Detecting LGTs using a novel probabilistic model integrating duplications, LGTs, losses, rate variation, and sequence evolution

A. Tofigh, J. Sjöstrand, B. Sennblad, L. Arvestad, and J. Lagergren

Abstract

The debate over the prevalence of lateral gene transfers (LGTs) has been intense. There is now to a large extent consensus around the view that LGT is an important evolutionary force as well as regarding its relative importance across species. This consensus relies, however, mainly on studies of individual gene families.

Up until now, the gold standard for identifying LGTs has been phylogenetic methods where LGTs are inferred from incongruities between a species tree and an associated gene tree. Even in cases where there is evidence of LGT, several concerns have often been raised regarding the significance of the evidence. One common concern has been the possibility that other evolutionary events have caused the incongruities. Another has been the significance of the gene trees involved in the inference; there may for instance be alternative, almost equally likely, gene trees that do not provide evidence for LGT. Independently of these concerns, there has been a need for methods that can be used to quantitatively characterize the level of LGT among sets of species, but also for methods able to pinpoint where in the species tree LGTs have occurred.

Here, we provide the first probabilistic model capturing gene duplication, LGT, gene loss, and point mutations with a relaxed molecular clock. We also provide all fundamental algorithms required to analyze a gene family relative to a given species tree under this model. Our algorithms are based on Markov chain Monte Carlo (MCMC) methodology but build also on techniques from numerical analysis and involve dynamic programming (DP).

1 Introduction

The importance and prevalence of lateral gene transfers (LGT) have been debated intensely. The interest in LGT is partly explained by its capacity to transfer pathogenic elements and antibiotic resistance between bacteria, but also the concern that it could transfer, e.g., pesticide resistance from genetically modified crops to other plants.

Transformation, transduction, and conjugation are the mechanisms through which lateral gene transfer can be accomplished among bacteria. Especially transformation has played a pivotal role in several ground-breaking biological experiments. Although DNA had not yet been identified as the carrier of genetic information, pneumococcal strains were

observed to be possible to transform by Griffith in 1928 [16]. Later, the Avery-MacLeod-McCarty experiment showed that DNA is the substance causing bacterial transformation. The possibility of lateral gene transfers among bacteria was realized already in 1946 [27, 28] and demonstrated to occur between different bacterial species in 1959 [31]. A number of studies have established that LGT occurs among prokaryotes, see for example [32] and [4]. Evidence has also been presented for the occurrence of lateral gene transfers from prokaryotes to eukaryotes and even between eukaryotes, see [23] for a recent review.

There has been an intense debate concerning the relative benefits of different methods for phylogenetic tree reconstruction. Today, however, it is common to describe the development of phylogeny algorithms as a progression starting in 1965 with parsimony methods [5, 26, 12], continuing with Maximum Likelihood (ML) methods introduced by Felsenstein [11], and where the most recent contribution is Bayesian methods [21].

The first phylogenetic incongruence methods were constructed to identify gene duplications based on the parsimony principle. Goodman *et al.* [15] pioneered the field by introducing the term *reconciliation* for an embedding of a gene tree into a species tree explaining the evolution of the former. In later contributions, parsimony-based phylogenetic incongruence methods have been described for LGT alone [19], but also for the combination of gene duplications and LGT [18]. The application of Kishino-Hasegawa tests in [30] is another example of a phylogenetic incongruence method for LGT.

The statistical significance of the investigated phylogenetic trees has been a common concern in the context of phylogenetic incongruence methods. Recently, partly prompted by such concerns, Bayesian phylogenetic incongruence methods were developed for duplication analysis. In [37], the GSR model was presented; it is a probabilistic model integrating gene duplication, sequence evolution, and a relaxed molecular clock for substitution rates. Based on the GSR model and using Markov Chain Monte Carlo (MCMC) methodology, a Bayesian analysis tool, PrIME-GSR, was constructed, which takes a known species tree into account and performs simultaneous gene tree reconstruction and reconciliation.

The extreme view that LGT hardly exists (implying that discrepancies between gene and species trees are due to random effects or to insufficiently sophisticated tree reconstruction methods, or possibly due to other events such as duplications) has lost most of its supporters, and instead, LGT is recognized as a major evolutionary force. In fact, due to its prevalence among prokaryotes, the appropriateness of using trees to represent the evolution of some sets of species has been questioned [14, 7, 36], see also [8] and references therein. Here, we will adopt an intermediate view that has emerged in recent years with respect to prokaryotic evolution, namely, that although LGTs are common, they occur with a frequency which is sufficiently low to render tree based representations of organismal evolution meaningful [3].

In the context of hosts and parasites, Huelsenbeck *et al.* developed a Bayesian framework for the detection of host switching using MCMC and taking advantage of sequence information from the host and the parasite species [20]. The model in [20] assumes a one-to-one correspondence between hosts and parasites and does not consider duplications. To our knowledge no probabilistic phylogenetic method has been proposed for simultaneous analysis of duplications and LGTs, although a probabilistic model based on the birth-death process [24] was used in [18] to generate synthetic data, and a similar model was used in [6] to estimate gene family sizes. However, the former never analyzed data

with respect to the model; the latter was only concerned with gene family size, not trees, and LGTs were modeled by introduction events without any explicit points of origin.

We provide the first probabilistic model capturing gene duplication, LGT, gene loss, and point mutations with a relaxed molecular clock. We also provide all fundamental algorithms required to analyze a gene family relative to a given species tree under this model. In the next section, our probabilistic model is presented. In Section 3, we describe an MCMC approach for estimating the posterior distribution of our model. It turns out that computing the generation probability of a gene tree G with edge lengths l, $\Pr[G, l|\theta]$ (where θ consists of parameters of the model), is crucial. We carefully describe how this probability can be expressed, and also, how it can be approximated by introducing discretization points in the species tree. Section 4 contains derivations of differential equations for several important distributions, for instance the probability of extinction. The corresponding distributions can be evaluated at the discretization points using numerical techniques. Also in section 4, we describe how a dynamic programming (DP) algorithm can be constructed for the approximation of $\Pr[G, l|\theta]$ by taking advantage of differential equations, which are also formulated in the section. Differential equations and algorithms that enable approximation of the probability that G, l has been generated using k LGTs are presented in Section 5. Finally, preliminary results from experiments on synthetic data are presented in Section 6

2 A new model for duplication, LGT, loss, rate variation, and sequence evolution

The *duplication-transfer-loss gene sequence evolution model with iid rates across gene tree edges*, which we denote DTLSR, is a joint generalization of models used in [18] (which are here described for the first time) and the GSR model [37]. DTLSR integrates the following probabilistic sub-models, which will be described more fully below:

1. A probabilistic duplication-transfer-loss model (DTL-model) describing a gene evolving over a species tree through gene duplication, LGT, and gene loss, thereby generating a gene tree.

2. A substitution rate model describing rate variation over the gene tree.

3. A sequence evolution model describing how nucleotide substitutions occur.

Let the species tree S and the gene tree G generated by the duplication-transfer-loss process be planted trees, i.e., trees with a root of degree one. These trees also have divergence times associated with their vertices. Because S and its divergence times are considered given, they will be omitted from our notation for probabilities, i.e., $\Pr[\cdot|S]$ will be written $\Pr[\cdot]$.

A gene tree vertex represents either a speciation, a duplication, or an LGT event; the divergence time for a speciation vertex is given by the corresponding species tree vertex, while the divergence time for a duplication or an LGT vertex is given by the duplication-transfer-loss process. Divergence times associated with vertices of a tree induce edge times in the natural way.

We use the substitution rate model in order to obtain a relaxed molecular clock [35, 25, 1, 34, 9, 29, 33], which allows for more biological realism. The substitution rate model also turns out to facilitate a more efficient and more accurate MCMC implementation. In the next three subsections, we briefly describe each of the DTLSR sub-models.

2.1 Gene duplication, LGT, and gene loss

In the probabilistic DTL-model, a gene tree G evolves over a species tree S with given divergence times. Over any edge $\langle x, y\rangle$ in the species tree, each gene lineage is exposed to gene duplications, LGTs, and gene losses with rates δ, τ, and μ, respectively. That is, in an interval of length h on a species tree edge $\langle x, y\rangle$ the probabilities of a single gene lineage being exposed to a duplication, an LGT, and a loss are, respectively,

$$\delta h, \tau h, \text{ and } \mu h. \tag{1}$$

Moreover, the probability of two or more events happening in such an interval is $o(h)$. When a gene u is exposed to a duplication event, it is replaced by two children, which both continue evolving over the same species tree edge as did u. When a gene u is exposed to an LGT, it is replaced by two children: one continuing to evolve over the same species tree edge $\langle x, y\rangle$ as did u, and one evolving over another species tree edge chosen uniformly from those concurrent with $\langle x, y\rangle$ at the time of the LGT event. A loss of the gene u removes it from the process as well as from the generated tree, in which also its former parent is suppressed. Each gene lineage reaching a speciation vertex y in S splits into two independent processes, each evolving down distinct outgoing edges of y. The process continues recursively down to the leaves where it stops.

The process also generates a *realization* explaining how the gene tree has evolved by mapping each gene tree vertex to a pair with one component being the species tree edge or vertex where the event happened and the other component being the time when the event creating the vertex happened. We will later introduce several types of realizations and the type of realization generated by the process will be called c-realizations. Computing the probability of a given gene tree under the model is non-trivial and we will use a combination of dynamic programming and techniques from numerical analysis to accomplish this task.

2.2 Substitution rates

The purpose of the substitution rate model is to transform dated trees with leaves representing extant entities, such trees being necessarily ultra-metric (i.e., all root-to-leaf paths have the same length), into trees consistent with a relaxed molecular clock. This provides a biologically realistic prior distribution for *edge lengths*—the convolution of edge times and substitution rates conventionally used in substitution models. We achieve a relaxed molecular clock by assuming that edge substitution rates are *independently and identically* Γ*-distributed* variables with mean m and variance ν [29, 38]. We denote this gamma distribution ρ.

Let l, r, and t denote functions associating an edge length, an edge specific rate, and an edge time, respectively, to each edge of G so that, e.g., $l(u, v)$ is the edge length of the

edge $\langle u, v\rangle$. The relation between lengths, rates, and times over all edges will be denoted by $l = rt$, or conversely $r = l/t$.

2.3 Sequence evolution

Each edge in the gene tree has, as explained above, been assigned an *edge length* by the duplication-transfer-loss process and the substitution rate process. Sequence evolution over the gene tree with these edge lengths can be modeled using any of the standard substitution models used in phylogenetics [11].

3 MCMC and discretizing the gene tree probability

MCMC is commonly used to estimate the posterior of phylogenetic trees for given gene sequences [21]. In this application of the MCMC methodology, it is natural to let the states consist of trees with edge lengths and additional parameters. When considering to use MCMC to estimate posterior probabilities under the GSR model [37], the most immediate idea is to also include a reconciliation of the gene and species tree (which explains how the gene tree evolved by mapping it into the species tree, the explanation may contain duplications and losses but not LGTs) as a component of the state; this approach would, however, lead to several technical complications. Fortunately, it is possible to evade these problems by estimating an integral over all reconciliations [37]. Here we will use a similar approach, although in our case, estimating the integral is significantly harder due to the inclusion of LGTs.

To simplify notation, let $\theta = (\delta, \tau, \mu, m, \nu)$ denote the parameters of the DTLSR model (there may also be additional parameters associated with the sequence evolution model, but we omit these from the present notation). Our Markov chain will have states of the form (G, l, θ) where G is a gene tree, l denotes edge lengths, and θ denotes parameters of the DTLSR model. Ratios between posterior probabilities of the form $p[G, l, \theta|D]$ need to be computed in order to determine acceptance probabilities of proposed states in our Markov chain. This posterior probability can be rewritten as follows:

$$p[G, l, \theta|D] = \frac{\Pr[D|G, l] \Pr[G, l|\theta]\, p[\theta]}{\Pr[D]}, \tag{2}$$

where the parameters θ are assigned independent priors (which will be uniform or some other distribution that we can compute). As usual in MCMC estimation of posterior probabilities, the denominators will cancel in any ratio between two such probabilities. Moreover, the factor $\Pr[D|G, l]$ can be computed using the standard DP algorithm introduced by Felsenstein [11]. The last component of our MCMC algorithm for estimating the posterior of the DTLSR model is a procedure to estimate $\Pr[G, l|\theta]$.

In the next subsections, we will show how to estimate $\Pr[G, l|\theta]$ by summing over realizations that only associate gene tree vertices to points from a set of discretization points on the species tree. We will clearly describe the two approximations we make. The expression below is formally incorrect (since our density function is discontinuous at the vertices of S) but fits intuitively with the formal description we will give, and it also leads

to a functional MCMC algorithm for estimating posteriors of gene trees. The integration is over realizations t and the summation over discretized realizations:

$$\begin{aligned}\Pr[G, l|\theta] &\approx \int_t p[G, l, t|\theta]dt \\ &= \int_t p[(r = l/t)|m, \nu]\ p[G, t|\delta, \tau, \mu]dt \\ &\approx \sum_t p[(r = l/t)|m, \nu]\ p[G, t|\delta, \tau, \mu].\end{aligned}$$

3.1 Definitions

We will now introduce several concepts that will be useful in the rest of the article. When the notation introduced in this subsection is used, the tree will be clear from the context. For each species tree T and each vertex $x \in V(T)$, there will be an associated divergence time $t(x)$. Associated with each edge $\langle x, y\rangle$ of a species tree is the interval $I(x, y) = [t(y), t(x)]$. The leaves of a species tree have divergence time 0 and each internal vertex has divergence time > 0. We will assume that all speciations have taken place at distinct times although all our results can easily be modified to allow concurrent speciations.

The following definitions are standard. An edge $\langle x, y\rangle$ has *tail* x and *head* y and it is an *outgoing* edge of x. For a pair of edges e and f of the same tree, if the head of e is the tail of f, then e is the *parent* of f and f a *child* of e. If there is an edge $\langle x, y\rangle$ in the tree T, then x is the parent of y and denoted $p_T(y)$. The *proper ancestor* relation is the transitive closure of the parent relation. That a is a proper ancestor of b in the tree T is denoted $a >_T b$, and b is also said to be a proper descendant of a. If a equals b or is a proper ancestor of b in T, then a is an *ancestor* of b in T, which is denoted $a \geq_T b$. Two vertices are said to be comparable if one is a descendant of the other, and incomparable otherwise. Finally, the planted subtree of T containing u, its parent $p_T(u)$, and all descendants of u is denoted T^u.

3.2 A discrete approximation of the probability of a gene tree

In this subsection, we show how to properly express $\Pr[G, l|\theta]$. A key step is to discretize the species tree, which will also give us subintervals of the edges of S in which the discretization points can be considered to be midpoints. We will use two approximation steps in order to compute $\Pr[G, l|\theta]$. The first approximation is an assumption that only one of any two comparable vertices in G can be created by the events occurring in a particular subinterval. The second approximation is obtained by approximating the density function in any point of the subinterval by the density function's value in the subinterval's midpoint.

We will now in two steps introduce discretization points in S to obtain a second species tree S' and then also a third species tree S''. Let S' be the tree obtained from S by recursively for each $t \in \{t(x) : x \in V(S)\}$ subdividing each edge $\langle x, z\rangle$ of S such that $t(z) < t < t(x)$ by introducing a new vertex y and letting the divergence time of y be defined by $t(y) = t$. See Figure 1a and 1b for an example.

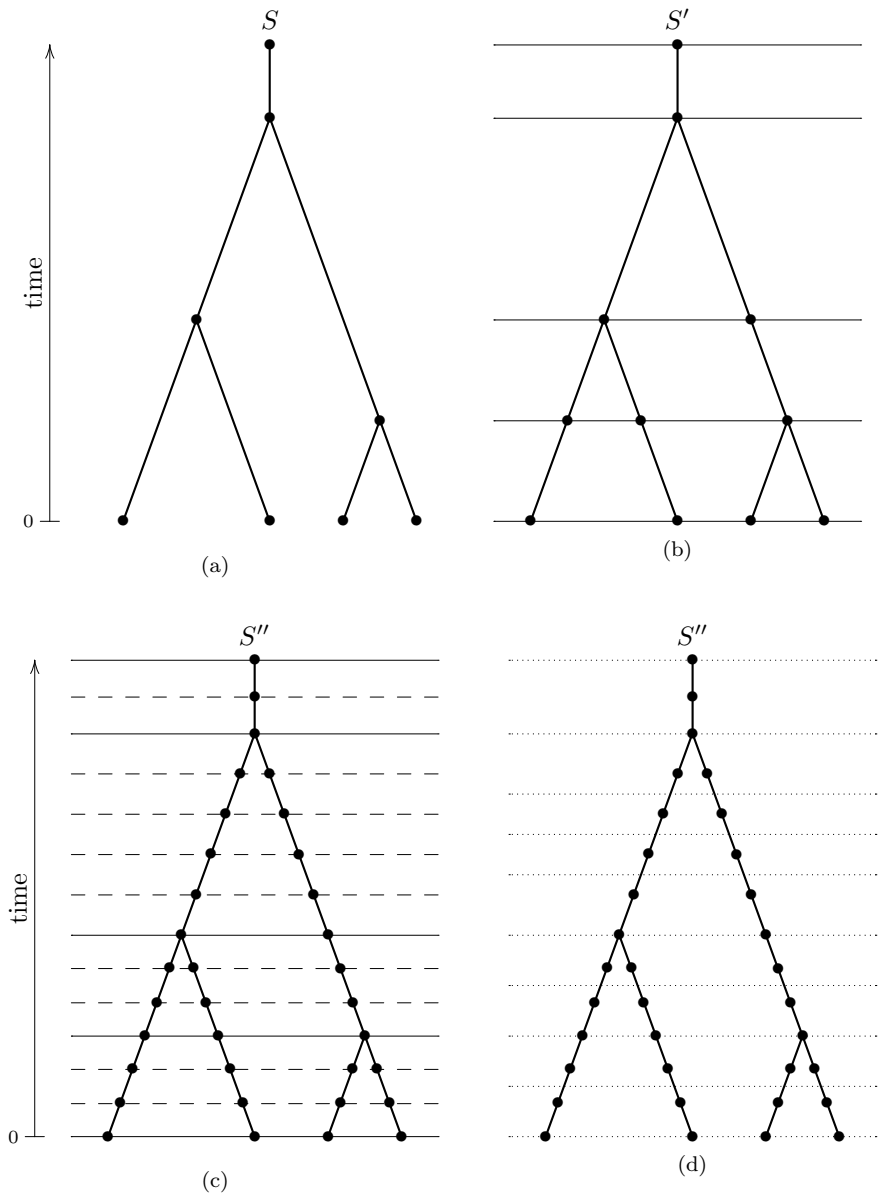


Figure 1: **The subdivisions and subintervals of the species tree.** (a) A species tree S. (b) The tree S'. (c) The tree S''. (d) The subintervals Δ associated with the vertices of S''. Note how each discretization point in $\mathcal{D}$ is a "midpoint" of a subinterval.

We subdivide S' by introducing new vertices at a number of discretization points $\mathcal{D}$. The set of discretization points can in principle be arbitrary, although the accuracy of the algorithm will depend on it. It is, for instance, natural to use all multiples of an interval length d as the set of discretization points, i.e.,

$$\mathcal{D} = \{dk : k \in N^+ \text{ and } dk \leq t(\text{root}(S'))\}.$$

We will for convenience assume that $\mathcal{D}$ and $\{t(x) : x \in V(S')\}$ are disjoint. Let S'' be the tree obtained from S' by recursively for each $t \in \mathcal{D}$ subdividing each edge $\langle x, z\rangle$ such that $t(z) < t < t(x)$ by introducing a new vertex y and letting the divergence time of y be defined by $t(y) = t$. See Figure 1c for an example.

A *continuous realization* (c-realization) of G is a function $c : V(G) \to \{\langle e, t\rangle : e \in E(S') \text{ and } t \in I(e)\} \cup \{\langle x, t(x)\rangle : x \in V(S)\}$ such that for each $u >_G v$,

$$c_t(u) > c_t(v),$$

where $c_t(u)$ denotes the projection of the second component of $c(u)$. The projection of the first component of $c(u)$ is denoted $c_V(u)$. A *speciation realization* (s-realization) of G is a function $s : U \to V(S)$ where $U \subseteq V(G)$ such that for each $u, v \in U$, $u >_G v$ implies

$$t(s(u)) > t(s(v)).$$

A *discrete realization* (d-realization) of G is a function $d : V(G) \to (V(S'') \setminus V(S')) \cup V(S)$ such that for each $u >_G v$,

$$t(d(u)) > t(d(v)).$$

Notice that for each edge e of S'', there is a unique edge $\langle x, y\rangle$ of S' such that the path in S'' between the vertices x and y contains e; we say that the edge $\langle x, y\rangle$ *captures* the edge e. Analogously for each vertex $z \in V(S'') \setminus V(S')$, there is a unique edge $\langle x, y\rangle$ of S' such that the path in S'' between the vertices x and y contains z; we say that the edge $\langle x, y\rangle$ *captures* the vertex z.

For each vertex $x \in V(S'') \setminus V(S')$ that is captured by the edge $e \in E(S')$, we associate what can be called a subinterval of e as follows. Assume that y is the single child of x. First, if $p_{S''}(x) \in V(S')$, define $t_p(x)$ to be $t(p_{S''}(x))$, and otherwise define $t_p(x)$ to be $(t(p_{S''}(x)) + t(x))/2$. Second, if $y \in V(S')$, define $t_c(x)$ to be $t(y)$, and otherwise define $t_c(x)$ to be $(t(x) + t(y))/2$. Finally, let $\Delta(x) = [t_c(x), t_p(x))$ and let $|\Delta(x)|$ denote the length of the the interval $\Delta(x)$, i.e., $|\Delta(x)| = t_p(x) - t_c(x)$. See Figure 1d for an example.

Let s be an s-realization of G. A c-realization c of G is a c-extension of s if $c_V|_{c_V^{-1}(V(S))} = s$. Similarly, a d-realization d of s is a d-extension of s if $d|_{d^{-1}(V(S))} = s$. A c-realization is *sparse* if for each $x \in V(S'')$ and each pair of vertices $u, v \in V(G)$,

$$u >_G v \text{ and } c_t(u) \in \Delta(x) \text{ implies } c_t(v) \notin \Delta(x).$$

Let $\mathbf{s_G}$ be the set of s-realizations of G. Let $\mathbf{d_G}$ be the set of d-realizations of G. For any s-realization of G, let $\chi_{\mathbf{c}}(s)$, $\chi_{\mathbf{s}}(s)$, and $\chi_{\mathbf{d}}(s)$ be the sets of c-extensions, sparse c-extensions, and d-extensions of s, respectively. Since the vertices of S create discontinuities in the density $p[G, l, c|\theta]$, we express the probability $\Pr[G, l|\theta]$ as the following sum:

$$\sum_{s \in \mathbf{s_G}} \Pr[G, l, s|\theta] = \sum_{s \in \mathbf{s_G}} \int_{c \in \chi_{\mathbf{c}}(s)} p[G, l, c|\theta] dc.$$

We approximate the RHS by

$$\sum_{s \in \mathbf{s_G}} \int_{c \in \chi_{\mathbf{s}}(s)} p[G, l, c|\theta] dc.$$

Notice that, for any c-realization or sparse c-realization c,

$$p[G, l, c|\theta] = \prod_{\langle u,v \rangle \in E(G)} p[l(u, v), c(v)|c(u), \theta].$$

As a notational convenience, we define $|\Delta(x)| = 1$ for $x \in V(S')$. Our approximation of the probability $\Pr[G, l|\theta]$ can now be summarized as

$$\begin{aligned}
\Pr[G, l|\theta] &= \sum_{s \in \mathbf{s_G}} \int_{c \in \chi_{\mathbf{c}}(s)} p[G, l, c|\theta] dc \\
&\approx \sum_{s \in \mathbf{s_G}} \int_{c \in \chi_{\mathbf{s}}(s)} p[G, l, c|\theta] dc \\
&= \sum_{s \in \mathbf{s_G}} \int_{c \in \chi_{\mathbf{s}}(\mathbf{s})} \prod_{\langle u,v \rangle \in E(G)} p[l(u, v), c(v)|c(u), \theta] dc \\
&\approx \sum_{s \in \mathbf{s_G}} \sum_{d \in \chi_{\mathbf{d}}(\mathbf{s})} \prod_{\langle u,v \rangle \in E(G)} p[l(u, v), d(v)|d(u), \theta] \cdot |\Delta(d(v))|. \qquad (3)
\end{aligned}$$

In the next section, we will show how to compute the right hand side of the above equation.

4 Computing the probability of a gene tree using DP

In this section, we will show how to compute the RHS of (3) using DP. In the DP algorithm, two distributions will turn out to be useful. First, the probability of extinction, for which we will derive differential equations in Subsection 4.1. Second, in Subsection 4.2, we will derive differential equations for the probability of a single gene u evolving, between two points in the species tree, to a single descendant v that may give rise to descendants in the extant species (u may also have other descendants contemporary to v but these will go extinct before reaching the leaves of the species tree). In the last subsection, we derive a DP algorithm for computing (3) from differential equations.

Let $\phi = \delta + \tau + \mu$. When the notation introduced in this paragraph is used it will be clear from the context whether the species tree concerned is S' or S''. Two vertices x and y are said to be *contemporary* if $t(x) = t(y)$. Two edges $\langle x, y \rangle$ and $\langle x', y' \rangle$ are said to be *contemporary* if $t(x) = t(x')$ and $t(y) = t(y')$ (in fact, in both S' and S'', $t(x) = t(x')$ implies $t(y) = t(y')$ and the other way around). An *edge generation* is a maximal set of pairwise contemporary edges. The edge generation containing e is denoted $\mathcal{G}_E(e)$ and the edges contemporary to e, i.e., $\mathcal{G}_E(e) \setminus \{e\}$, is denoted $\mathcal{C}_E(e)$. For two edges e, f, if e is the parent of f, then $\mathcal{G}_E(e)$ is the *parental generation* of $\mathcal{G}_E(f)$.

4.1 The probability of extinction

In this subsection, we derive differential equations for the probability of extinction, which can be solved numerically. For $e \in E(S')$ and $t \in I(e)$, let $Q_e(t)$ be the probability of extinction when starting with a single gene at time t on edge e. The following system of differential equations follow from standard techniques for Poisson processes [10, 2] and the fact that when an LGT occurs, the edge to which the transfer is made is chosen uniformly:

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_e(t) = \delta(Q_e(t))^2 + \tau\left(\sum_{f\in\mathcal{C}_E(e)}\frac{1}{|\mathcal{C}_E(e)|}Q_e(t)Q_f(t)\right) + \mu - \phi Q_e(t). \tag{4}$$

For $e = \langle x, y\rangle \in E(S')$, the initial values for the system of equations above are given by

$$Q_e(t(y)) = \begin{cases} 0 & \text{if } y \text{ is a leaf,} \\ Q_f(t(y)) & \text{if } f \text{ is the single child of } e, \\ Q_f(t(y))Q_g(t(y)) & \text{if } f \text{ and } g \text{ are the two children of } e. \end{cases}$$

For one generation of edges of S', the systems of equations for Q_e can be solved using standard Runge-Kutta numerical solvers [17] once the systems for proper descendant generations have been solved. That is, we can solve these equations first for the generation of edges incident to the leaves and then continue upwards to the root of the species tree. For the edge generation $\mathcal{G}_E(e)$, we solve $Q_e(t)$ for all $t \in \{t(x) : x \in V(S'')\} \cap I(e)$.

4.2 The probability of exactly one mortal descendant

In this subsection, we apply the same approach as in the previous subsection. In this case, we are interested in the probability of a single gene u evolving, between to points in the species tree, to a single descendant v that may give rise to descendants in the extant species (u may also have other descendants contemporary to v but none of these should give rise to extant descendants).

By a *ghost* we mean a gene in the probabilistic DTL-model that will not have any descendants among the leaves of the species tree. In contrast, a *mortal* is a gene that may or may not yield descendants among the leaves of the species tree. For a pair of edges e, f of S' such that $s \in I(e)$, $t \in I(f)$, and $t < s$, define $Q_{ef}(s,t)$ as the probability of starting on e at time s and having one mortal in f at time t and all other descendants at time t being ghosts.

Let $e, f \in E(S')$ be two contemporary edges. As before, the following system of differential equations can be obtained using standard techniques:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s}Q_{ef}(s,t) &= 2\delta Q_e(s)Q_{ef}(s,t) - \phi Q_{ef}(s,t) \\ &\quad + \tau\sum_{g\in\mathcal{C}_E(e)}\frac{1}{|\mathcal{C}_E(e)|}\Big(Q_{gf}(s,t)Q_e(s) + Q_{ef}(s,t)Q_g(s)\Big).\end{aligned}$$

The initial values for the above equations are given by

$$Q_{ef}(t,t) = \begin{cases} 1 & \text{if } e = f, \\ 0 & \text{otherwise.} \end{cases}$$

For one generation of edges of S', the systems of equations for Q_{ef} and Q_e can be solved together using standard Runge-Kutta numerical solvers. For the edge generation $\mathcal{G}_E(e)$, we solve $Q_{fg}(s,t)$ for all $f,g \in \mathcal{G}_E(e)$ and $s,t \in \{t(x) : x \in V(S'')\} \cap I(e)$.

We will now show how to compute Q_{ef} when e is a proper ancestor of f, i.e., when e and f belong to different edge generations of S'. Let $e = \langle x, y\rangle$ and assume that g is the unique edge that is contemporary with e and has two children g' and g''. For any $s \in I(e)$ and $t \in I(f)$, $Q_{ef}(s,t)$ can be written

$$Q_{ef}(s,t) = Q_{eg}(s,t(y))\Big(Q_{g'f}(t(y),t)Q_{g''}(t(y)) + Q_{g''f}(t(y),t)Q_{g'}(t(y))\Big) + \sum_{h \in \mathcal{C}_E(g)} Q_{eh}(s,t(y))Q_{hf}(t(y),t).$$

The above equations are solved for all $s \in \{t(x) : x \in V(S'')\} \cap I(e)$ and all $t \in \{t(x) : x \in V(S'')\} \cap I(f)$. These equations can be solved for all pairs of edge generations and discretization points recursively from the leaves of the species tree towards the root.

4.3 The final recursion

In this subsection, we derive a DP algorithm for computing (3) from the differential equations in the previous subsections.

We will need to compute the probability of extinction at the vertices of S'' and the probability of evolving to exactly one mortal between any pair of vertices in S''. For $e = \langle y, z\rangle \in E(S'')$ and $x \in V(S'')$ which is the head of the edge f in S'' and satisfies $t(x) \leq t(z)$, we define $p_{11}(e,x)$ as follows

$$p_{11}(e,x) = Q_{e'f'}(t(y),t(x)),$$

where e' and f' are the edges of S' that capture e and f, respectively.

Now, to compute the probability of a gene tree G, we sum the probabilities of every possible mapping of the gene tree vertices on the vertices of S''. For $x \in V(S'') \setminus L(S'')$ and $u \in V(G) \setminus L(G)$, define $a(x,u)$ as the probability of G_u given that the event creating u occurred at x. For $e \in E(S'')$ and $u \in V(G)$, define $s(e,u)$ as the probability of the planted tree G^u when starting at the tail of e. These two probabilities can be computed as follows. If x is a speciation, then

$$a(x,u) = s(e,v)s(f,w) + s(e,w)s(f,v),$$

where e, f are the outgoing edges of x and v, w are the children of u. If x is not contemporary to any speciations, then

$$a(x,u) = 2\delta s(e,v)s(e,w) + \tau \sum_{f \in \mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|}\Big(s(e,v)s(f,w) + s(e,w)s(f,v)\Big),$$

where e is the outgoing edge of x. We define $a(x,u)$ to be zero in the two other possible cases, i.e., when x is a leaf or has out-degree one and is also contemporary to a speciation.

For $e = \langle x, y\rangle \in E(S'')$ and $u \in V(G)$, we can compute $s(e,u)$ as follows:

$$s(e,u) = \begin{cases} p_{11}(e,\sigma(u))\rho\left(\frac{l(p(u),u)}{t(x)}\right) & \text{if } u \in L(G), \\ \sum_{z\in\mathcal{Q}(x)} p_{11}(e,z)\rho\left(\frac{l(p(u),u)}{t(x)-t(z)}\right) a(z,u) & \text{otherwise,} \end{cases}$$

where $\mathcal{Q}(x)$ is the set of all vertices z of S'' such that $t(z) < t(x)$.

5 Counting Transfers

In this section we present differential equations and algorithms that enable approximation of the probability that G, l has been generated using k LGTs. We wish to compute the probability that G is generated and that exactly k LGTs has occurred on the paths to the leaves of G during the generation of G. In order to accomplish this, we will compute almost the same probabilities as in the previous section, but with the addition of an index k to keep track of the number of LGTs used. In this sense, this section is an extension of the previous section. The probability of extinction is the same as before, since we are not counting the number of LGTs only creating ghosts.

Let e and f be two contemporary edges of S''. For $t \le s \in I(e)$, define $Q_{efk}(s,t)$ to be the probability of starting on e at time s having some number of ghosts at time t and, except for these ghosts, having only produced one mortal on f at time t using exactly k LGTs (so there is a path containing k LGTs that end on f at time t). For $k > 0$, the following holds

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}Q_{efk}(t) = {} & 2\delta Q_e(s)Q_{efk}(s,t) - \phi Q_{efk}(s,t) \\ & + \tau \sum_{g\in\mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|}\Big(Q_{gf(k-1)}(s,t)Q_e(s) + Q_{efk}(s,t)Q_g(s)\Big).\end{aligned}$$

For $k = 0$ a similar expression can be obtained

$$\frac{\mathrm{d}}{\mathrm{d}t}Q_{ef0}(t) \quad = \quad 2\delta Q_e(s)Q_{ef0}(s,t) - \phi Q_{ef0}(s,t) + \tau \sum_{g\in\mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|}Q_{ef0}(s,t)Q_g(s).$$

As before, the initial values for the above equations are given by

$$Q_{efk}(t,t) = \begin{cases} 1 & \text{if } e = f \text{ and } k = 0, \\ 0 & \text{otherwise.} \end{cases}$$

We will now show how to compute Q_{efk} when $e = \langle x, y\rangle$ and f belong to different edge generations. Assume that g is the unique edge in $\mathcal{G}_E(e)$ that has two children g' and g''. For any edge $h \in \mathcal{C}_E(g)$, let h' denote the unique child of h. For any $s \in I(e)$ and

$t \in I(f)$, $Q_{efk}(s,t)$ can be written

$$Q_{efk}(s,t) = \sum_{k'+k''=k} \Bigg(Q_{egk'}(s,t(y)) \Big(Q_{g'fk''}(t(y),t) Q_{g''}(t(y)) + Q_{g''fk''}(t(y),t) Q_{g'}(t(y)) \Big) + \sum_{h \in \mathcal{C}_E(g)} Q_{ehk'}(s,t(y)) Q_{h'fk''}(t(y),t) \Bigg).$$

For $e = \langle y, z \rangle \in E(S'')$ and $x \in V(S'')$ such that $t(x) \leq t(z)$, we define $p_{11k}(e,x)$ as follows

$$p_{11k}(e,x) = Q_{e'f'k}(t(y),t(x)),$$

where e' and f' are the edges of S' that captures e and f, respectively.

To compute the probability of a gene tree G with k LGTs, we sum the probabilities of every possible mapping of the gene tree vertices on the vertices of the subdivision S''. For $x \in V(S) \setminus L(S)$ and $u \in V(G) \setminus L(G)$, define $a_k(x,u)$ as the probability of generating G_u using exactly k LGTs given that the event creating u occurred at x. For $e \in E(S')$ and $u \in V(G)$, define $s_k(e,u)$ as the probability of generating the planted tree G^u using k LGTs when starting at the tail of e. These two probabilities can be computed as follows.

If x is a speciation, then

$$a_k(x,u) = \sum_{k'+k''=k} s_{k'}(e,v) s_{k''}(f,w) + s_{k'}(e,w) s_{k''}(f,v),$$

where e, f are the outgoing edges of x and v, w are the children of u. If x is not contemporary to any speciations, i.e., contemporary to any vertices of S, then

$$a_k(x,u) = \sum_{k'+k''=k} 2\delta s_{k'}(e,v) s_{k''}(e,w) + \tau \sum_{k'+k''=k-1} \sum_{f \in \mathcal{C}_E(e)} \frac{1}{|\mathcal{C}_E(e)|} \Big(s_{k'}(e,v) s_{k''}(f,w) + s_{k''}(e,w) s_{k'}(f,v) \Big),$$

where e is the outgoing edge of x. We define $a_k(x,u)$ to be zero in the two other cases, i.e., when x is a leaf or has out-degree one and is also contemporary to a speciation.

For $e = \langle x, y \rangle \in E(S')$ and $u \in V(G)$, we can compute $s_k(e,u)$ as follows:

$$s_k(e,u) = \begin{cases} p_{11k}(e,\sigma(u)) \rho \left(\frac{l(p(u),u)}{t(x)} \right) & \text{if } u \in L(G), \\ \sum_{k'+k''=k} \sum_{z \in \mathcal{Q}(x)} p_{11k'}(e,z) \rho \left(\frac{l(p(u),u)}{t(x)-t(z)} \right) a_{k''}(z,u) & \text{otherwise,} \end{cases}$$

where $\mathcal{Q}(x)$ is the set of all vertices z of S' such that $t(z) < t(x)$.

6 Experimental results

In this section, we present results of preliminary experiments performed on synthetic data. For our species tree, we selected a subset of the taxa in the yeast tree which together with

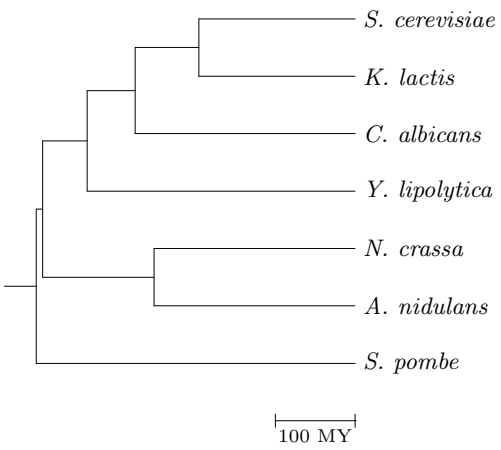


Figure 2: **The yeast tree used in the experiments.** The root-to-leaf time of the tree is approximately 400 million years. In the experiments, the time was rescaled so that the root-to-leaf time became 1.0 and the total birth-rate became 0.17 (relative to the new time scale).

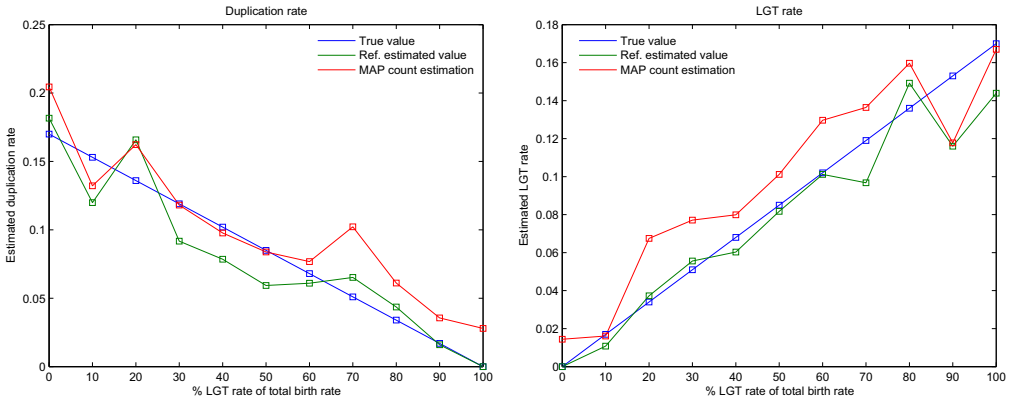
divergence times was presented in [37]. The species tree is shown in Figure 2. The species tree was rescaled so that the root-to-leaf time equaled 1.0. In addition, an edge of length 0.1 preceding the root was introduced in order to allow duplication events to occur prior to the first speciation. Naturally, LGT events may not take place along this edge.

We generated 11 distinct sets of gene trees according to the probabilistic DTL-model, each comprising 100 trees. Analysis of the data from [37] yields an estimated death rate of approximately 0.17. When generating the trees, both the death rate μ and the total birth rate, i.e., $\delta + \tau$, were kept fixed at 0.17, while the LGT rate τ varied between 0% and 100% of the total birth rate in steps of 10% increments. All gene trees were produced starting with a single lineage at the earliest point of the species tree.

We used the resulting gene tree topologies and divergence times to generate sequences using the JTT amino acid substitution model [22]. Edge rates were drawn *iid* from a gamma distribution with mean 0.5 and variance 0.1, with no rate variation among sites. The output of this procedure was aligned amino acid sequences of length 1,000.

In order to verify the soundness of our probabilistic model, we analyzed the posterior distributions of the duplication and LGT rates. The gene tree topologies were kept fixed to the true topologies, while the remaining parameters were inferred by the MCMC process.

As we are not yet able to perform automatized convergence testing, we conducted a series of pilot tests to analyze mixing and simulation length requirements. We selected several of the smallest and largest trees from each set and ran three seperate chains with 1,000,000 iterations per tree, sampling every 100th iteration. A small number of trees proved too small for stable performance, each such tree had only two leaves. These were consequently removed from further consideration. Parameter trace plots on remaining trees indicated good mixing. Convergence was evaluated using the Gelman-Rubin test [13],



(a) (b)

Figure 3: **Results of tests performed on synthetic data.** The estimated duplication and LGT rates are plotted against the true rates. The blue curve plots the true rates showing the ideal estimation curves. The green curve shows our reference estimate of the rates using previous knowledge about the true duplication vertices and transfer edges. The red curve is obtained by first estimating the number of events using the posterior distribution of the rates, and analogous to the green curve, using these values to obtain estimates of the rates.

where all parameters had a joint test statistic ≤ 1.02, with the exception of the LGT rate in one instance reaching 1.08. All tests were conducted with the first 10% of the samples removed as burn-in. We concluded that these settings are sufficient to provide convergence in most cases, and used them in subsequent analyses.

The output of each tree, i.e., the merged chain triplet, was used to estimate the posterior distribution. We then analyzed the marginal distributions of the duplication and LGT rates, and after applying a moderate smoothing kernel, the MAP rate for each gene tree was used to estimate the number of duplication and LGT events.

The true number of duplication vertices and transfer edges divided by the total length of the species tree was used to estimate the birth rates of each gene tree. The average value for each set was used as a reference estimate of the birth rates. Similarly, the estimated number of duplications and LGTs derived from the posterior distributions were divided by the total length of the species tree to obtain estimates of the birth rates from our MCMC procedure. Figure 3 shows the results.

References

[1] S. Aris-Brosou and Z. Yang. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal RNA phylogeny. *Syst Biol*, 51(5):703–714, Oct 2002.

[2] N.T.J. Bailey. *The elements of stochastic processes with applications to the natural sciences*. Wiley-Interscience, 1990.

[3] E. Bapteste, E. Susko, J. Leigh, D. MacLeod, R.L. Charlebois, and W.F. Doolittle. Do orthologous gene phylogenies really support tree-thinking? *BMC Evol Biol*, 5(1):33, 2005.

[4] J.R. Brown. Ancient horizontal gene transfer. *Nat Rev Genet*, 4(2):121–132, Feb 2003.

[5] J.H. Camin and R.R. Sokal. A method for reducing branching sequences in phylogeny. *Evolution*, 19:311–326, 1965.

[6] M. Csűrös and I. Miklós. A probabilistic model for gene content evolution with duplication, loss and horizontal transfer. In *In the tenth annual international conference on Research in Computational Molecular Biology (RECOMB)*, pages 206–220. Springer, 2006.

[7] W.F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2129, Jun 1999.

[8] W.F. Doolittle and E. Bapteste. Pattern pluralism and the tree of life hypothesis. *Proc Natl Acad Sci U S A*, 104(7):2043–2049, Feb 2007.

[9] A.J. Drummond, S.Y. Ho, M.J. Phillips, and A. Rambaut. Relaxed phylogenetics and dating with confidence. *PLoS Biol*, 4(5):e88, May 2006.

[10] W. Feller. *An introduction to probability theory and its applications*, volume 1. Wiley, 1968.

[11] J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–376, 1981.

[12] J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2003.

[13] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Stat Sci*, 7(4):457–472, 1992.

[14] J.P. Gogarten, W.F. Doolittle, and J.G. Lawrence. Prokaryotic evolution in light of gene transfer. *Mol Biol Evol*, 19(12):2226–2238, Dec 2002.

[15] M. Goodman, J. Cselusniak, G.W. Moore, A.E. Romero-Herrera, and G. Matsuda. Fitting the gene lineage into its species lineage: A parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool*, 28:132–168, 1979.

[16] F. Griffith. The significance of pneumococcal types. *J Hyg*, 27:113–159, 1928.

[17] E. Hairer, S.P. Nørsett, and G. Wanner. *Solving ordinary differential equations I: nonstiff problems*. Springer-Verlag, 1993.

[18] M. Hallett, J. Lagergren, and A. Tofigh. Simultaneous identification of duplications and lateral transfers. *Proceedings of the eighth annual international conference on Research in Computational Molecular Biology (RECOMB)*, pages 347–356, 2004.

[19] M.T. Hallett and J. Lagergren. Efficient algorithms for lateral gene transfer problems. *Proceedings of the fifth annual international conference on Research in Computational Biology (RECOMB)*, 2001.

[20] J.P. Huelsenbeck, B. Rannala, and B. Larget. A Bayesian framework for the analysis of cospeciation. *Evolution*, 54(2):352–364, Apr 2000.

[21] J.P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, Aug 2001.

[22] D.T. Jones, W.R. Taylor, and J.M. Thornton. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, 8(3):275–282, Jun 1992.

[23] P.J. Keeling and J.D. Palmer. Horizontal gene transfer in eukaryotic evolution. *Nat Rev Genet*, 9(8):605–618, Aug 2008.

[24] David G. Kendall. On the generalized "birth-and-death" process. *Ann Math Stat*, 19:1–15, 1948.

[25] H. Kishino, J.L. Thorne, and W.J. Bruno. Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol Biol Evol*, 18(3):352–361, Mar 2001.

[26] A.G. Kluge and J.S. Farris. Quantitative phyletics and the evolution of anurans. *Syst Zool*, 18(1):1–32, 1969.

[27] J. Lederberg and E. Tatum. Gene recombination in Escherichia coli. *Nature*, 158:558, October 1946.

[28] J. Lederberg and E.L. Tatum. Novel genotypes in mixed cultures of biochemical mutants of bacteria. In *Cold Spring Harb Symp Quant Biol*, volume 11, pages 113–114, 1946.

[29] T. Lepage, D. Bryant, H. Philippe, and N. Lartillot. A general comparison of relaxed molecular clock models. *Mol Biol Evol*, 24(12):2669–2680, Dec 2007.

[30] E. Lerat, V. Daubin, H. Ochman, and N.A. Moran. Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol*, 3(5):e130, May 2005.

[31] K. Ochiai, T. Yamanaka, K. Kimura, and O. Sawada. Inheritance of drug resistance (and its transfer) between Shigella strains and between Shigella and E. coli strains. *Nihon Iji Shimpo*, 1861:34, 1959.

[32] H. Ochman, J.G. Lawrence, and E.A. Groisman. Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405(6784):299–304, May 2000.

[33] B. Rannala and Z. Yang. Inferring speciation times under an episodic molecular clock. *Syst Biol*, 56(3):453–466, Jun 2007.

[34] J.L. Thorne and H. Kishino. Divergence time and evolutionary rate estimation with multilocus data. *Syst Biol*, 51(5):689–702, Oct 2002.

[35] J.L. Thorne, H. Kishino, and I.S. Painter. Estimating the rate of evolution of the rate of molecular evolution. *Mol Biol Evol*, 15(12):1647–1657, Dec 1998.

[36] C.R. Woese. On the evolution of cells. *Proc Natl Acad Sci U S A*, 99(13):8742–8747, Jun 2002.

[37] Ö. Åkerborg, B. Sennblad, L. Arvestad, and J. Lagergren. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci U S A*, 106(14):5714–5719, Apr 2009.

[38] Ö. Åkerborg, B. Sennblad, and J. Lagergren. Birth-death prior on phylogeny and speed dating. *BMC Evol Biol*, 8:77, 2008.