Models, algorithms and programs for phylogeny reconciliation

Jean-Philippe Doyon, Vincent Ranwez, Vincent Daubin and Vincent Berry

Submitted: 3rd March 2011; Received (in revised form): 24th June 2011

Abstract
Gene sequences contain a gold mine of phylogenetic information. But unfortunately for taxonomists this information does not only tell the story of the species from which it was collected. Genes have their own complex histories which record speciation events, of course, but also many other events. Among them, gene duplications, transfers and losses are especially important to identify. These events are crucial to account for when reconstructing the history of species, and they play a fundamental role in the evolution of genomes, the diversification of organisms and the emergence of new cellular functions. We review reconciliations between gene and species trees, which are rigorous approaches for identifying duplications, transfers and losses that mark the evolution of a gene family. Existing reconciliation models and algorithms are reviewed and difficulties in modeling gene transfers are discussed. We also compare different reconciliation programs along with their advantages and disadvantages.

Keywords: phylogeny; gene duplication; loss; lateral gene transfer; parsimony; probability; reconciliation

INTRODUCTION
The systematic reconstruction of gene phylogenies from a wide variety of organisms reveals an unforeseen diversity of histories, which are hard to understand solely on the basis of simple species evolution patterns. Each gene history is a complex series of events including duplications, losses and lateral gene transfers (LGT). Observed differences among gene trees underline the importance of modeling factors that specifically affect gene evolution.

In this article, we discuss genome evolution models at the gene level so as to specifically account for gene duplication, gene loss and LGT mechanisms. We will consider that these events happen independently, e.g. an LGT adding an extra copy of a gene in a genome is not necessarily followed by a loss in this genome of one copy of this gene. For this reason, we do not review works on dependent events such as multiple gene duplication [1]. The development of such models is crucial to gain insight into the evolution of unicellular organisms where LGT has played a major role [2] and, more generally, to clarify homology relationships among genes. Indeed, the combination of duplication, transfer and loss over the history of life may have been such that no single phylogenetic marker can be considered reliable for inferring the history of species. The following studies arguably rely on the development of models of duplication, transfer and loss [3]: reconstructing the tree of life; understanding the principles of genome evolution, the role of transfer in species adaptation and the contribution of duplication and transfer to the evolution of new functions.

Reconciliation models consider a species tree within which a gene can evolve (Figure 1). Leaves
of the species tree and gene tree are associated and specific events are invoked to allow the gene to evolve within the species tree so as to explain its phylogeny. Partial models accounting for duplication and loss [4–6], or LGT and loss [7–9], have been described. These models are realistic in particular biological cases (resp. multicellular organisms where LGT is rare and gene families for which functional redundancy can be detrimental). Here we focus on models that account for Duplications and Losses (DL models) or on models that also consider Transfers (DTL models).

Reconciliation is a popular approach for inferring orthology relationships [10–14], even though its accuracy strongly depends on the phylogeny reliability [15, 16]. It has applications in other areas such as DTL rate estimation [17, 18], gene tree inference [19–22] and genome phylogeny reconstruction from discordant gene trees [5, 23]. Reconciliation can also be used to study co-evolution between parasites and their hosts (parasitology), and between organisms and their living areas (biogeography) [24–26].

**RECONCILIATION MODELS AND ALGORITHMS**

As for phylogenetic reconstruction, parsimony and probabilistic frameworks have been developed for reconciliation inference. Parsimony methods are based on explicit discrete models of gene evolution and search for an optimal reconciliation given the elementary costs of individual evolutionary events. Probabilistic methods rely on continuous models and seek a reconciliation with maximum likelihood or maximum posterior probability. Parsimony methods are faster but use less realistic models than probabilistic methods.

**Evolutionary scenarios with duplications and losses**

**Parsimony models**

There may be numerous reconciliations between a gene tree G and a species tree S. For instance, Figure 1 depicts two reconciliations R1 and R2 for the same trees—R2 differs from R1 by the location in S of the duplication u of an ancestral gene of G (and by the induced losses). Several ways to represent reconciliations have been proposed. One of the most widespread is through a so-called ‘reconciled tree’ [4, 6, 27], here denoted RT, and defined as follows: (i) the clade (considering a node of a phylogenetic tree, its clade represents the set of contemporary taxa present on the leaves of the corresponding subtree) of each node of RT has to be present in S; (ii) for each internal node of RT, the clades of its children are either equal (duplication) or disjoint (speciation); and (iii) G has to be obtained from RT by pruning some of its subtrees (Figure 2).

Based on the reconciled tree formalism, [28] introduced an architecture that allows to describe the whole set of reconciliations between G and S. They proved that a simple polygonal time algorithm, called LCA mapping, enables identification of one of the ‘most parsimonious reconciliations’ (MPR). [29] proved the same result, but for a now obsolete model [30]. This well-known LCA algorithm [4, 6, 31] can be implemented to run in linear time with respect to the number of nodes in G [29, 32]. The LCA mapping maps each gene u of G onto the most recent species x of S such that each contemporary gene that descends from u belongs to a contemporary species that descends from x (Figure 3). According to the definition, a node u of G is a duplication if and only if it is mapped to the same vertex of S as one of its children. Otherwise,
is a speciation. Based on this rule, the so-called LCA reconciliation is computed as follows. For each node $u$ of $G$ mapped on a vertex $x$ of $S$, if $u$ is a duplication (according to the LCA mapping), it is located on the branch immediately above $x$, otherwise $u$ is a speciation placed on $x$. For instance, $R_1$ of Figure 1 is the LCA reconciliation, while $R_2$ is not.

The reconciled tree model (Figure 2; RT) is sometimes less intuitive than a drawing of $G$ embedded within $S$ (Figure 2; R1). A model that leads to such a representation is defined in [33], but as each gene of $G$ is mapped on a set of vertices of $S$, it cannot be immediately interpreted. A more intuitive model is defined in [34], where each ancestral gene of $G$ is mapped either on a vertex (speciation) or a branch (duplication) of $S$. This paper also provides an algorithm to explore the reconciliation space by moving from one reconciliation to another via elementary transformations.

Given a cost for each event (i.e. duplication and loss), the parsimony score for reconciliations can be either the sum of costs for duplications or the sum of costs for duplications and losses. The LCA reconciliation provides an MPR for the two scores, and is even the only one in the latter case [28, 35]. Given a choice for event costs, numerous reconciliations may exist with near-optimal scores. Some could be optimal if slightly different event costs are used [36]. However, most reconciliation analyses focus on the LCA reconciliation and ignore such near-optimal reconciliations.

All branches of a phylogenetic tree are not equally reliable, bootstrap and posterior probabilities are usual support measures. Reconciliation methods are biased when the inferred gene tree $G$ is not correct [37]. This uncertainty in $G$ can be taken into account by collapsing weakly supporting branches, thus creating polytomous nodes. A heuristic [38] and two exact algorithms [20, 39] have been proposed to search for an MPR when $G$ is polytomous. An MPR can still be computed in polynomial time when $S$ contains polytomies [40].

It is also possible to account for duplication and losses when reconstructing gene trees from sequences. Given a set of genomes and a reference species tree $S$, ‘SYNERGY’ [41] simultaneously computes orthology/paralogy relationships and reconstructs a gene tree for each family. The evolutionary distance used to cluster genes into a family combines sequence similarity and syntenic block conservation. Hereafter, the reconciliation score is used to root each gene tree.

**Probabilistic models**

A probabilistic model of reconciliation has been developed [33, 42], where each branch of the species tree $S$ is associated with a pre-computed duration (branch length) and estimated duplication and loss rates. The model simulates the evolution of a gene $u$ along a subtree of $S$ rooted on $x$ as follows: (i) $u$ evolves toward the child $x_1$ of $x$ following a birth-and-death process (see [43] for a survey) along the branch $(x, x_1)$, and (ii) for each descendant of $u$ that survived until species $x_1$, step (i) is repeated.
departs from $x_1$ (the same is done for $x_2$). The advantage of such a model is that it offers the possibility of considering events that left no trace (so-called ‘ghost event’), e.g. gene duplication where one or both copies become extinct.

This model is used to compute the probability of an evolutionary scenario that gives rise to the gene tree $G$ and reconciliation $R$. This probability is the likelihood of $R$, denoted $P(G,R)$, and can be computed in time $O(n_G n_S)$, where $n_G$ and $n_S$ are the number of nodes in $G$ and $S$. The reconciliation of maximum likelihood can be computed in time $O(n_G n_S \log^3 n_G)$ [44].

An efficient algorithm to compute the probability $P(G)$, which is the sum of $P(G,R)$ over all reconciliations $R$, is developed in [44]. This is an important breakthrough as it allows to compute $P(G)$, which is the sum of $P(G,R)$ over all reconciliations $R$. A maximum likelihood reconciliation can be computed in time $O(n_G n_S)$ where the substitution rate of any gene is expressed as the product of gene-specific and species-specific rates. This model is used to compute the probability of such a small number of reconciliations $R$. An analysis of 1278 gene trees from 12 fungal genomes concluded that (i) a close neighborhood of the MPR (i.e. the LCA reconciliation) contains the most likely reconciliations; and (ii) the likelihood $P(G,R)$ of such a small number of reconciliations $R$ were sufficient in this case to compute very precise approximations of $P(G)$ and $P(R | G)$. Though these results rely on simplified assumptions on genomic evolution (such as constant duplication and loss rates along each branch of $S$), these results highlight the strong relationship that may exist between a probabilistic model [33, 42] and a parsimony model.

When considering a discrete distribution model of duplications parameterized by the branch lengths of $S$, a maximum likelihood reconciliation can be computed in time $O(n_G^2 n_S)$ [48, 49]. This approach does not consider ‘ghost events’ or losses, which can be problematic when losses are prevalent [50].

As in the parsimony context, several probabilistic approaches for gene tree reconstruction that consider DL events and sequence evolution have been proposed. Such mixed evolutionary models rely on a dated species tree and integrate: (i) gene duplication and loss [33, 42]; (ii) sequence substitution; and (iii) substitution rate variation over the gene tree (i.e. relaxed molecular clock [51]). By sampling DL rates and parameters of the substitution rate model, a MCMC approach estimates joint posterior probabilities of molecular sequences and gene trees [19]. A maximum likelihood method is developed in [52], where the substitution rate of any gene is expressed as the product of gene-specific and species-specific rates. This approach is extended into a Bayesian framework [21], where DL rates and the above-mentioned parameters are learned using two distinct EM algorithms [53]. Arguing on similar results as in [47], the topology of $G$ is approximated using MPR.

**Evolutionary scenarios with duplications, losses and transfers**

Computing an MPR is hard when transfers are considered [54, 55], although it can be solved in polynomial time with realistic constraints [56] (e.g. bounding the number of transfers, genes per species lineage, etc.). This high contrast in complexity is due to chronological constraints induced by transfers. A transfer has to be locally consistent, which means that it occurs between two coexisting species. Two (or more) consecutive transfers also have to be globally consistent (Figure 4). If these constraints are omitted, time inconsistent scenarios can ensue. In time $O(n_G^2 n_L)$, [55] solves such a variant of the MPR problem, where the number of duplications and transfers are optimized whereas losses are only used a posteriori to discriminate between MPRs.

A recent promising approach for handling time constraints is to accept a dated tree $S$ as input. Time consistency can then be ensured locally by checking that donor and receiver branches of a transfer have intersecting time intervals (Figure 4). This approach has been used in five reconciliation algorithms [18, 36, 57–59], which differ in their way to handle global consistency and in the degree of generality of their model.

Global time consistency can be ensured by: (i) altering the position of the proposed transfers a posteriori; (ii) checking that all branches involved in a succession of transfers share a sub-interval of time; or (iii) using a subdivision of the branches of
S within time slices and allowing for transfers only between branches of the same time slice. The two-step strategy of (i) does not guarantee that an optimal reconciliation will be found. Approach (ii) only considers a subset of scenarios due to over-restrictive rules. For instance, reversing the direction of transfers $T1$ and $T2$ in Figure 4 leads to a globally consistent scenario. However, this scenario is rejected by (ii) as the three involved branches do not share a common time subinterval. Approach (iii) ensures that an optimal reconciliation will be found [58, 59].

Models have to be general enough to encompass the variability of all possible scenarios involving transfers. In particular, they have to consider the following two cases: (i) Transfers where the donor branch loses its gene copy (that is the gene copy of the donor left no trace in the contemporary species; i.e. it became extinct; TL event, in short); and (ii) scenarios with one (or more) speciation/duplication node $u$ located below its LCA vertex of $S$. Due to the possibility of transfers, forbidding case (ii) is no longer sure to be optimal. In Figure 5, for instance, $R1$ is more parsimonious than $R2$ due to points (i and ii) above. In $R1$, the gene lineage $(u,b)$ follows a TL event from $(x,A)$ toward $(z,B)$ and node $w$ is a speciation located below its LCA mapping (vertex $y$ of $S$).

Similar to [19], [36] developed a probabilistic mixed model (see the last paragraph of section ‘Probabilistic models’) and an MCMC approach to estimate posterior probabilities of gene trees and DT rates. The algorithm formulated to compute the probability of $G$, given parameters for the mixed model, is a major contribution here.

According to the features introduced above, Table 1 summarizes the pros and cons of the five reconciliation models presented in [18, 36, 57–60].

In contrast to DL models, where the LCA reconciliation is the sole MPR, several optimal reconciliations are possible for DTL models. This multiplicity may undermine the confidence we have in a single reconciliation drawn at random by a program. In this context, algorithms that enumerate all MPRs have been proposed [55, 59, 61].

**AVAILABLE PROGRAMS**

**Gene tree reconstruction with duplications and losses (DL model)**

$SPIMAP$ [62] and $PrimeGSR$ [63] are two gene tree reconstruction programs that implement the Bayesian frameworks, respectively, developed in...
Both programs take as input a dated species tree S in Newick format and aligned sequences in Fasta format. DL rates are estimated a priori by SPIMAP using a method similar to that of CAFE [53], while the rates are sampled during the MCMC implemented in PrimeGSR. Substitution rates are a priori estimated by SPIMAP, while PrimeGSR uses iid to model substitution rate variations [64]. Moreover, SPIMAP optionally performs bootstrapping and outputs the best reconciliation found.

**Reconciliation with duplications and losses (DL model)**

TreeMap [65] was the first program developed for reconciling a gene tree G with a species tree S. A graphical interface is provided with a number of options. However, it does not deal with dates for nodes of S, and as such cannot ensure the time consistency of transfers. Notung [20] reconciles G and S according to the DL model, where at least one of the trees is binary. It has an interface that displays orthology relationships and a command line version. It can also root G and resolve its polytomies (i.e. nodes with low support) by minimizing the parsimony score. The algorithm inferring duplications and speciations based on LCA mapping [31] is implemented in [66], which also roots G by minimizing the sum of inferred duplications.

Algorithms exploring the reconciliation space [34] and the probabilistic framework [47] are implemented in a program called Korak [67]. Given a tree G, a dated tree S, and DL rates, it computes the number of reconciliations, the likelihood of the LCA-based reconciliation, and the (exact/approximate) posterior probability of each visited reconciliation.

**Reconciliation with duplications, transfers and losses (DTL model)**

In order to compute reconciliations with consistent transfers, dates for nodes of S can be obtained by relaxed molecular clock techniques working from molecular sequences [64, 68, 69].

The reconciliation approach of [57] has been implemented in CoRe-Pa [60]. This software includes a reconciliation viewer, an editor for modifying G and S, as well as resampling facilities evaluation of the statistical relevancy of an MPR. It does not require inputted costs for the three evolutionary events. Instead, it tries to estimate them based on observed event frequencies.

The model of [58] has been implemented in a program called Jane [70], which also includes resampling facilities. In addition, it allows visual editing of a reconciliation (its cost is updated accordingly), and can be run from the command-line (for large-scale experiments). Reconciliations are built for a dated tree S, whose dates can be provided by the user. Alternatively, Jane uses a genetic algorithm to find optimal dates (with respect to reconciliation costs). Jane also enables control of the maximal distance between two species that can exchange genes. The latter option is especially relevant for co-evolution studies [25].

<table>
<thead>
<tr>
<th>Input gene/species trees</th>
<th>Model characteristics</th>
<th>Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree G nG nodes</td>
<td>Tree S nS nodes</td>
<td>Transfer with loss</td>
</tr>
<tr>
<td>Merkle et al. [60] and Merkle and Middendorf [57]</td>
<td>Binary or polytomous, time interval</td>
<td>Binary, dated</td>
</tr>
<tr>
<td>Libeskind-Hadas and Charleston [58]</td>
<td>Binary</td>
<td>Network, dated</td>
</tr>
<tr>
<td>Tofigh [36]</td>
<td>Binary</td>
<td>Binary, dated</td>
</tr>
<tr>
<td>Doyon et al. [59]</td>
<td>Binary</td>
<td>Binary, dated</td>
</tr>
<tr>
<td>David and Alm [18]</td>
<td>Binary</td>
<td>Binary, dated</td>
</tr>
</tbody>
</table>

The models of [36] and [18, 57–60] are continuous and discrete, respectively.

**Table 1:** Comparison of five reconciliation models accounting for duplications, losses and transfers
The reconciliation approach of [59] is implemented in a command-line program called Mowgli [71]. Mowgli computes an MPR and the number of equally optimal reconciliations. This provides an alternative (and usually much faster) way to measure the statistical significance of the returned MPR. The method of [18] is also implemented in a command-line program called AnGST. It deals with phylogenetic uncertainties in gene phylogenies by inferring G as a combination of bootstrap subtrees to yield the reconciliation of minimal cost.

We considered a species tree of 90 genomes (11 eukaryotic, 12 archaeal and 67 bacterial), a gene family tree of 4 genes, and costs of 2, 3 and 1, respectively, for a duplication, transfer and loss [18]. Figure 6 displays the reconciliations proposed by Mowgli, AnGST and Jane. Mowgli finds a reconciliation that is more parsimonious than those inferred by the other two software packages. As the three reconciliations differ according to the number and kind of events, the different models allow parsimony optimization at different degrees.

We note that some of the above software can differ from the models presented in the associated paper. For instance, on several datasets from [18] the reconciliations proposed by CoRe-Pa and AnGST have speciations located below the LCA mapping, while they are not supposed to [18, 57, 60].

Note finally that, although parsimony is the fastest approach to reconcile gene trees, there are cases where several most parsimonious reconciliations exist. The prevalence of this effect has not yet been measured and certainly deserves further attention.

**Key Points**
- Reconciliation is an approach used to depict the evolution of a gene family with respect to the evolution of the species.
- Several reconciliation models based on parsimony and probabilistic criteria have been proposed.
- Several DTL models have been proposed. They differ in their way to handle the time consistency of transfer and in their degree of generality.

**FUNDING**
The French Agence Nationale de la Recherche ‘Domaines Emergents’ (ANR-08-EMER-011, ‘PhylAriane’) and by the Languedoc-Roussillon ‘Chercheur d’Avenir’ program. This publication is contribution No ISEM 2011-077 of the Institut des Sciences de l’Evolution de Montpellier (UMR 5554 - CNRS).
References