappa q^m$$

and, for $0 < n$,

$$w(m \mid n) = q\, w(m \mid n - 1)$$
$$+ \{m > 0\}(1 - p - q)w(m - 1 \mid n - 1)$$
$$+ \{m > 0\}q\, w(m - 1 \mid n).$$

3.2. Phylogenetic model with surviving copies

We amend the phylogenetic model by explicitly inserting a hidden random variable η_v of *surviving* parental copies between the copy numbers ξ_u and ξ_v on every edge uv. In particular, with the xenolog-inparalog decomposition of Eq. (5),

$$\xi_v = \chi + \sum_{i=1}^{\xi_u} \zeta_i \quad \text{and} \quad \eta_v = \sum_{i=1}^{\xi_u}\{0 < \zeta_i\}.$$

Fig. 2 shows the example of evolving copy numbers ξ, η on a small phylogeny (on the path to leaf 3). For the ease of presentation, we continue with $\lambda_v > 0$ at every node v, and return to the no-duplication model afterwards. Using Theorem 2,

$$\mathbb{P}\left\{\eta_v = s \mid \xi_u = n\right\}$$
$$= \binom{n}{s}(1 - p_v)^s(p_v)^{n-s} \qquad \{s \le n\} \tag{8a}$$

$$\mathbb{P}\left\{\xi_v = m \mid \eta_v = s\right\}$$
$$= \binom{\kappa_v + m - 1}{m - s}(1 - q_v)^{\kappa_v + s}(q_v)^{m - s} \qquad \{s \le m\} \tag{8b}$$

with edge-specific loss, duplication, and gain parameters p_v, q_v, κ_v.

A complete history fixes all counts ξ_u and η_u: $\{\xi_1 = n_1, \ldots, \xi_R = n_R, \eta_1 = s_1, \ldots, \eta_{R-1} = s_{R-1}\}$. The joint distribution of our phylogenetically linked random variables is written explicitly as

$$\mathbb{P}\{\xi_1 = n_1, \ldots, \xi_R = n_R, \eta_1 = s_1, \ldots, \eta_{R-1} = s_{R-1}\}$$
$$= \mathbb{P}\{\xi_R = n_R\} \times \prod_{uv \in T}\left(\underbrace{\binom{n_u}{s_v}(1 - p_v)^{s_v}(p_v)^{n_u - s_v}}_{\mathbb{P}\left\{\eta_v = s_v \mid \xi_u = n_u\right\}}\right. \tag{9}$$
$$\left.\times \underbrace{\binom{\kappa_v + n_v - 1}{n_v - s_v}(1 - q_v)^{\kappa_v + s_v}(q_v)^{n_v - s_v}}_{\mathbb{P}\left\{\xi_v = n_v \mid \eta_v = s_v\right\}}\right),$$

All histories satisfying $s_v \le \min\{n_u, n_v\}$ on every edge $uv \in T$ and $\mathbb{P}\{\xi_R = n_R\} \ne 0$ have positive probability if p_u, q_u are bounded away from 0 and 1.

Let $\Xi = \{n_v : v \in \mathcal{L}\}$ be a profile comprising the observed copy numbers. The *profile likelihood* is the sum of all history
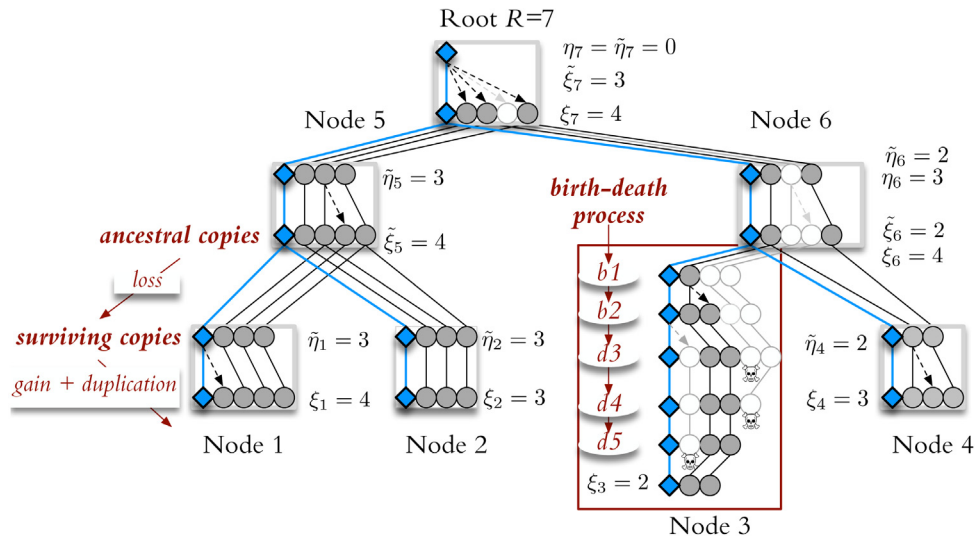
Fig. 2. Gene family evolution on a 4-leaf full binary phylogeny. Circles denote copies with lines connecting homologs among them. The blue diamonds mark a "virtual" copy outside the genome that is the progenitor of *xenolog* copies. The random variables ξ and η count all copies and surviving copies from the parent, respectively On the edge to Node 3 (a leaf), the birth–death process is illustrated with 5 events at random time intervals: duplication (b1, originating from the first inherited copy) and gain (b2, originating from outside), and losses (d3, d4, d5). The *ancestral* copy numbers $\tilde{\xi}$ and $\tilde{\eta}$ count only the copies that have homologs in at least one descendant (shaded circles). The full birth–death history includes the copies that are extinct in all descendants (white circles, shown along the path to leaf 3). An equivalent probability distribution is generated by a two-step manufacturing of ancestral copies that have descendants at the leaves: children inherit surviving copies by asymmetric loss, and surviving copies generate duplicates.

probabilities from (9) for the same profile:

$$L(\varXi)$$

$$= \mathbb{P}\{\varXi\}$$

$$= \sum_{n_u,s_u : u \notin \mathcal{L}} \mathbb{P}\{\xi_1 = n_1, \ldots, \xi_R = n_R, \eta_1 = s_1, \ldots, \eta_{R-1} = s_{R-1}\},$$

with infinitely many terms. Define the *partial profile* within every subtree as $\varXi_u = \{\forall v \in \mathcal{L}_u : \xi_v = n_v\}$ where $\mathcal{L}_u$ denotes the leaves in the subtree rooted at u, including the singleton $\mathcal{L}_u = \{u\}$ whenever u is a leaf. Define the likelihood of the partial profiles conditioned on ξ_u or η_u:

$$C_u(n) = \mathbb{P}\Big\{\varXi_u \mid \xi_u = n\Big\} \quad \text{and} \quad K_u(s) = \mathbb{P}\Big\{\varXi_u \mid \eta_u = s\Big\}.$$

At a leaf u, we have $C_u(n) = 1$ if $n = n_u$, the observed count, or $C_u(n) = 0$ if $n \neq n_u$. All other conditional likelihoods can be expressed using Eqs. (8a) and (8b) about the conditional distributions $\xi_u \mid \eta_u$ and $\eta_v \mid \xi_v$. At all nodes u,

$$K_u(s) = \sum_{k=0}^{\infty} \binom{\kappa_u + s + (k-1)}{k}(1-q_u)^{\kappa_u+s}(q_u)^k \times C_u(s+k); \quad (10a)$$

and at every ancestral node u,

$$C_u(n) = \prod_{uv \in T}\left(\sum_{s=0}^{n}\binom{n}{s}(1-p_v)^s(p_v)^{n-s} \times K_v(s)\right). \quad (10b)$$

The family distribution at the root R is needed to sum across the likelihoods $C_R(n)$ to get the profile likelihood

$$L(\varXi) = \mathbb{P}\{\varXi\} = \sum_{n=0}^{\infty}\mathbb{P}\{\xi_R = n\}\mathbb{P}\Big\{\varXi_R \mid \xi_R = n\Big\}$$

$$= \sum_{n=0}^{\infty}\mathbb{P}\{\xi_R = n\} \times C_R(n).$$

Assume that the root copy number follows a Pólya distribution with some parameters $\kappa_R, q_R > 0$:

$$L(\varXi) = \binom{\kappa_R + n - 1}{n}(1-q_R)^r(q_R)^n \times C_R(n). \quad (11)$$

After defining $\eta_R = 0$, Eq. (11) is the same formula for the likelihoods K_R as on the edges, and $L(\varXi) = K_R(0)$.

3.3. Empty profile likelihood

Typically, the input sample does not include families with an *empty profile* that has $\xi_v = 0$ at all leaves v. We do not insist on empty profiles having a particular biological interpretation – after all, one could add empty profiles based on genes annotated outside the input genomes – but they should not be ignored. First, the extinction, or complete loss of descendants for a gene that existed in some organism at some time is a biologically plausible scenario. The model, in fact, quantifies the per-copy extinction probability exactly (by ϵ in Theorem 4). Secondly, the multitude of unsequenced genomes certainly have many genes waiting to be discovered. Thirdly, future evolution also results in entirely new gene families, as the (presumably finite) landscape of functional coding sequences is explored by mutation and descent. In the first case, the empty profile has a non-empty history. The latter two cases are conceptualized by a placeholder empty profile with an empty history. The model captures our observation bias in the input sample by the empty profile's probability.

Theorem 4 (*Empty Profile Likelihood*). *Define the (per-copy) survival parameters $\tilde{p}_u, \tilde{q}_u$ and the (per-copy) extinction probability ϵ_u:*

(i) *$\epsilon_u = 0$ for all leaves u, and for every non-leaf u, $\epsilon_u = \prod_{uv \in T} \tilde{p}_v$.*

(ii) *At every non-root v,*

$$\tilde{p}_v = \big(p_v + (1-p_v)\epsilon_v(1-\tilde{q}_v)\big) = \frac{p_v(1-\epsilon_v) + \epsilon_v(1-q_v)}{1 - q_v\epsilon_v}.$$

(iii) *At every node u,*

$$\tilde{q}_u = q_u\frac{1-\epsilon_u}{1-q_u\epsilon_u}.$$

The probability of the empty profile is

$$L(0) = \prod_{u=1}^{R}(1 - \tilde{q}_R)^{\kappa_R} = \prod_{u=1}^{R}\left(\frac{1 - q_u}{1 - q_u\epsilon_u}\right)^{\kappa_u}.$$

Let the input sample consist of the observed profiles for families $f = 1, \ldots, F$: $\Xi^f = \left\{\xi_u = n_{f,u} : u \in \mathcal{L}\right\}$. If the empty profiles are unobservable, then the likelihood of a single family profile is conditioned on the fact that at least one copy number is positive:

$$L^*(\Xi^f) = \mathbb{P}\left\{\forall u \in \mathcal{L}_R : \xi_u = n_{f,u} \mid \exists u \in \mathcal{L}_R : \xi_u \neq 0\right\} = \frac{L(\Xi^f)}{1 - L(0)},$$

using the uncorrected likelihoods $L(\Xi)$ without conditioning on being empty, and in particular the empty profile likelihood $L(0)$ from Theorem 4. Applying the correction to the entire sample:

$$L^* = \prod_{f=1}^{F} L^*(\Xi^f) = \frac{\prod_{f=1}^{F} L(\Xi^f)}{(1 - L(0))^F}. \tag{12}$$

The correction of Eq. (12) is akin to Felsenstein's likelihood correction formula for restriction site evolution (Felsenstein, 1992).

3.4. Computing the profile likelihood

Since the ancestors' copy number $\{\xi_u = n\}$ may be possible for all nonnegative integers n, the likelihood recurrences of (10) involve infinite sums for K_u, and infinitely many $C_u(n)$. We can, however, factor out the histories with parallel losses for a finite calculation. Define $\tilde{\xi}_u$ at every ancestral node u as the number of copies that are not lost simultaneously in all descendant lineages to $\mathcal{L}_u$. Let $\tilde{\eta}_u$ denote the number of surviving ancestral copies: those that are not lost either on the edge leading to u or in the subtree T_u. In other words, the *ancestral copy numbers* $\tilde{\eta}$ and $\tilde{\eta}$ count only the progenitors of extant copies at the leaves, as opposed to the *ancestors'* copy numbers ξ, η that count all homologs in the ancestors' genomes. See Fig. 2 for an illustration. Define $\epsilon_u, \tilde{p}_v$ and $\tilde{q}_u$ as in Theorem 4. Since copies are extinct independently with probability ϵ_u, for $0 \leq \ell \leq n$, $\mathbb{P}\{\tilde{\xi}_u = \ell \mid \xi_u = n\} = \binom{n}{\ell}(1 - \epsilon_u)^\ell(\epsilon_u)^{n-\ell}$.

Theorem 5 (*Likelihood Computation*). *Given a profile Ξ, define the conditional likelihoods*

$$\tilde{K}_u(s) = \mathbb{P}\left\{\Xi_u \mid \tilde{\eta}_u = s\right\} \quad \text{and} \quad \tilde{C}_u(\ell) = \mathbb{P}\left\{\Xi_u \mid \tilde{\xi}_u = \ell\right\}$$

at all nodes u. In particular, the profile likelihood is $L(\Xi) = \tilde{K}_R(0)$ at the root R. Define the sum of observed leaf copy numbers within every subtree: $m_u = \sum_{v \in \mathcal{L}_u} n_v$.

(i) *For all $s > m_u$, $\tilde{K}_u(s) = 0$, and for all $\ell > m_u$, $\tilde{C}_u(\ell) = 0$.*
(ii) *At every node u, for all $0 \leq s \leq m_u$,*

$$\tilde{K}_u(s) = \sum_{\ell=s}^{m_u} \tilde{C}_u(\ell) \times \binom{\kappa_u + \ell - 1}{\ell - s}(1 - \tilde{q}_u)^{\kappa_u + s}(\tilde{q}_u)^{\ell - s}. \tag{13}$$

(iii) *If u is a leaf, then $\tilde{C}_u(\ell) = \{\ell = n_u\}$. If u is an ancestral node with children $uv, uw \in T$, then for all $0 \leq \ell \leq m_u = m_v + m_w$,*

$$\tilde{C}_u(\ell) = \sum_{s=0}^{\min\{\ell, m_v\}} \tilde{K}_v(s) \times \tilde{K}_w^\ell(\ell - s)$$
$$\times \binom{\ell}{s}\left(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^s\left(\frac{\tilde{p}_v - \tilde{p}_v\tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell-s} \tag{14}$$

with $\tilde{K}_w^\ell(\ell) = \tilde{K}_w(\ell)$, and, for all $0 \leq d < \ell$,

$$\tilde{K}_w^\ell(d) = (1 - \tilde{p}_w) \times \tilde{K}_w^\ell(d + 1) + \tilde{p}_w \times \tilde{K}_w^{\ell-1}(d). \tag{15}$$

Note that Eq. (13) also applies to a duplication-loss ($\lambda_u, \mu_u > 0$) model with no gain ($\kappa_u = 0$). Then, since $\tilde{\xi}_u$ is the sum of $s = \tilde{\eta}_u$ geometric distributions, it has a negative binomial distribution with parameters s and $\tilde{q}$. So, $\tilde{K}_u(0) = C_u(0)$, and for all $1 \leq s \leq m_u$,

$$\tilde{K}_u(s) = \sum_{\ell=0}^{m_u}\binom{\ell - 1}{\ell - s}(1 - \tilde{q}_u)^s\tilde{q}_u^{\ell-s}.$$

Non-binary phylogeny

A *degenerate* phylogeny T represents the parent–child relationships in a non-binary rooted tree. In such a phylogeny, the ancestral nodes may have 2 or more children. In practice, it makes sense to put multifurcating nodes at deep ancestors to represent the ambiguity of resolving short edges, and a ternary root is common if the phylogeny was derived from an unrooted tree. The likelihood recurrences of Theorem 5 can accommodate any d-ary node, by considering survival in $1, 2, 3, \ldots d$ child lineages incrementally (for any child ordering). More specifically, a d-ary node can be resolved arbitrarily into a series of bifurcations by keeping the distribution parameters on the edges leading to the d leaves, and 0-length inner edges (thus with $p = q = 0$). Theorem 6 uses a straightforward resolution into a right-leaning caterpillar tree.

Theorem 6 (*Likelihood Recurrences for Multifurcating Node*). *Let u be a node in a degenerate phylogeny with $d \geq 2$ distinct children $uv_1, \ldots, uv_d \in T$ enumerated in any order. Let $\epsilon_{u,-i} = \prod_{j=i}^{d} \tilde{p}_{v_j}$ for $i = 1, \ldots, d$, so that $\epsilon_u = \epsilon_{u,-1}$. Define the likelihoods $\tilde{C}_u^{-i}(\ell)$ and $\tilde{K}_{v_{i..d}}^\ell(s)$ conditioned on s surviving copies in the subtrees of $v_i, v_{i+1}, \ldots, v_d$:*

$$\tilde{C}_u^{-d}(\ell) = K_{v_d}(\ell) \qquad \{0 \leq \ell \leq m_{v_d}\} \tag{16a}$$

and, for all $0 < i \leq d$ and for all $0 \leq \ell \leq m_{v_{i-1}} + \cdots + m_{v_d}$

$$\tilde{C}_u^{-(i-1)}(\ell) = \sum_{s=0}^{\min\{\ell, m_{v_{i-1}}\}} \tilde{K}_{v_{i-1}}(s) \times \tilde{K}_{v_{i..d}}^\ell(\ell - s)$$
$$\times \binom{\ell}{s}\left(\frac{1 - \tilde{p}_{v_{i-1}}}{1 - \epsilon_{u,-(i-1)}}\right)^s\left(\frac{\tilde{p}_{v_{i-1}} - \tilde{p}_{v_{i-1}}\epsilon_{u,-i}}{1 - \epsilon_{u,-(i-1)}}\right)^{\ell-s}, \tag{16b}$$

with

$$\tilde{K}_{v_{i..d}}^\ell(\ell) = \tilde{C}_u^{-i}(\ell) \tag{16c}$$
$$\tilde{K}_{v_{i..d}}^\ell(k) = (1 - \epsilon_{u,-i})\tilde{K}_{v_{i..d}}^\ell(k + 1) + \epsilon_{u,-i}\tilde{K}_{v_{i..d}}^{\ell-1}(k) \quad \{0 \leq k < \ell\} \tag{16d}$$

Then $\tilde{C}_u(\ell) = \tilde{C}_u^{-1}(\ell)$.

Proof. Omitted. □

Partial genomes

An incomplete genome at a leaf u is characterized by the fraction $(1 - \epsilon_u)$ of the genome that is annotated. Assuming a simple model of randomly missing copies, we have

$$\mathbb{P}\left\{\tilde{\xi}_u = k \mid \xi_u = n\right\} = \binom{n}{k}(1 - \epsilon_u)^k(\epsilon_u)^{n-k},$$

where $\tilde{\xi}$ is the number of annotated copies, and ξ is the true copy number in the complete genome. In other words, the recurrences of Theorems 5 for the likelihood and 4 for the empty profile remain the same, with the only change that $\tilde{q}_u \neq q_u$ at such a leaf with $\epsilon_u > 0$. Without constraints, however, the trio (p_u, q_u, ϵ_u) is

not identifiable: the distribution parameters

$$p = p_u + (1-p_u)\epsilon_u \frac{1-q_u}{1-q_u\epsilon_u} \qquad q = q_u \frac{1-\epsilon_u}{1-q_u\epsilon_u} \qquad \epsilon = 0,$$

produce the exact same distribution at u as (p_u, q_u, ϵ_u).

3.5. Posterior probabilities for ancestral copy numbers

Let $\Xi = \{\xi_v = n_v : v \in \mathcal{L}\}$ be an arbitrary profile of copy numbers observed at the leaves. Theorem 5 provides the recurrences for the conditional likelihoods $\tilde{C}_u$ and $\tilde{K}_u$ of the partial profile Ξ_u conditioned on the ancestral copies $\tilde{\xi}_u$ and $\tilde{\eta}_u$, respectively. Define the complementary *outside likelihoods*

$$\tilde{B}_u(\ell) = \mathbb{P}\{\Xi - \Xi_u, \tilde{\xi}_u = \ell\} \quad \text{and} \quad \tilde{J}_u(s) = \mathbb{P}\{\Xi - \Xi_u, \tilde{\eta}_u = s\}, \quad (17)$$

where $\Xi - \Xi_u = \{\xi_v = n_v : v \in \mathcal{L} - \mathcal{L}_u\}$ denotes the profile outside the subtree rooted at node u.

Theorem 7 (*Outside Likelihoods*)**.** *Let $\Xi = \{\xi_v = n_v : v \in \mathcal{L}\}$ be an arbitrary profile, and define the outside likelihoods as in Eq. (17). The following recurrences hold.*

(i) *At the root, $\tilde{J}_R(0) = 1$ and $\tilde{J}_R(s) = 0$ for $s > 0$.*
(ii) *At any node u, for all $0 \le \ell \le m_u$,*

$$\tilde{B}_u(\ell) = \sum_{s=0}^{\ell} \tilde{J}_u(s) \times \binom{\kappa_u + \ell - 1}{\ell - s}(1 - \tilde{q}_u)^{\kappa_u + s}(\tilde{q}_u)^{\ell - s}. \quad (18)$$

(iii) *At every non-root node v with parent u and sibling w (i.e, $uv, uw \in T$), for all $0 \le s \le m_v$,*

$$\tilde{J}_v(s) = \sum_{\ell=s}^{m_u} \tilde{B}_u(\ell) \times \tilde{K}_w^{\ell}(\ell - s)$$
$$\times \binom{\ell}{s}\left(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^s \left(\frac{\tilde{p}_v - \tilde{p}_v\tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell - s}. \quad (19)$$

Theorem 7 with Theorem 5 deliver the posterior probabilities in computable forms.

Corollary 8 (*Posterior Probabilities*)**.** *Fix an arbitrary profile Ξ and let $\tilde{C}_u, \tilde{K}_u, \tilde{B}_u, \tilde{J}u$ denote the inside and outside likelihoods at every node u.*

(i) *The profile likelihood can be computed by either formulas*

$$L(\Xi) = \mathbb{P}\{\Xi\} = \sum_{\ell=0}^{m_u} \tilde{B}_u(\ell) \times \tilde{C}_u(\ell) = \sum_{s=0}^{m_u} \tilde{J}_u(s) \times \tilde{K}_u(s). \quad (20)$$

(ii) *The posterior distribution of $\tilde{\xi}_u$ is $\mathbb{P}\{\tilde{\xi}_u = \ell \mid \Xi\} = \frac{\tilde{B}_u(\ell) \times \tilde{C}_u(\ell)}{L(\Xi)}$.*
(iii) *The posterior distribution of $\tilde{\eta}_u$ is $\mathbb{P}\{\tilde{\eta}_u = s \mid \Xi\} = \frac{\tilde{J}_u(s) \times \tilde{K}_u(s)}{L(\Xi)}$.*

3.6. Partial derivatives of the likelihood

Suppose that we are interested in the corrected likelihood for a sample of family profiles $\{\Xi^f : f = 1, \ldots, F\}$. By Eq. (12), the derivative of the corrected log-likelihood, with respect to any distribution parameter θ is

$$\frac{\partial}{\partial \theta}\left(\ln L^*\right) = \left(\sum_{f=1}^{F} \frac{L'(\Xi^f)}{L(\Xi^f)}\right) + F \frac{L'(0)}{1 - L(0)}, \quad (21)$$

where $L'(\Xi) = \frac{\partial L(\Xi)}{\partial \theta}$ denotes the derivative of the uncorrected profile likelihood.

It is tempting to choose the optimized distribution parameters directly as κ_u and the survival parameters $\tilde{p}_u, \tilde{q}_u$ for the

maximization of the corrected log-likelihood $\ln L^*$. They uniquely determine the parameters p_u, q_u, and, consequently, the edge-specific rate parameters. The values of $\tilde{p}$ and $\tilde{q}$ are, however, not arbitrary across the tree.

Theorem 9 (*Unicity of Survival Parameters*)**.** *Let T be a phylogeny equipped with gain rates $0 < \kappa_u$ and survival parameters $0 < \tilde{p}_u, \tilde{q}_u < 1$ at every node u. If, at every non-root ancestral node u,*

$$\tilde{p}_u > (1 - \tilde{q}_u) \prod_{uv \in T} \tilde{p}_v, \quad (22)$$

then there exists a phylogenetic birth–death model on the same phylogeny with distribution parameters $0 < p_u, q_u < 1$ and same gain rates κ_u. Otherwise, no solution exists with positive p_u on every edge.

In light of Theorem 9, we should aim at using the partial derivatives with respect to $\tilde{p}$ and $\tilde{q}$ as an intermediate step toward inferring the dependence on p and q. Using Corollary 8, we can determine the partial derivatives with respect to the survival distribution parameters.

Theorem 10 (*Partial Derivatives of Then Profile Likelihood*)**.** *Let Ξ be an arbitrary phyletic profile.*

(i) *At every node $1 \le u \le R$,*

$$\frac{\partial L(\Xi)}{\partial \tilde{q}_u} = L(\Xi) \times \left(\frac{1}{\tilde{q}_u} \mathbb{E}[\tilde{\xi}_u \mid \Xi] \right.$$
$$\left. - \left(\frac{1}{\tilde{q}_u} + \frac{1}{1 - \tilde{q}_u}\right)\mathbb{E}[\tilde{\eta}_u \mid \Xi] - \frac{1}{1 - \tilde{q}_u}\kappa_u \right).$$

(ii) *At every non-root node $1 \le v < R$,*

$$\frac{\partial L(\Xi)}{\partial \tilde{p}_v} = L(\Xi) \times \left(\frac{1}{\tilde{p}_v}\mathbb{E}[\tilde{\xi}_u \mid \Xi] - \left(\frac{1}{\tilde{p}_v} + \frac{1 - \epsilon}{1 - \tilde{p}_v}\right)\mathbb{E}[\tilde{\eta}_v \mid \Xi]\right),$$

where $\epsilon = \epsilon_u/\tilde{p}_v = (\prod_{uw \in T} \tilde{p}_w)/\tilde{p}_v$ is the product of $\tilde{p}_w$ across the siblings with the same parent $uv, uw \in T$. (Simply $\epsilon = \tilde{p}_w$ if there is only one sibling $uw \in T$.)

(iii) *The partial derivatives with respect to κ_u are, for all $1 \le u \le R$,*

$$\frac{\partial L(\Xi)}{\partial \kappa_u}$$
$$= L(\Xi) \times \left(\ln(1 - \tilde{q}_u) + \sum_{i=0}^{m_u - 1} \frac{\mathbb{P}\{\tilde{\xi}_u > i \mid \Xi\} - \mathbb{P}\{\tilde{\eta}_u > i \mid \Xi\}}{\kappa_u + i}\right).$$

(iv) *The partial derivatives for the empty profile $\Xi = 0$ are*

$$\frac{\partial L(0)}{\partial \tilde{q}_u} = -L(0)\frac{\kappa_u}{1 - \tilde{q}_u},$$
$$\frac{\partial L(0)}{\partial \tilde{p}_u} = 0 \quad \text{and} \quad \frac{\partial L(0)}{\partial \kappa_u} = L(0) \times \ln(1 - \tilde{q}_u).$$

Note that using the posterior distributions from Corollary 8, we readily obtain the posterior expectations

$$\mathbb{E}[\tilde{\xi}_u \mid \Xi] = \sum_{\ell=0}^{m_u} \ell \times \mathbb{P}\{\tilde{\xi}_u = \ell \mid \Xi\} \quad \text{and}$$

$$\mathbb{E}[\tilde{\eta}_u \mid \Xi] = \sum_{s=0}^{m_u} s \times \mathbb{P}\{\tilde{\eta}_\xi = s \mid \Xi\},$$

as well as the distribution tails $\mathbb{P}\{\tilde{\xi}_u > i \mid \Xi\} = \sum_{\ell=i+1}^{m_u} \mathbb{P}\{\tilde{\xi}_u = \ell \mid \Xi\}$ and $\mathbb{P}\{\tilde{\eta}_u > i \mid \Xi\} = \sum_{\ell=i+1}^{m_u} \mathbb{P}\{\tilde{\eta}_u = \ell \mid \Xi\}$ which are needed in Theorem 10 and the following Corollary 11 that combines Theorem 10 with Eq. (21).

Corollary 11 (*Partial Derivatives of Sample Likelihood by Survival Parameters*). *Let* $\ln L$ *and* $\ln L^*$ *denote the uncorrected and corrected log-likelihood for a sample of family profiles* $\{\Xi^f : f = 1, \ldots, F\}$. *Define the posterior expected counts across the sample*

$$\tilde{N}_u = \sum_{f=1}^{F} \mathbb{E}\big[\tilde{\xi}_u \mid \Xi^f\big] \qquad \tilde{S}_u = \sum_{f=1}^{F} \mathbb{E}\big[\tilde{\eta}_u \mid \Xi^f\big]$$

$$\tilde{N}_u^{>i} = \sum_{f=1}^{F} \mathbb{P}\big\{\tilde{\xi}_u > i \mid \Xi^f\big\} \qquad \tilde{S}_u^{>i} = \sum_{f=1}^{F} \mathbb{P}\big\{\tilde{\eta}_u > i \mid \Xi^f\big\}$$

at every node $1 \le u \le R$.

(i) *At every node* $1 \le u \le R$,

$$\frac{\partial}{\partial \tilde{q}_u} \ln L = \frac{\tilde{N}_u - \tilde{S}_u}{\tilde{q}_u} - \frac{\tilde{S}_u + \kappa_u F}{1 - \tilde{q}_u}$$

$$\frac{\partial}{\partial \tilde{q}_u} \ln L^* = \frac{\tilde{N}_u - \tilde{S}_u}{\tilde{q}_u} - \frac{\tilde{S}_u + \kappa_u \frac{F}{1-L(0)}}{1 - \tilde{q}_u}$$

(ii) *For a non-root node* $1 \le v < R$, *let* u *be its parent*:

$$\frac{\partial}{\partial \tilde{p}_v} \ln L = \frac{\partial}{\partial \tilde{p}_v} \ln L^* = \frac{\tilde{N}_u - \tilde{S}_v}{\tilde{p}_v} - \frac{(1-\epsilon)\tilde{S}_v}{1 - \tilde{p}_v}$$

$$= \frac{\tilde{N}_u - (1-\epsilon_u)\tilde{S}_v}{\tilde{p}_u} - \frac{(1-\epsilon_u)\tilde{S}_v}{1 - \tilde{p}_v}$$

where $\epsilon = \big(\prod_{uw \in T} \tilde{p}_w\big)/\tilde{p}_v$.

(iii) *At every node* $1 \le u \le R$,

$$\frac{\partial}{\partial \kappa_u} \ln L = F \ln(1 - \tilde{q}_u) + \sum_{i=0}^{m_u - 1} \frac{\tilde{N}_u^{>i} - \tilde{S}_u^{>i}}{\kappa_u + i}$$

$$\frac{\partial}{\partial \kappa_u} \ln L^* = F \frac{\ln(1 - \tilde{q}_u)}{1 - L(0)} + \sum_{i=0}^{m_u - 1} \frac{\tilde{N}_u^{>i} - \tilde{S}_u^{>i}}{\kappa_u + i}$$

Powerful numerical algorithms for function maximization (conjugate gradient and variable metric methods like Broyden–Fletcher–Goldfarb–Shanno) exploit the gradient for quick convergence to optimum. The likelihood optimization for a phylogenetic birth–death model can rely on the computation of both the likelihood (Theorem 5), and the gradient with respect to the parameters κ_u, p_u, q_u across the tree. By Theorem 1, the probabilistic model is uniquely determined by the parameter set, up to equivalent rate scalings. Maximizing the likelihood with respect to the survival distribution parameters $\tilde{p}$ and $\tilde{q}$ from Theorem 10 is not straightforward because Theorem 9 imposes monotonicity constraints between parameters on adjoining edges. Let u be an arbitrary node at some depth d (root is at depth 0). For a distribution parameter such as $\theta_v = p_v$ or $\theta_v = q_v$,

$$\frac{\partial L(\Xi)}{\partial \theta_v} = \sum_{u=1}^{R} \frac{\partial L(\Xi)}{\partial \tilde{q}_u} \frac{\partial \tilde{q}_u}{\partial \theta_v} + \sum_{u=1}^{R} \frac{\partial L(\Xi)}{\partial \tilde{p}_u} \frac{\partial \tilde{p}_u}{\partial \theta_v},$$

by the chain rule. In particular, p_v and q_v influence $\tilde{p}_u$ and $\tilde{q}_u$ at nodes u along the path between the root and v. Consequently, the above sums include only the ancestors of v, and the partial derivatives can be computed in a preorder traversal. We state the procedure in a generic theorem about recovering the derivatives of any function Φ of the distribution parameters.

Theorem 12 (*Gradient Computation*). *Let* Φ *be an arbitrary differentiable function of the survival parameters* $\{\tilde{p}_u, \tilde{q}_u\}_{u=1}^{R}$. *Let* $\Phi^{(\theta_v)} = \frac{\partial f}{\partial \theta_v}$ *denote the partial derivative with respect to any distribution parameter* θ_v. *The partial derivatives* $\Phi^{(p_v)}$ *(for non-root* v*),* $\Phi^{(q_v)}$ *(for any* v*) and* $\Phi^{(\epsilon_v)}$ *(for non-leaf* v*) can be computed in a preorder traversal by the following recurrences.*

(i) *At the root* $v = R$,

$$\Phi^{(q_R)} = \frac{1 - \epsilon_R}{(1 - q_R \epsilon_R)^2} \Phi^{(\tilde{q}_R)} \tag{23a}$$

$$\Phi^{(\epsilon_R)} = \frac{1 - q_R}{(1 - q_R \epsilon_R)^2} \big(-q_R \Phi^{(\tilde{q}_R)}\big). \tag{23b}$$

(ii) *At every non-root node* $1 \le v < R$,

$$\Phi^{(p_v)} = \frac{1 - \epsilon_v}{1 - q_v \epsilon_v} \big(\Phi^{(\tilde{p}_v)} + \epsilon \Phi^{(\epsilon_u)}\big) \tag{24a}$$

$$\Phi^{(q_v)} = \frac{1 - \epsilon_v}{(1 - q_v \epsilon_v)^2} \big(\Phi^{(\tilde{q}_v)} - (1 - p_v)\epsilon_v \big(\Phi^{(\tilde{p}_v)} + \epsilon \Phi^{(\epsilon_u)}\big)\big) \tag{24b}$$

and, if v *is not a leaf,*

$$\Phi^{(\epsilon_v)} = \frac{1 - q_v}{(1 - q_v \epsilon_v)^2} \big((1 - p_v)\big(\Phi^{(\tilde{p}_v)} + \epsilon \Phi^{(\epsilon_u)}\big) - q_v \Phi^{(\tilde{q}_v)}\big) \tag{24c}$$

with the parent u *and*

$$\epsilon = \prod_{w \,:\, uw \in T} \{w \ne v\} \tilde{p}_w = \frac{\epsilon_u}{\tilde{p}_v}.$$

(For a binary node, $\epsilon = \tilde{p}_w$ *with the sole sibling* w*).*

Theorem 12 can be employed with the individual family profiles using $\Phi = L(\Xi^f)$ and plugging $L'(\Xi^f) = \Phi^{(\theta)}$ into the corrected log-likelihood formula of (21) for each $f = 1, \ldots, F$ in the sum, as well as for $L'(0)$. But it may be more efficient to carry out the procedure only once at the end, using $\Phi = \ln L^*$ directly with its partial derivatives from Corollary 11. For the purposes of likelihood maximization, use a parametrization with the logistic and exponential functions as

$$p_u = \frac{1}{1 + e^{-\alpha_u}} \quad q_u = \frac{1}{1 + e^{-\beta_u}} \quad \kappa_u = e^{\omega_u}$$

with unconstrained real-valued parameters

$$\alpha_u = \ln \frac{p_u}{1 - p_u} \quad \beta_u = \ln \frac{q_u}{1 - q_u} \quad \omega_u = \ln \kappa_u.$$

The partial derivatives are computed by the chain rule as

$$\Phi^{(\alpha_u)} = \Phi^{(p_u)} \frac{\partial p_u}{\partial \alpha_u} = p_u(1 - p_u) \Phi^{(p_u)} \qquad \{0 < u < R\}$$

$$\Phi^{(\beta_u)} = \Phi^{(q_u)} \frac{\partial q_u}{\partial \beta_u} = q_u(1 - q_u) \Phi^{(q_u)} \qquad \{0 < u \le R\}$$

$$\Phi^{(\omega_u)} = \Phi^{(\kappa_u)} \frac{\partial \kappa_u}{\partial \omega_u} = \kappa_u \Phi^{(\kappa_u)} \qquad \{0 < u \le R\}$$

3.7. Likelihoods in the no-duplication model

In the case of $\lambda_v = 0$ on all edges $uv \in T$, the joint distribution of the random variables multiplies Poisson and binomial masses:

$$\mathbb{P}\{\xi_1 = n_1, \ldots, \xi_R = n_R, \eta_1 = s_1, \ldots, \eta_{R-1} = s_{R-1}\}$$

$$= \mathbb{P}\{\xi_R = n_R\} \times \prod_{uv \in T} \left(\binom{n_u}{s_v}(1 - p_v)^{s_v}(p_v)^{n_u - s_v} \times e^{-r_v} \frac{(r_v)^{n_v - s_v}}{(n_v - s_v)!}\right).$$

This time we assume a Poisson distribution at the root: $\mathbb{P}\{\xi_R = n\} = e^{-r} r^n / (n!)$, and, as before $s_R = 0$ for retrieving the likelihood $L(\Xi) = K_R(0)$. The recurrences for the likelihood and the empty profile are adjusted accordingly. In particular,

$$K_u(s) = \sum_{k=0}^{\infty} e^{-r_u} \frac{(r_u)^k}{k!} \times C_u(s + k), \tag{25}$$

but the recurrence for C_u stays the same.

Theorem 13 (*Empty Profile in the No-Duplication Model*). *Define ϵ_u as in Theorem 4, with $\tilde{p}_v = (p_v + (1 - p_v)\epsilon_v)$ at every non-root node v, and $\tilde{r}_u = r_u(1 - \epsilon_u)$ at every node u. The probability of the empty profile is*

$$L(0) = \prod_{u=1}^{R} e^{-\tilde{r}_u} = \prod_{u=1}^{R} \exp\Big(-r_u(1 - \epsilon_u)\Big).$$

The likelihood computations of Theorems 5 and 7 adapt easily to the no-duplication model, with $\tilde{\xi}$ and $\tilde{\eta}$ defined as before. Two recurrences change: at every node u, and for all $0 \le s \le m_u$,

$$\tilde{K}_u(s) = \sum_{\ell=s}^{m_u} \tilde{C}_u(\ell) \times e^{-\tilde{r}_u} \frac{(\tilde{r}_u)^{\ell-s}}{(\ell-s)!},$$

and, for all $0 \le \ell \le m_u$,

$$\tilde{B}_u(\ell) = \sum_{s=0}^{\ell} \tilde{J}_u(s) \times e^{-\tilde{r}_u} \frac{(\tilde{r}_u)^{\ell-s}}{(\ell-s)!}.$$

Consequently, the derivatives of the profile likelihood are

$$\frac{\partial L(\Xi)}{\partial \tilde{r}_u} = \frac{\partial}{\partial \tilde{r}_u}\Big(\sum_{\ell=0}^{m_u} \tilde{B}_u(\ell) \times \tilde{C}_u(\ell)\Big)$$

$$= \sum_{0 \le s \le \ell \le m_u} \tilde{J}_{u(s)} \times \tilde{C}_u(\ell) \times e^{-\tilde{r}_u} \frac{(\tilde{r}_u)^{\ell-s}}{(\ell-s)!}\Big(\frac{\ell-s}{\tilde{r}_u} - 1\Big)$$

$$= L(\Xi) \times \Big(\frac{\mathbb{E}\big[\tilde{\xi}_u \mid \Xi\big] - \mathbb{E}\big[\tilde{\eta}_u \mid \Xi\big]}{\tilde{r}_u} - 1\Big).$$

at every node $1 \le u \le R$. In particular, for the empty profile $\Xi = 0$,

$$\frac{\partial L(0)}{\partial \tilde{r}_u} = \frac{\partial}{\partial \tilde{r}_u}\Big(\prod_{v=1}^{R} e^{-\tilde{r}_v}\Big) = -L(0).$$

by Theorem 13. Substituting into Eq. (21) for the derivatives of corrected log-likelihood on a sample of family profiles gives

$$\frac{\partial(\ln L^*)}{\partial \tilde{r}_v} = \frac{\tilde{N}_u - \tilde{S}_u}{\tilde{r}_u} - \frac{F}{1 - L(0)}.$$

The analogue of Theorem 12 is the following claim.

Theorem 14 (*Gradient in the No-Duplication Model*). *Let Φ be an arbitrary differentiable function of the distribution parameters $\{\tilde{p}_u, \tilde{r}_u\}_{u=1}^{R}$ in a no-duplication model. Let $\Phi^{(\theta_v)} = \frac{\partial f}{\partial \theta_v}$ denote the partial derivative with respect to any distribution parameter θ_v. The partial derivatives $\Phi^{(p_v)}$ (for non-root v), $\Phi^{(r_v)}$ (for any v) and $\Phi^{(\epsilon_v)}$ (for non-leaf v) can be computed in a preorder traversal by the following recurrences.*

(i) *At the root $v = R$,*

$$\Phi^{(r_R)} = (1 - \epsilon_R)\Phi^{(\tilde{r}_R)} \tag{26a}$$

$$\Phi^{(\epsilon_R)} = -r_R\Phi^{(\tilde{r}_R)}. \tag{26b}$$

(ii) *At every non-root node $1 \le v < R$,*

$$\Phi^{(p_v)} = (1 - \epsilon_v)\Big(\Phi^{(\tilde{p}_v)} + \epsilon\Phi^{(\epsilon_u)}\Big) \tag{27a}$$

$$\Phi^{(r_v)} = (1 - \epsilon_v)\Phi^{(\tilde{r}_v)}. \tag{27b}$$

and, if v is not a leaf,

$$\Phi^{(\epsilon_v)} = (1 - p_v)\Big(\Phi^{(\tilde{p}_v)} + \epsilon\Phi^{(\epsilon_u)}\Big) - r_v\Phi^{(\tilde{r}_v)} \tag{27c}$$

with the parent u and

$$\epsilon = \prod_{w:uw \in T} \{w \ne v\}\tilde{p}_w = \frac{\epsilon_u}{\tilde{p}_v}.$$

Note that the different duplication models can be used in the same tree: some edges can have $\lambda = 0$, and some $\lambda > 0$. In the recurrences for $\tilde{K}_v$ and $\tilde{B}_v$, either the Poisson (if $\lambda_v = 0$) or the Pólya (if $\lambda_v > 0$) formulas apply, and the computed derivatives are ∂r_v or $\partial \kappa_v$, respectively.

3.8. Algorithmic complexity

The set of conditional likelihoods $\tilde{C}_u(\ell)$ and $\tilde{K}_u(s)$ for a given profile Ξ can be computed in a postorder traversal of the phylogeny using Theorem 5. The recurrences for $\tilde{K}_u(s)$ from (13) are straightforward to implement by embedded loops over $0 \le s \le \ell \le m_u$. Define

$$h_u(s, t) = \binom{\kappa_u + s + t - 1}{t}(1 - \tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^t.$$

```
// Computing K̃_u(s) for all s
1   for ℓ ← 0, 1, . . . , m_u
2       for s ← 0, 1, . . . , ℓ
3           K̃_u(s) ← K̃_u(s) + C̃_u(ℓ) × h_u(s, ℓ − s)
```

For the recurrence of (13), compute $K_w^{s+t}(s)$ looping over t and s in the opposite direction. Let

$$g_v(s, t) = \binom{s + t}{s}\Big(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\Big)^s\Big(\frac{\tilde{p}_v - \tilde{p}_v\tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\Big)^t.$$

```
// Computing C̃_u(ℓ) for all ℓ at u with children uv, uw ∈ T
1   for t ← m_w, m_w − 1, . . . , 0
2       K_w^t(t) ← K_w(t)
3       for s ← 0, 1, . . . , m_v
4           C̃_u(s + t) ← C̃_u(s + t) + K̃_v(s) × K̃_w^{s+t}(t) × g_v(s, t)
5           K_w^{(s+1)+t}(t) ← (1 − p̃_w)K̃_w^{(s+1)}(t + 1) + p̃_w K̃_w^{s+t}(t)
```

Note that $g_u(s, t)$ and $h_u(s, t)$ can be computed in constant time. For instance, $\ln h_u(s, 0) = (\kappa_u + s)\ln(1 - \tilde{q}_u)$, and for $t > 0$,

$$\ln h_u(s, t) = (\kappa_u + s)\ln(1 - \tilde{q}_u) + t\ln\tilde{q}_u$$
$$+ \ln\Gamma(\kappa_u + s + t) - \ln\Gamma(\kappa_u + s) - \ln\Gamma(t + 1)$$

with the Gamma function $\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$ (so that $\Gamma(t + 1) = t!$).

The outside likelihoods $\tilde{B}_u(\ell)$ and $\tilde{J}_u(s)$ from Theorem 7 are computed in a preorder traversal. Concomitantly, the posterior distributions for ancestral copy numbers $\tilde{\xi}_u$ and $\tilde{\eta}_u$ are obtained by Corollary 8 in the same traversal. In addition, during the same preorder traversal, the partial derivatives can be computed with respect to all $\tilde{p}_u, \tilde{q}_u, \kappa_u$ parameters. In order to get the gradient of the corrected log-likelihood $\Phi = \ln L^*$ over a sample of F families, first compute the partial derivatives $\Phi^{(\tilde{p}_u)}$, $\Phi^{(\tilde{q}_u)}$, and $\Phi^{(\kappa_u)}$ of the corrected log-likelihood from the derivatives for the individual profile likelihoods using Eq. (21). Subsequently, the recurrences of Theorem 12 compute all $\Phi^{(p_u)}$ and $\Phi^{(q_u)}$ in a single preorder traversal. The running time is quadratic in the total number of observed copies.

Theorem 15 (*Running Time for Likelihood Computation*). *Let $\Xi = \{\xi_u = n_u : u \in \mathcal{L}\}$ be an arbitrary profile across $L = |\mathcal{L}|$ leaves. Let $\bar{n} = \frac{1}{L}\sum_{u=1}^{L} n_u$ be the average of the leaf copy numbers within the profile, and let h be the phylogeny's height. The profile likelihood and the posterior distributions for $\tilde{\xi}_u$ and $\tilde{\eta}_u$ at all ancestral nodes u can be computed in $O\big(hL(L\bar{n}^2 + 1)\big)$ time.*

The height $h = O(\log L)$ for almost all random phylogenies in the Yule model (where the tree is generated by a birth process), as well as in Aldous' beta-splitting model (Aldous, 1996) (when $\beta > -1$); in the biologically less plausible uniform and unbalanced beta-splitting models, $h = O(\sqrt{L})$.

4. Conclusion

We report fast algorithms for likelihood, posterior and gradient calculations for linear birth–death processes on a phylogeny. For a tree with L leaves and a profile with $L\bar{n}$ total copies, the likelihood computations take $O(hL^2\bar{n}^2)$ time, which matches previous algorithms' running time (Csűrös and Miklós, 2006; Csűrös and Miklós, 2009). Imposing a maximum *ancestral* copy number $\tilde{\xi} \leq n_{max}$ to truncate the summations (in Theorems 5 and 7) yields $\Theta(Ln_{\max}^2)$ running time. Other methods (Hahn et al., 2005; Iwasaki and Takagi, 2007; Spencer et al., 2006; Ames et al., 2012; Fukunaga and Iwasaki, 2021) that adapt Felsenstein's peeling algorithm with maximum *ancestor* copy number $\xi \leq n_{max}$ have the same asymptotic running time, provided that transition probabilities can be computed easily. If transition probabilities are not readily available, however, then the exponentiation of the rate matrix on all edges takes $\Theta(Ln_{\max}^3)$ time, restricting applicability to smaller families.

A subtle, but important feature of our methods is that they incorporate observation bias for empty profiles. Neglecting the likelihood correction may be problematical, because it inflates the appearance of conservation, and thus results in biased ancestral reconstruction and parameter inference.

In our approach, posterior distributions for ancestral copy numbers are computed alongside the likelihoods. To our knowledge, the gradient computation algorithm (Theorem 12) is entirely novel of its kind. It extracts the partial derivatives from the posterior copy number distributions in linear time, necessitating a single evaluation of the likelihood. In contrast, approximate gradient calculation (implemented in Count, Csűrös, 2010, for example) for $\Theta(L)$ distribution parameters entails $\Theta(L)$ evaluations of the likelihood.

Our mathematical framework for phylogenetic gain-loss-duplication models provides the clean decomposition of Eq. (9), involving a network of dependent random variables. The elementary decomposition can be employed with standard Bayesian and likelihood methods, leading to efficient algorithms for a notoriously hard bioinformatics problem.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Appendix

We give the proofs for Theorem 2, Corollary 3, Theorems 4–7, Corollary 8, Theroems 9, 10, Corollary 11 and Theorem 12 in Appendices A.1–A.7. The results about the no-duplication model are proven in Appendix A.8. Finally, Appendix A.9 discusses the proof of Theorem 15 for the running time. Table 1 shows the notations in the proofs.

A.1. Proof of Theorem 2

Proof. First, suppose that $\lambda > 0$. Decompose $\xi(t)$ as in (5). For $\xi(0) = n$,

$$\xi(t) = \chi(t) + \sum_{i=1}^{n} \zeta_i(t)$$

Table 1
Notations for phylogenetic birth-and-death model.

T	Phylogeny (binary tree) with integer-indexed nodes
R	Number of nodes in the tree (also the root's index)
u, v, w	Nodes in the tree
T_u	Subtree rooted at node u
$\mathcal{L}$ and $\mathcal{L}_u$	Leaf set for T and T_u
ξ_u	Copy number at node u
η_v	Parental copies surviving at node v
χ, ζ_i	Xenolog and inparalog copies (Fig. 1)
κ	Relative gain rate (from outside sources)
γ	Relative gain rate in no-duplication model
λ	Per-copy duplication rate
μ	Per-copy loss rate
t	Edge length (time)
q_u	Duplication parameter on edge leading to node u
p_u	Loss parameter on edge leading to node u
r_u	Gain parameter in no-duplication model
$h_n(t)$	Point mass function for xenolog copy number at time t
$g_n(t)$	Point mass function for inparalog copy number at time t
$\varXi$	Phylogenetic profile (vector of non-negative copy numbers at the leaves)
$\varXi_u$	Profile in the subtree T_u
$\varXi^f$	Profile for family f
ϵ_u	Extinction probability for a copy at node u (Theorem 4)
$\tilde{p}_u$	Survival loss parameter (Theorem 4)
$\tilde{q}_u$	Survival duplication parameter (Theorem 4)
$\tilde{\xi}_u$	Ancestral copy number (non-extinct copies)
$\tilde{\eta}_v$	Surviving ancestral copies
n_w	Copy number observed at leaf w
m_u	Sum of copy numbers at the leaves $\mathcal{L}_u$ below node u

where χ follows Pólya with parameters (κ, q), and ζ_i are iid shifted geometric with parameters (p, q). Now define the random variable $\eta(t)$ as the number of *surviving* copies

$$\eta(t) = \sum_{i=1}^{n}\{\zeta_i(t) > 0\}.$$

Since ζ_i are independent with $\mathbb{P}\{\zeta_i(t) = 0\} = p$,

$$\mathbb{P}\Big\{\eta(t) = s \mid \xi(0) = n\Big\} = \binom{n}{s}(1-p)^s p^{n-s}. \tag{28}$$

Since the $\zeta_i = 0$ are immaterial in the sum, we can condition on $\eta(t)$:

$$\mathbb{P}\Big\{\xi(t) = m \mid \eta(t) = s, \xi(0) = n\Big\} = \mathbb{P}\Big\{\xi(t) = m \mid \eta(t) = s\Big\}$$
$$= \mathbb{P}\{\chi + \zeta_1' + \cdots + \zeta_s' = m\},$$

where $\zeta_1' - 1$ are iid random variables following a Pólya distribution with parameter 1:

$$\mathbb{P}\{\zeta_i' - 1 = k\} = (1-q)q^k = \binom{1+k-1}{k}(1-q)^1 q^k.$$

Looking specifically at the generator functions:

$$F_0(z) = \sum_{i=0}^{\infty} \mathbb{P}\{\chi = i\}z^i = \left(\frac{1-q}{1-qz}\right)^\kappa$$

$$F_i(z) = \sum_{i=0}^{\infty} \mathbb{P}\{\zeta_i' = i\}z^i = \frac{z(1-q)}{1-qz}$$

so

$$F(z) = \sum_{m=s}^{\infty} \mathbb{P}\Big\{\xi(t) = m \mid \eta(t) = s\Big\}z^m$$
$$= F_0(z)\prod_{i=1}^{s} F_i(z) = z^s \left(\frac{1-q}{1-qz}\right)^{\kappa+s}.$$

Hence, $\xi(t) - \eta(t)$ follows a Pólya distribution with parameter $(\kappa + \eta(t))$, and the same tail parameter q. Now,

$$\mathbb{P}\Big\{\xi(t) = m \ \Big| \ \xi(0) = n\Big\}$$

$$= \sum_s \mathbb{P}\Big\{\xi(t) = m \ \Big| \ \eta(t) = s\Big\}\mathbb{P}\Big\{\eta(t) = s \ \Big| \ \xi(0) = n\Big\}$$

$$= \sum_s \mathbb{P}\Big\{\xi(t) - \eta(t) = m - s \ \Big| \ \eta(t) = s\Big\}\mathbb{P}\Big\{\eta(t) = s \ \Big| \ \xi(0) = n\Big\}$$

$$= \sum_{s=0}^{\min\{n,m\}} \binom{n}{s}(1-p)^s p^{n-s}\binom{(\kappa+s)+(m-s)-1}{m-s}$$
$$\times (1-q)^{\kappa+s}q^{m-s},$$

as claimed.

When $\lambda = 0$, define $\chi(t)$ and $\zeta_i(t)$ for xenologs and inparalogs:

$$\mathbb{P}\{\chi(t) = k\} = e^{-r}\frac{r^k}{k!},$$

and $\zeta_i(t)$ for $i > 0$ are Bernoulli random variables with

$$\mathbb{P}\{\zeta_i(t) = 0\} = p \qquad \mathbb{P}\{\zeta_i(t) = 1\} = 1 - p.$$

We condition on $\eta(t) = \sum_{i=1}^n\{\zeta_i(t) > 0\} = \sum_{i=1}^n \zeta_i(t)$ with the same binomial distribution as in (28): now $\xi(t) - \eta(t)$ has a Poisson distribution. $\square$

Proof of Corollary 3. By Theorem 2, the generating function for the transition probabilities is

$$G_n(z) = \sum_{m=0}^\infty w(m \mid n)z^m$$

$$= \Big(\frac{1-q}{1-qz}\Big)^\kappa\Big(p + (1-p)\frac{(1-q)z}{1-qz}\Big)^n$$

$$= \Big(\frac{1-q}{1-qz}\Big)^\kappa\Big(\frac{p + z(1-p-q)}{1-qz}\Big)^n.$$

The generating function satisfies

$$G_n(z) \times (1 - qz) = G_{n-1}(z) \times (p + z(1-p-q)).$$

Noting that $zG_n(z) = \sum_{m=1}^\infty w(m-1 \mid n)z^m$, the equality of the coefficients implies that

$$w(m \mid n) - qw(m-1 \mid n)$$
$$= pw(m \mid n-1) + (1-p-q)w(m-1 \mid n-1),$$

which gives the stated recurrence. $\square$

A.2. Proof Theorem 4

Proof. Let C_u, K_u denote the likelihoods for the empty profile:

$$C_u(n) = \mathbb{P}\Big\{\forall v \in \mathcal{L}_u : \xi_v = 0 \ \Big| \ \xi_u = n\Big\}$$

$$K_u(s) = \mathbb{P}\Big\{\forall v \in \mathcal{L}_u : \xi_v = 0 \ \Big| \ \eta_u = s\Big\}.$$

Let Q_u denote the product of $(1 - \tilde{q}_v)^{\kappa_v}$ across all edges in the subtree of u: $Q_u = 1$ at a leaf, and at an ancestral node u with children v, w

$$Q_u = \big(Q_v(1 - \tilde{q}_v)^{\kappa_v}\big)\big(Q_w(1 - \tilde{q}_w)^{\kappa_w}\big).$$

We prove that for all nodes u,

$$K_u(s) = Q_u \times (\epsilon_u)^s(1 - \tilde{q}_u)^{\kappa_u+s} \qquad C_u(n) = Q_u \times (\epsilon_u)^n.$$

(With $0^0 = 1$ and $0^n = 0$ for $n > 0$.) In particular, at the root R, the probability of the empty profile is

$$L(0) = K_R(0) = Q_R(1 - \tilde{q}_R)^{\kappa_R} = \prod_{u=1}^R\Big(\frac{1-q_u}{1-q_u\epsilon_u}\Big)^{\kappa_u}.$$

We prove the claim by induction in the node height, starting with the leaves.

Base case. At a leaf u (height 0), we have $C_u(0) = 1$ and $C_u(n) = 0$ for $n > 0$. Since $\epsilon_u = 0$, $C_u(n) = 0^n = \epsilon_u^n$ holds at all n.

At any node v, with $\kappa = \kappa_v, q = q_v, \epsilon = \epsilon_v, Q = Q_v$ and $\tilde{q} = q\frac{1-\epsilon}{1-q\epsilon}$:

$$K_v(s) = \sum_{n=s}^\infty\binom{\kappa + n - 1}{n - s}(1-q)^{\kappa+s}q^{n-s}C_v(n)$$

$$= Q\sum_{n=s}^\infty\binom{\kappa + n - 1}{n - s}(1-q)^{\kappa+s}q^{n-s}\epsilon^n$$

$$= Q\epsilon^s\Big(\frac{1-q}{1-q\epsilon}\Big)^{\kappa+s}$$

$$= Q\epsilon^s(1 - \tilde{q})^{\kappa+s}.$$

Induction. Suppose u is an ancestral node with two non-null children v and w. The height of u is $(h + 1)$ for some $h \geq 0$: suppose that the induction claim holds for all nodes at heights up to h. Both children have heights at most h, so

$$K_v(s) = Q_v(1 - \tilde{q}_v)^{\kappa_v}(\epsilon_v(1 - \tilde{q}_v))^s$$
$$K_w(s) = Q_w(1 - \tilde{q}_w)^{\kappa_w}(\epsilon_w(1 - \tilde{q}_w))^s.$$

Therefore,

$$C_u(n) = \Big(\sum_{s=0}^n\binom{n}{s}(1 - p_v)^s(p_v)^{n-s}K_v(s)\Big)$$

$$\times \Big(\sum_{s=0}^n\binom{n}{s}(1 - p_w)^s(p_w)^{n-s}K_w(s)\Big)$$

$$= \big(Q_v(1 - \tilde{q}_v)^{\kappa_v}\big)\big(p_v + (1 - p_v)\epsilon_v(1 - \tilde{q}_v)\big)^n$$

$$\times \big(Q_w(1 - \tilde{q}_w)^{\kappa_w}\big)\big(p_w + (1 - p_w)\epsilon_w(1 - \tilde{q}_w)\big)^n$$

$$= Q_u\big(\tilde{p}_v\tilde{p}_w\big)^n = Q_u(\epsilon_u)^n. \quad \square$$

A.3. Proof of Theorem 5

Proof.

(i) Given the definition of $\tilde{\eta}_u$ and $\tilde{\xi}_u$, the Pigeonhole Principle implies that their maximal value is $m_u = \sum_{v\in\mathcal{L}_u} n_v$, the sum of copy numbers at the leaves descending from u.

(ii) By Eq. (8b), the generating function for the conditional distribution of $\tilde{\xi}_u \mid \tilde{\eta}_u$ is

$$\tilde{F}_s(z) = \sum_{\ell=0}^\infty\mathbb{P}\Big\{\tilde{\xi}_u = \ell \ \Big| \ \tilde{\eta}_u = s\Big\}z^\ell$$

$$= \sum_{n=s}^\infty\binom{\kappa_u + n - 1}{n - s}(1 - q_u)^{\kappa_u+s}(q_u)^{n-s}\sum_{i=0}^{n-s}\binom{n-s}{i}$$
$$\times (1 - \epsilon_u)^i(\epsilon_u)^{n-s-i}z^{s+i}$$

$$= z^s\sum_{k=0}^\infty\binom{\kappa_u + s + k - 1}{k}(1 - q_u)^{\kappa_u+s}(q_u)^k$$
$$\times (\epsilon_u + (1 - \epsilon_u)z)^k$$

$$= z^s\Big(\frac{1 - \tilde{q}_u}{1 - \tilde{q}_u z}\Big)^{\kappa_u+s},$$

where we used $1 - \tilde{q}_u = \frac{1-q_u}{1-q_u\epsilon_u}$. Hence, $(\tilde{\xi}_u - \tilde{\eta}_u)$ has a Pólya distribution with parameters $(\kappa_u + \tilde{\eta}_u)$ and $\tilde{q}_u$:

$$\mathbb{P}\Big\{\tilde{\xi}_u = \ell \ \Big| \ \tilde{\eta}_u = s\Big\} = \binom{\kappa_i + \ell - 1}{\ell - s}(1 - \tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^{\ell-s}.$$

$$(29)$$

Now we have the recurrences for $\tilde{K}_u$:

$$\tilde{K}_u(s) = \mathbb{P}\left\{ \varXi_u \;\middle|\; \tilde{\eta}_u = s \right\}$$

$$= \sum_{\ell \geq s} \tilde{C}_u(\ell) \times \binom{\kappa_u + \ell - 1}{\ell - s}(1 - \tilde{q}_u)^{\kappa_u + s}(\tilde{q}_u)^{\ell - s},$$

as claimed.

(iii) The $\tilde{\xi}_u = \ell$ ancestral copies get sorted in the two child lineages with probabilities $(1 - \tilde{p}_v)\tilde{p}_w/(1 - \tilde{p}_v\tilde{p}_w)$, $(1 - \tilde{p}_w)\tilde{p}_v/(1 - \tilde{p}_v\tilde{p}_w)$, and $(1 - \tilde{p}_v)(1 - \tilde{p}_w)/(1 - \tilde{p}_v\tilde{p}_w)$ as surviving only on the left v, only on the right w, or on both sides. Hence,

$$\mathbb{P}\left\{ \tilde{\eta}_v = s \;\middle|\; \tilde{\xi}_u = \ell \right\} = \binom{\ell}{s}\left(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^s\left(\tilde{p}_v\frac{1 - \tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell - s} \tag{30}$$

$$\mathbb{P}\left\{ \tilde{\eta}_w = s \;\middle|\; \tilde{\xi}_u = \ell \right\} = \binom{\ell}{s}\left(\frac{1 - \tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^s\left(\tilde{p}_w\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell - s}$$

for $0 \leq s \leq \ell$. Define $\tilde{\psi}_u$ as the ancestral copies from $\tilde{\xi}_u$ that survive in both child lineages:

$$\mathbb{P}\left\{ \tilde{\psi}_u = k \;\middle|\; \tilde{\eta}_v = s \right\} = \binom{s}{k}(1 - \tilde{p}_w)^k(\tilde{p}_w)^{s-k} \tag{31}$$

$$\mathbb{P}\left\{ \tilde{\psi}_u = k \;\middle|\; \tilde{\eta}_w = s \right\} = \binom{s}{k}(1 - \tilde{p}_v)^k(\tilde{p}_v)^{s-k}$$

for $0 \leq k \leq s$. The two random variables $\tilde{\eta}_v, \tilde{\eta}_w$ are not independent when conditioned on $\tilde{\xi}_u$, since $\tilde{\eta}_w = \tilde{\xi}_u - \tilde{\eta}_v + \tilde{\psi}_u$:

$$\mathbb{P}\left\{ \tilde{\eta}_v = s, \tilde{\eta}_w = t \;\middle|\; \tilde{\xi}_u = \ell \right\}$$

$$= \mathbb{P}\left\{ \tilde{\eta}_v = s \;\middle|\; \tilde{\xi}_u = \ell \right\}\mathbb{P}\left\{ \tilde{\psi}_u = (s+t) - \ell \;\middle|\; \tilde{\eta}_v = s \right\}.$$

Combining (30) and (31) gives us the recurrence for $\tilde{C}_u$:

$$\tilde{C}_u(\ell) = \mathbb{P}\left\{ \varXi_u \;\middle|\; \tilde{\xi}_u = \ell \right\}$$

$$= \sum_{\substack{s+t \geq \ell}}^{s,t \leq \ell} \mathbb{P}\left\{ \varXi_v \;\middle|\; \tilde{\eta}_v = s \right\}$$

$$\times \mathbb{P}\left\{ \varXi_w \;\middle|\; \tilde{\eta}_w = t \right\}\mathbb{P}\left\{ \tilde{\eta}_v = s, \tilde{\eta}_w = t \;\middle|\; \tilde{\xi}_u = \ell \right\}$$

$$= \sum_{s=0}^{\ell}\left(\tilde{K}_v(s) \times \binom{\ell}{s}\left(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^s\left(\tilde{p}_v\frac{1 - \tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell - s}\right.$$

$$\left.\times \sum_{k=0}^{s}\tilde{K}_w(\ell - s + k) \times \binom{s}{k}(1 - \tilde{p}_w)^k(\tilde{p}_w)^{s-k}\right).$$

The inner sum can be computed in $O(1)$ amortized time by dynamic programming: for all $0 \leq d \leq \ell$, let

$$\tilde{K}_w^{\ell}(d) = \sum_{k=0}^{\ell - d}\tilde{K}_w(d+k) \times \binom{\ell - d}{k}(1 - \tilde{p})^k\tilde{p}^{\ell - d - k},$$

with $\tilde{p} = \tilde{p}_w$, so that

$$\tilde{C}_u(\ell) = \sum_{s=0}^{\ell}\tilde{K}_v(s) \times \tilde{K}_w^{\ell}(\ell - s)$$

$$\times \binom{\ell}{s}\left(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^s\left(\frac{\tilde{p}_v - \tilde{p}_v\tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell - s}$$

The initial values are

$$\tilde{K}_w^{\ell}(\ell) = \tilde{K}_w(\ell).$$

Let $s = \ell - d$. Since

$$\binom{s}{k}(1 - \tilde{p})^k\tilde{p}^{s-k}$$

$$= \{k < s\}\tilde{p}\binom{s-1}{k}(1 - \tilde{p})^k\tilde{p}^{(s-1)-k}$$

$$+ \{0 < k\}(1 - \tilde{p})\binom{s-1}{k-1}(1 - \tilde{p})^{k-1}\tilde{p}^{(s-1)-(k-1)},$$

we have the recursions for $d < \ell$:

$$\tilde{K}_w^{\ell}(d)$$

$$= \sum_{k=0}^{\ell - d}\tilde{K}_w(d+k) \times \binom{\ell - d}{k}(1 - \tilde{p})^k\tilde{p}^{\ell - d - k}$$

$$= \tilde{p}\sum_{k=0}^{\ell - d - 1}\tilde{K}_w(d+k)\binom{\ell - d - 1}{k}(1 - \tilde{p})^k\tilde{p}^{\ell - d - 1 - k}$$

$$+ (1 - \tilde{p})\sum_{k=1}^{\ell - d}\tilde{K}_w(d+k)\binom{\ell - d - 1}{k - 1}$$

$$\times (1 - \tilde{p})^{k-1}\tilde{p}^{(\ell - d - 1) - (k-1)}.$$

By setting $d + k = (d+1) + (k-1)$ in the second term,

$$\tilde{K}_w^{\ell}(d) = \tilde{p}\tilde{K}_w^{\ell-1}(d) + (1 - \tilde{p})\tilde{K}_w^{\ell}(d+1). \quad \square$$

A.4. Proof of Theorem 7

Proof. At the root, $\tilde{J}_R(s) = \mathbb{P}\{\tilde{\eta}_R = s\} = \{s = 0\}$ by our model. Let u be an arbitrary node, and let $\kappa = \kappa_u, \tilde{q} = \tilde{q}_u$. Using Eq. (29) for $\tilde{\xi}_u|\tilde{\eta}_u$,

$$\tilde{B}_u(\ell) = \mathbb{P}\{\varXi - \varXi_u, \tilde{\xi}_u = \ell\}$$

$$= \sum_s \mathbb{P}\left\{ \varXi - \varXi_u, \tilde{\xi}_u = \ell \;\middle|\; \tilde{\eta}_u = s \right\}\mathbb{P}\{\tilde{\eta}_u = s\}$$

$$= \sum_s \mathbb{P}\{\varXi - \varXi_u, \tilde{\eta}_u = s\}\mathbb{P}\left\{ \tilde{\xi}_u = \ell \;\middle|\; \tilde{\eta}_u = s \right\}$$

$$= \sum_s \tilde{J}_u(s)\binom{\kappa + \ell - 1}{\ell - s}(1 - \tilde{q})^{\kappa + s}(\tilde{q})^{\ell - s}.$$

Now let $uv \in T$ be a parent–child pair, and let $uw \in T$ be the sibling lineage (with $v \neq w$). Since $\varXi - \varXi_v = (\varXi - \varXi_u) \cup \varXi_w$,

$$\tilde{J}_v(s) = \mathbb{P}\{\varXi - \varXi_v, \tilde{\eta}_v = s\}$$

$$= \sum_{\ell} \mathbb{P}\left\{ \varXi - \varXi_v, \tilde{\eta}_v = s \;\middle|\; \tilde{\xi}_u = \ell \right\}\mathbb{P}\{\tilde{\xi}_u = \ell\}$$

$$= \sum_{\ell} \mathbb{P}\{\varXi - \varXi_u, \tilde{\xi}_u = \ell\} \times \mathbb{P}\left\{ \tilde{\eta}_v = s \;\middle|\; \tilde{\xi}_u = \ell \right\}$$

$$\times \sum_t \mathbb{P}\left\{ \tilde{\eta}_w = t \;\middle|\; \tilde{\xi}_u = \ell, \tilde{\eta}_v = s \right\}\mathbb{P}\left\{ \varXi_w \;\middle|\; \tilde{\eta}_w = t \right\}$$

$$= \sum_{\ell} \tilde{B}_u(\ell) \times \binom{\ell}{s}\left(\frac{1 - \tilde{p}_v}{1 - \tilde{p}_v\tilde{p}_w}\right)^s\left(\frac{\tilde{p}_v - \tilde{p}_v\tilde{p}_w}{1 - \tilde{p}_v\tilde{p}_w}\right)^{\ell - s}$$

$$\times \tilde{K}_w^{\ell}(\ell - s),$$

where we used Eqs. (30) and (31) as in the proof of Theorem 5. $\square$

A.5. Proof of Theorem 9

Proof. First, define ϵ_u at all nodes using $\tilde{p}$: if u is a leaf, then $\epsilon_u = 0$, and at an ancestral node u, $\epsilon_u = \prod_{uv \in T}\tilde{p}_v$. Since all $\tilde{p}_u$ are positive, $\epsilon_u > 0$ at every ancestral node u.

Let u be an arbitrary node and let $\tilde{q} = \tilde{q}_u, \tilde{p} = \tilde{p}_u$. Since $0 < (1-\tilde{q})\epsilon_u < 1$, the equation $\tilde{q} = q_u \frac{1-\epsilon_u}{1-q_u\epsilon_u}$ has a unique positive solution

$$q_u = \frac{\tilde{q}}{1-(1-\tilde{q})\epsilon_u} = \frac{\tilde{q}}{\tilde{q}+(1-\tilde{q})(1-\epsilon_u)} < 1.$$

Furthermore, the equation $\tilde{p} = p_u + (1-p_u)\epsilon_u(1-\tilde{q})$ has a unique solution

$$p_u = \frac{\tilde{p}-\epsilon_u(1-\tilde{q})}{1-\epsilon_u(1-\tilde{q})} < 1.$$

Since $\epsilon_u = \prod_{uv \in T} \tilde{p}_v$, by the assumption of (22), $p_u > 0$. If the assumption is violated by $\tilde{p} < \epsilon_u(1-\tilde{q})$, then $p_u < 0$, which is illegal.

Since $0 < p_u < 1$ and $0 < q_u < 1$ can be selected at every node, Theorem 1 implies that a corresponding phylogenetic birth–death model exists that is unique up to equivalent rate scalings. $\square$

A.6. Proof of Theorem 10

Proof. By Eqs. (20) and (18),

$$\frac{\partial L(\varXi)}{\partial \tilde{q}_u}$$
$$= \sum_{\ell=0}^{m_u} \tilde{C}_u(\ell)\frac{\partial \tilde{B}_u(\ell)}{\partial \tilde{q}_u}$$
$$= \sum_{\ell=0}^{m_u} \tilde{C}_u(\ell)\frac{\partial}{\partial \tilde{q}_u}\left(\sum_{s=0}^{\ell}\tilde{J}_u(s)\times\binom{\kappa_u+\ell-1}{\ell-s}(1-\tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^{\ell-s}\right).$$

So,

$$\frac{\partial L(\varXi)}{\partial \tilde{q}_u}$$
$$= \sum_{0 \le s \le \ell \le m_u}\tilde{C}_u(\ell)\times\tilde{J}_u(s)$$
$$\times\binom{\kappa_u+\ell-1}{\ell-s}(1-\tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^{\ell-s}\left(\frac{\ell-s}{\tilde{q}_u}-\frac{\kappa_u+s}{1-\tilde{q}_u}\right). \quad (32)$$

By Theorem 4, the empty profile likelihood is $L(0) = \prod_{u=1}^{R}(1-\tilde{q}_u)^{\kappa_u}$, so

$$\frac{\partial L(0)}{\partial \tilde{q}_u} = -L(0)\frac{\kappa_u}{1-\tilde{q}_u}.$$

For derivatives with respect to $\tilde{p}_v$ on an edge between a non-root node v and its parent $uv \in T$, consider the recurrences of Theorems 5 and 6. Both can be written as

$$\tilde{C}_u(\ell) = \sum_{s=0}^{\ell}\tilde{K}_v(s)\times\tilde{K}^\ell_{-v}(\ell-s)\times\binom{\ell}{s}\left(\frac{1-\tilde{p}_v}{1-\tilde{p}_v\epsilon}\right)^s\left(\frac{\tilde{p}_v-\tilde{p}_v\epsilon}{1-\tilde{p}_v\epsilon}\right)^{\ell-s}.$$

At a binary node u (Theorem 5), $\epsilon = \tilde{p}_w$ with the sibling $uw \in T$, and $\tilde{K}^\ell_{-v}(k) = \tilde{K}^\ell_w(k)$. If u has more than 2 children $v_1, \dots, v_d$, then order them so that v is the first, and apply Theorem 6: $\epsilon = \epsilon_{u,-2} = \prod_{j=2}^{d}\tilde{p}_{v_j}$ and $\tilde{K}^\ell_{-v}(k) = \tilde{K}^\ell_{v_{2..d}}(k)$ from Eq. (16b). Hence, using Corollary 8,

$$\frac{\partial L(\varXi)}{\partial \tilde{p}_v} = \sum_{\ell=0}^{m_u}\tilde{B}_u(\ell)\frac{\partial \tilde{C}_u(\ell)}{\partial \tilde{p}_v}$$
$$= \sum_{\ell=0}^{m_u}\tilde{B}_u(\ell)\sum_{s=0}^{\min\{\ell,m_u\}}\tilde{K}_v(s)\times\tilde{K}^\ell_{-v}(\ell-s)$$
$$\times\frac{\partial}{\partial \tilde{p}_v}\left(\binom{\ell}{s}\left(\frac{1-\tilde{p}_v}{1-\tilde{p}_v\epsilon}\right)^s\left(\frac{\tilde{p}_v-\tilde{p}_v\epsilon}{1-\tilde{p}_v\epsilon}\right)^{\ell-s}\right).$$

Therefore,

$$\frac{\partial L(\varXi)}{\partial \tilde{p}_v}$$
$$= \sum_{0 \le s \le \ell \le m_u}\tilde{B}_u(\ell)\times\tilde{K}_v(s)\times\tilde{K}^\ell_{-v}(\ell-s)$$
$$\times\binom{\ell}{s}\left(\frac{1-\tilde{p}_v}{1-\tilde{p}_v\epsilon}\right)^s\left(\frac{\tilde{p}_v-\tilde{p}_v\epsilon}{1-\tilde{p}_v\epsilon}\right)^{\ell-s}\left(\frac{\ell-s}{\tilde{p}_v}-\frac{s(1-\epsilon)}{1-\tilde{p}_v}\right). \quad (33)$$

The derivatives for the empty profile likelihood are trivial, since $L(0)$ does not depend on any of the $\tilde{p}_v$.

By Corollary 8 and Theorem 7,

$$\frac{\partial L(\varXi)}{\partial \kappa_u}$$
$$= \sum_{\ell=0}^{m_u}\tilde{C}_u(\ell)\sum_{s=0}^{\ell}\tilde{J}_u(s)\frac{\partial}{\partial \kappa_u}\left((1-\tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^{\ell-s}\binom{\kappa_u+\ell-1}{\ell-s}\right).$$

Since

$$\frac{\partial}{\partial \kappa}\left(\ln\binom{\kappa+\ell-1}{\ell-s}\right) = \frac{\frac{\partial}{\partial \kappa}\binom{\kappa+\ell-1}{\ell-s}}{\binom{\kappa+\ell-1}{\ell-s}},$$

and

$$\frac{\partial}{\partial \kappa}\left(\ln\binom{\kappa+\ell-1}{\ell-s}\right) = \sum_{i=0}^{\ell-s-1}\frac{\partial\ln(\kappa+s+i)}{\partial \kappa} = \sum_{i=s}^{\ell-1}\frac{1}{\kappa+i},$$

we have

$$\frac{\partial L(\varXi)}{\partial \kappa_u}$$
$$= \ln(1-\tilde{q}_u)\times L(\varXi)$$
$$+ \sum_{0 \le s \le \ell \le m_u}\tilde{C}_u(\ell)\times\tilde{J}_u(s)$$
$$\times\binom{\kappa_u+\ell-1}{\ell-s}(1-\tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^{\ell-s}\left(\sum_{i=s}^{\ell-1}\frac{1}{\kappa_u+i}\right).$$
$$= \ln(1-\tilde{q}_u)\times L(\varXi) + \sum_{i=0}^{m_u-1}\frac{1}{\kappa_u+i}\sum_{s=0}^{i}\sum_{\ell=i+1}^{m_u}\tilde{C}_u(\ell)\times\tilde{J}_u(s)$$
$$\times\binom{\kappa_u+\ell-1}{\ell-s}(1-\tilde{q}_u)^{\kappa_u+s}(\tilde{q}_u)^{\ell-s}$$
$$= \ln(1-\tilde{q}_u)\times L(\varXi)$$
$$+ \sum_{i=0}^{m_u-1}\frac{1}{\kappa_u+i}\left(\mathbb{P}\{\tilde{\xi}_u > i; \varXi\}-\mathbb{P}\{\tilde{\eta}_u > i; \varXi\}\right).$$

For the empty profile,

$$\frac{\partial L(0)}{\partial \kappa_u} = \frac{\partial}{\partial \kappa_u}\left(\prod_{v=1}^{R}(1-\tilde{q}_v)^{\kappa_v}\right) = L(0)\times\ln(1-\tilde{q}_u) \quad \square$$

A.7. Proof of Theorem 12

Proof. Let $u_0 u_1 \cdots u_{d-1}$ denote the path between $u_d = v$ and the root $u_0 = R$ with edges $u_i u_{i+1} \in T$. Since p_v and q_v influence $\tilde{p}_u$ and $\tilde{q}_u$ at $u = v$ and at all the other ancestors $u = u_i$, but not at any other node,

$$\varPhi^{(\theta_v)} = \sum_{i=0}^{d}\left(\varPhi^{(\tilde{q}_{u_i})}\frac{\partial \tilde{q}_{u_i}}{\partial \theta_v}+\varPhi^{(\tilde{p}_{u_i})}\frac{\partial \tilde{p}_{u_i}}{\partial \theta_v}\right).$$

Recall the definitions $\tilde{q}_v = q_v\frac{1-\epsilon_v}{1-q_v\epsilon_v}$ and $\tilde{p}_v = \frac{p_v(1-\epsilon_v)+\epsilon_v(1-q_v)}{1-q_v\epsilon_v}$ (substituting $p_R = 0$ at the root). We have thus

$$\frac{\partial \tilde{p}_v}{\partial p_v} = \frac{1-\epsilon_v}{1-q_v\epsilon_v} \qquad\qquad \frac{\partial \tilde{q}_v}{\partial p_v} = 0$$

$$\frac{\partial \tilde{p}_v}{\partial q_v} = \frac{-(1-p_v)\epsilon_v(1-\epsilon_v)}{(1-q_v\epsilon_v)^2} \qquad \frac{\partial \tilde{q}_v}{\partial q_v} = \frac{1-\epsilon_v}{(1-q_v\epsilon_v)^2}$$

$$\frac{\partial \tilde{p}_v}{\partial \epsilon_v} = \frac{(1-p_v)(1-q_v)}{(1-q_v\epsilon_v)^2} \qquad \frac{\partial \tilde{q}_v}{\partial \epsilon_v} = \frac{-q_v(1-q_v)}{(1-q_v\epsilon_v)^2}.$$

(i) If $v = R$ is the root, then

$$\Phi^{(q_R)} = \Phi^{(\tilde{p}_R)}\frac{\partial \tilde{p}_R}{\partial q_R} + \Phi^{(\tilde{q}_R)}\frac{\partial \tilde{q}_R}{\partial q_R}$$
$$= -\Phi^{(\tilde{p}_R)}\frac{\epsilon_R(1-\epsilon_R)}{(1-q_R\epsilon_R)^2} + \Phi^{(\tilde{q}_R)}\frac{1-\epsilon_R}{(1-q_R\epsilon_R)^2},$$

and, for $R > 1$,

$$\Phi^{(\epsilon_R)} = \Phi^{(\tilde{p}_R)}\frac{\partial \tilde{p}_R}{\partial \epsilon_R} + \Phi^{(\tilde{q}_R)}\frac{\partial \tilde{q}_R}{\partial \epsilon_R}$$
$$= \Phi^{(\tilde{p}_R)}\frac{1-q_R}{(1-q_R\epsilon_R)^2} - \Phi^{(\tilde{q}_R)}\frac{q_R(1-q_R)}{(1-q_R\epsilon_R)^2},$$

as claimed in (23).

(ii) Now suppose that v is not the root. At any ancestor u_i with $i < d$, the distribution parameters of v affect the extinction probability ϵ_{u_i}. For a distribution parameter $\theta_v = p_v$, $\theta_v = q_v$, or $\theta_v = \epsilon_v$,

$$\frac{\partial \tilde{p}_{u_i}}{\partial \theta_v} = \frac{\partial \tilde{p}_{u_i}}{\partial \epsilon_{u_i}}\frac{\partial \epsilon_{u_i}}{\partial \tilde{p}_{u_{i+1}}}\frac{\partial \tilde{p}_{u_{i+1}}}{\partial \theta_v} \quad \text{and} \quad \frac{\partial \tilde{q}_{u_i}}{\partial \theta_v} = \frac{\partial \tilde{q}_{u_i}}{\partial \epsilon_{u_i}}\frac{\partial \epsilon_{u_i}}{\partial \tilde{p}_{u_{i+1}}}\frac{\partial \tilde{p}_{u_{i+1}}}{\partial \theta_v},$$

with

$$\frac{\partial \epsilon_{u_i}}{\partial \tilde{p}_{u_{i+1}}} = \frac{\partial}{\partial \tilde{p}_{u_{i+1}}}\prod_{u_iw\in T}\tilde{p}_w = \frac{\epsilon_{u_i}}{\tilde{p}_{u_{i+1}}}.$$

Let $u = u_{d-1}$ be the parent of $v = u_d$. Since $\frac{\partial \tilde{q}_v}{p_v} = 0$,

$$\Phi^{(p_v)} = \left(\Phi^{(\tilde{p}_v)} + \Phi^{(\epsilon_u)}\frac{\partial \epsilon_u}{\partial \tilde{p}_v}\right)\frac{\partial \tilde{p}_v}{\partial p_v} = \left(\Phi^{(\tilde{p}_v)} + \epsilon\Phi^{(\epsilon_u)}\right)\frac{1-\epsilon_v}{1-q_v\epsilon_v}$$

with $\epsilon = \frac{\epsilon_u}{\tilde{p}_v} = \frac{\epsilon_{u_{d-1}}}{\tilde{p}_{u_d}}$. The other two recurrences include $\Phi^{(\tilde{q}_v)}$, as well:

$$\Phi^{(q_v)} = \left(\Phi^{(\tilde{p}_v)} + \Phi^{(\epsilon_u)}\frac{\partial \epsilon_u}{\partial \tilde{p}_v}\right)\frac{\partial \tilde{p}_v}{\partial q_v} + \Phi^{(\tilde{q}_v)}\frac{\partial \tilde{q}_v}{\partial q_v}$$
$$= \left(\Phi^{(\tilde{p}_v)} + \epsilon\Phi^{(\epsilon_u)}\right)\frac{-(1-p_v)\epsilon_v(1-\epsilon_v)}{(1-q_v\epsilon_v)^2}$$
$$+ \Phi^{(\tilde{q}_v)}\frac{1-\epsilon_v}{(1-q_v\epsilon_v)^2};$$

$$\Phi^{(\epsilon_v)} = \left(\Phi^{(\tilde{p}_v)} + \epsilon\Phi^{(\epsilon_u)}\right)\frac{\partial \tilde{p}_v}{\partial \epsilon_v} + \Phi^{(\tilde{q}_v)}\frac{\partial \tilde{q}_v}{\partial \epsilon_v}$$
$$= \left(\Phi^{(\tilde{p}_v)} + \epsilon\Phi^{(\epsilon_u)}\right)\frac{(1-p_v)(1-q_v)}{(1-q_v\epsilon_v)^2}$$
$$- \Phi^{(\tilde{q}_v)}\frac{q_v(1-q_v)}{(1-q_v\epsilon_v)^2},$$

as shown in (24). □

A.8. No-duplication model

Proof of Theorem 13. Let C_u, K_u denote the likelihoods for the empty profile:

$$C_u(n) = \mathbb{P}\left\{\forall v \in \mathcal{L}_u : \xi_v = 0 \mid \xi_u = n\right\}$$
$$K_u(s) = \mathbb{P}\left\{\forall v \in \mathcal{L}_u : \xi_v = 0 \mid \eta_u = s\right\}.$$

Let Q_u denote the product of $e^{-\tilde{r}_v}$ across all edges in the subtree of u: $Q_u = 1$ at a leaf, and at an ancestral node u with children v, w

$$Q_u = \left(Q_v e^{-\tilde{r}_v}\right)\left(Q_w e^{-\tilde{r}_w}\right).$$

We prove that for all nodes u,

$$K_u(s) = Q_u \times (\epsilon_u)^s \times e^{-\tilde{r}_u} \qquad C_u(n) = Q_u \times (\epsilon_u)^n.$$

(With $0^0 = 1$ and $0^n = 0$ for $n > 0$.) In particular, at the root R, $L(0) = \tilde{K}_R(0)$.

We adjust the induction proof of Theorem 4. At any node u, with $r = r_u$, $\epsilon = \epsilon_u$ and $Q = Q_u$, by Eq. (25),

$$K_u(s) = \sum_{k=0}^{\infty}e^{-r}\frac{r^k}{k!}Q\epsilon^{s+k} = Q\epsilon^se^{-r}\sum_{k=0}^{\infty}\frac{(r\epsilon)^k}{k!}$$
$$= Q\epsilon^se^{-r(1-\epsilon)} = Q\epsilon^se^{-\tilde{r}}$$

with $\tilde{r} = r(1-\epsilon)$. The inductive case for C_u is adjusted:

$$C_u(n) = \prod_{uv\in T}\left(Q_ve^{-\tilde{r}_v}\left(p_v + (1-p_v)\epsilon_v\right)^n\right) = Q_u(\epsilon_u)^n. \quad \square$$

Proof of Theorem 14. Since $\tilde{p}_u = p_u + (1-p_u)\epsilon_u$ and $\tilde{r}_u = r_u(1-\epsilon_u)$,

$$\frac{\partial \tilde{p}_u}{\partial p_u} = 1 - \epsilon_u \qquad\qquad \frac{\partial \tilde{r}_u}{\partial r_u} = 1 - \epsilon_u$$
$$\frac{\partial \tilde{p}_u}{\partial \epsilon_u} = 1 - p_u \qquad\qquad \frac{\partial \tilde{r}_u}{\partial \epsilon_u} = -r_u.$$

The rest of the proof is based on applications of the chain rule as in the proof of Theorem 12. □

A.9. Proof of Theorem 15

Proof. Let $N = \sum_{w\in\mathcal{L}}n_w = m_R$ be the sum of copy numbers across the leaves. At an ancestral node u, the calculations of $\tilde{C}_u(\ell)$ for all $0 \le \ell \le m_u$ and of $\tilde{J}_u(s)$ for all $0 \le s \le m_u$ take $(1+m_v)(1+m_w)$ iterations. Calculating $\tilde{K}(s)$ for all $0 \le s \le m_u$ and $\tilde{B}_u(\ell)$ for all $0 \le \ell \le m_u$ is done in $(1+m_u)(2+m_u)/2$ iterations. The total running time can be thus bounded asymptotically as $O\left(\sum_{u=1}^{R}m_u^2\right)$, or as $O(R) = O(L)$ if $N^2 < R$, the number of nodes. Summing by the height of the nodes $h(u)$,

$$\sum_{u=1}^{R}m_u^2 = \sum_{i=0}^{h-1}\sum_{u:h(u)=i}(m_u)^2 = \sum_{i=0}^{h-1}\sum_{u:h(u)=i}\left(\sum_{v\in\mathcal{L}_u}n_v\right)^2$$
$$\le \sum_{i=0}^{h-1}\left(\sum_{u:h(u)=i}\sum_{v\in\mathcal{L}_u}n_v\right)^2$$
$$\le \sum_{i=0}^{h-1}\left(\sum_{v\in\mathcal{L}}n_v\right)^2 = hN^2.$$

For the last inequality, note that if $h(v) = h(w)$ then their subtrees do not intersect and $\mathcal{L}_v \cap \mathcal{L}_w = \emptyset$. □

References

Aldous, D., 1996. Probability distributions on cladograms. In: Random Discrete Structures. Springer, pp. 1–18.

Ames, R.M., oney, D., GHatge, V.P., Whelan, S., Lovell, S.C., 2012. Deterining the evolutionary history of gene families. Bioinformatics 28, 48–55.

Aziz, R.K., et al., 2008. The RAST server: Rapid annotations using subsystems technology. BMC Genomics 9 (75), http://dx.doi.org/10.1186/1471-2164-9-75.

Crawford, F.W., Minin, V.N., M.A, Suchard, 2014. Estimation for general birth–death processes. J. Amer. Statist. Assoc. 109, 730–747.

Csűrös, M., 2010. Count: Evolutionary analysis of phylogenetic profiles with parsimony and likelihood. Bioinformatics 26 (15), 1910–1912. http://dx.doi.org/10.1093/bioinformatics/btq315.

Csűrös, M., Miklós, I., 2006. A probabilistic model for gene content evolution with duplication, loss, and horizontal transfer. In: Proc. Tenth Annual International Conference on Research in Computational Molecular Biology. RECOMB, In: Springer Lecture Notes in Bioinformatics, vol. 3909, pp. 206–220. http://dx.doi.org/10.1007/11732990_18.

Csűrös, M., Miklós, I., 2009. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. Mol. Biol. Evol. 26 (9), 2087–2095. http://dx.doi.org/10.1093/molbev/msp123.

Dey, G., Meyer, T., 2015. Phylogenetic profiling for probing the modular architecture of the human genome. Cell Syst. 1, 106–115. http://dx.doi.org/10.1016/j.cels.2015.08.006.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Syst. Zool. 22 (3), 240–249.

Felsenstein, J., 1992. Phylogenies from restriction sites, a maximum likelihood approach. Evolution 46, 159–173.

Fitch, W.M., 2000. Homology a personal view on some of the problems. TIG 16 (5), 227–231.

Fukunaga, T., Iwasaki, W., 2021. Mirage: Estimation of ancestral gene-copy numbers by considering different evolutionary patterns among gene families. Bioinform. Adv. 1, 1–20. http://dx.doi.org/10.1093/bioadv/vbab014.

Hahn, M.W., Bie, T.De., Stajich, J.E., Nguyen, C., Cristianini, N., 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. Genome Res. 15, 1153–1160. http://dx.doi.org/10.1101/gr.3567505.

Iwasaki, W., Takagi, T., 2007. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. Bioinformatics 23 (13), i230–i239. http://dx.doi.org/10.1093/bioinformatics/btm165.

Karev, G.P., Wolf, Y.I., Berezovskaya, F.S., Koonin, E.V., 2004. Gene family evolution: An in-depth theoretical and simulation analysis of non-linear birth–death-innovation models. BMC Evol. Biol. 4 (32), http://dx.doi.org/10.1186/1471-2148-4-32.

Karlin, S., McGregor, J., 1958. Linear growth birth, and death processes. J. Math. Mech. 7 (4), 643–662.

Kendall, D.G., 1949. Stochastic processes and population growth. J. R. Stat. Soc. Ser. B Stat. Methodol. 11 (2), 230–282.

Kruglyak, S., Durrett, R.T., Schig, M.D., Aquadro, C.F., 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc. Natl. Acad. Sci. USA 95, 10774–10778. http://dx.doi.org/10.1073/pnas.95.18.10774.

Leguesdron, P., Pellaumail, J., Rubino, G., Sericola, B., 1993. Transient analysis of the M/M/1 queue. Adv. Appl. Probab. 25 (3), 702–713.

Mayrose, I., Barker, M.S., Otto, S.P., 2010. Probabilistic models of chromosome number evolution and the inference of polyploidy. Syst. Biol. 59, 132–144.

Nee, S., May, R.M., Harvey, P.H., 1994. The reconstructed evolutionary process. Philos. Trans. R. Soc. London. Series B 344, 305–311.

Novozhilov, A.S., Karev, G., Koonin, E.V., 2006. Biological applications of the theory of birth-and-death processes. Brief. Bioinform. 7, 70–85. http://dx.doi.org/10.1093/bib/bbk006.

Nye, T.M.W., 2009. Modelling the evolution of multi-gene families. Stat. Methods Med. Res. 18, 487–504. http://dx.doi.org/10.1177/0962280208099450.

Rosenberg, N.A., Tsolaki, A.G., Tanaka, M.M., 2003. Estimating change rates of genetic markers using serial samples: Applications to the transposon IS6110 in mycobacterium tuberculosis. Theor. Popul. Biol. 63, 347–363.

Sela, I., Wolf, Y.I., Koonin, E.V., 2016. Theory of prokaryotic evolution. Proc. Natl. Acad. Sci. USA 113, 11,399–11,407. http://dx.doi.org/10.1073/pnas.1614083113.

Sonnhammer, E.L.L., Koonin, E.V., 2002. Orthology, paralogy and proposed classification for paralog subtypes. TIG 18 (12), 619–620.

Spencer, M., Susko, E., Roger, A.J., 2006. Modelling prokaryote gene content. Evol. Bioinform. Online 2, 165–186.

Stadler, T., 2009. On incomplete sampling under birth–death models and connections to the sampling-based coalescent. J. Theoret. Biol. 261, 58–66. http://dx.doi.org/10.1016/j.jtbi.2009.07.018.

Takács, L., 1962. Introduction to the Theory of Queues. Oxford University Press, New York.

Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A., 2006. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173, 1511–1520.

Tavaré, S., 1989. The genealogy of the birth, death, and immigration process. In: Feldman, M.W. (Ed.), Mathematical Evolutionary Theory. Princeton University Press, http://dx.doi.org/10.1515/9781400859832-004 (Chapter 3).

Thorne, J.L., Kishino, H., Felsenstein, J., 1991. An evolutionary model for maximum likelihod alignment of DNA sequences. J. Mol. Evol. 33, 114–124.

Xu, J.X., Guttorp, P., Kato-Maeda, M., VN, Minin, 2015. Likelihood-based inference for discretely observed birth–death-shift processes, with applications to evolution of mobile genetic elements. Biometrics 71, 1009–1021. http://dx.doi.org/10.1111/biom.12352.