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

Corresponding author. Vincent Berry, Université Montpellier 2, Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, France. Tel: 0(33) 46741 8548; Fax: 0(33) 46741 8500; E-mail: vberry@lirmm.fr

Jean-Philippe Doyon received the PhD degree in computer science from the University of Montréal. His research focuses on comparative genomics, especially combinatorial and probabilistic aspects of reconciliations between gene and species trees.

Vincent Ranwez worked almost 3 years in a biopharmaceutical start-up, before joining up the ISEM laboratory as an assistant professor. His main research topics are phylogenomics, supertree methods and tree reconciliation.

Vincent Daubin received his Ph.D. in evolutionary genomics at the University of Lyon, France. He is now a Centre National de la Recherche Scientifique (CNRS) researcher in the Laboratory of Biometry and Evolutionary Biology in Lyon. His research focuses on exploring phylogenomic approaches to understand the evolution of genomes.

Vincent Berry is Professor at the Polytech school of the University of Montpellier. He performs researches in phylogenomics, and on algorithmic aspects of trees in a computational biology context.

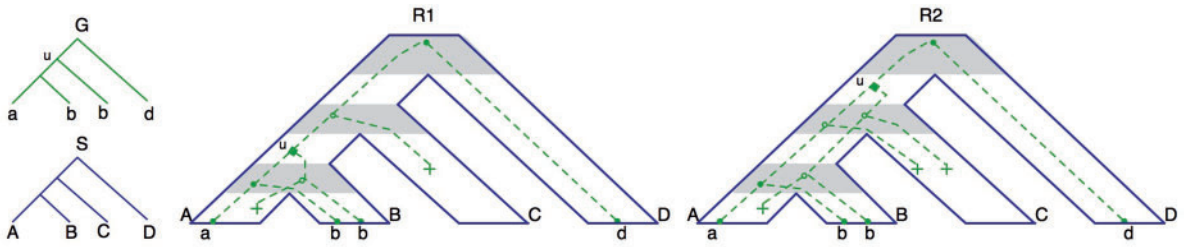

Figure 1: Drawing representation of a reconciliation. A species tree S and a gene tree G, where each lower case letter denotes an observed gene of an extant species (gene 'a' belongs to species 'A', etc). R1 and R2 are two reconciliations that embed G into S. A grayed zone (resp. tube) corresponds to a vertex (resp. branch) of S and G is depicted with dotted lines. Nodes of the embedded tree represent: duplication (lozenge), loss (+), speciation either present in G (filled circle) or not (open circle). Observe that node *u* of G is a duplication for R1 and R2, although located on different branches of S.

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 polynomial 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,

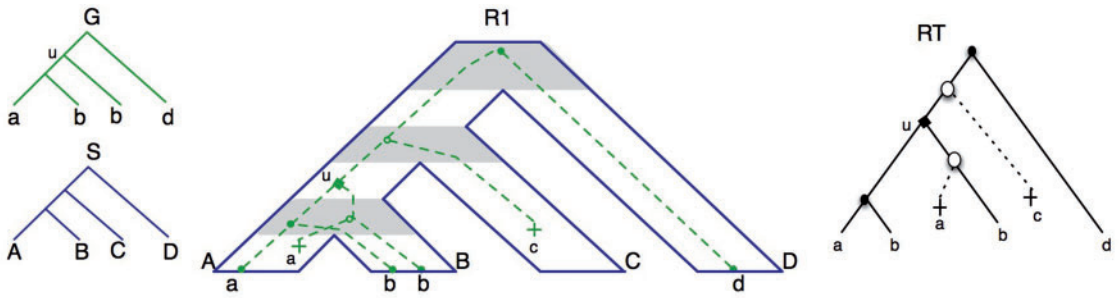

Figure 2: Two alternative representations of a reconciliation. RT is the reconciled tree corresponding to reconciliation RI. Dotted edges of RT represent losses and their removal leads to the original gene tree G.

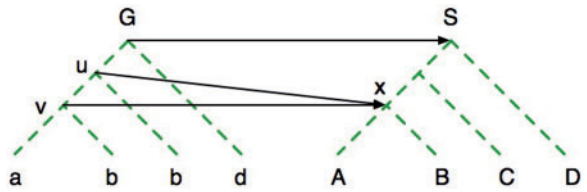

Figure 3: The LCA mapping between trees G and S. Ancestral node *u* of G is inferred to be a duplication since it maps to the same node *x* in S as one of its children (child v). (See also Figure 2; RI).

u 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, R1 of Figure 1 is the LCA reconciliation, while R2 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

departing 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 $P(R|G) = P(G,R)/P(G)$ in time $O(n_G^2 n_S)$. This posterior probability is useful to evaluate the reliability of the most likely scenario with respect to other reconciliations, particularly those having near-optimal likelihood. A similar algorithm computes the probability that a given ancestral gene of G is a speciation in time $O(n_G^2 n_S)$. By sampling duplication and loss rates [45] developed an MCMC algorithm that uses the latter one to estimate posterior probabilities of orthology relationships among sequences of a gene family. This method is implemented in a program called *PrimeGEM* [46].

The reconciliation space exploration algorithm of [34] is used in [47] to compute the exact posterior probability P(R|G) of each visited reconciliation 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 $O(n_G^4 n_S)$ [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 $O(n_S^2 n_G)$, [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 a 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).

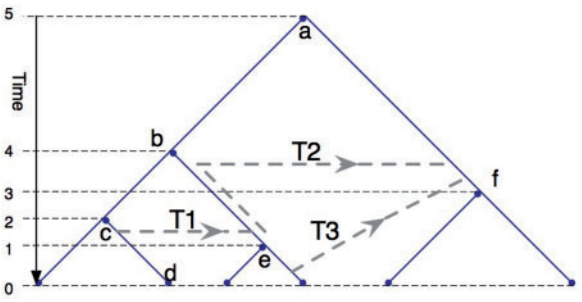

Figure 4: Time (in)consistencies with respect to transfers. A species tree S with a relative time order on its vertices (i.e. 0 is the present time and vertex c is at time 2). The three arrowed dotted lines represent transfers between two branches of S and the arrows their directions [e.g. T1 has branch (c, d) as donor and (b, e) as receiver]. T1 and T2 are both (individually) locally consistent, in contrast to T3. However, together T1 and T2 are not globally consistent since given the drawn reconciliation, the fact that T1 precedes T2 (in the gene tree) implies that the donor of T2 precedes (in time) the recipient of T1.

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

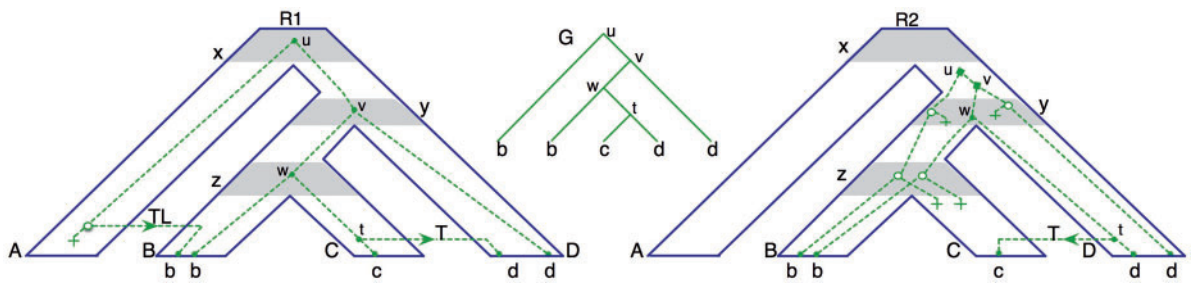

Figure 5: Model limitations with respect to optimal reconciliations. Consider a cost of 0 for speciation, a cost of 1 for other events, and the two reconciliations R1 (costing 3) and R2 (costing 7). If cases (i and ii) above are allowed, both R1 and R2 are valid reconciliations and thus R2 is not an MPR. If both cases (i and ii) are forbidden, R1 is not a valid reconciliation and R2 then becomes an MPR.

Table 1: Comparison of five reconciliation models accounting for duplications, losses and transfers

	Input gene/species trees		Model characteristics		Algorithm	
	Tree G n_G nodes	**Tree S n_S nodes**	**Transfer with loss**	**Location of spec./dup.**	**Global consistency**	**Time complexity**
Merkle *et al.* [60] and Merkle and Middendorf [57]	Binary or polytomous, time interval	Binary, dated	No	Only on/above the LCA	Not guaranteed (considered *a posteriori*)	$O(\max(n_S,n_G)^3)$
Libeskind-Hadas and Charleston [58]	Binary	Network, dated	No	Anywhere	Guaranteed, with time slices	Polynomial in n_S,n_G
Tofigh [36]	Binary	Binary, dated	Yes	Anywhere	Guaranteed, with time slices	$O(n_S^5 n_G)$
Doyon *et al.* [59]	Binary	Binary, dated	Yes	Anywhere	Guaranteed, with time slices	$O(n_S^2 n_G)$
David and Alm [18]	Binary	Binary, dated	No	Only on/above the LCA	Guaranteed, with simple rules	$O(n_S^3 n_G)$

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

[21] and [19]. 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].

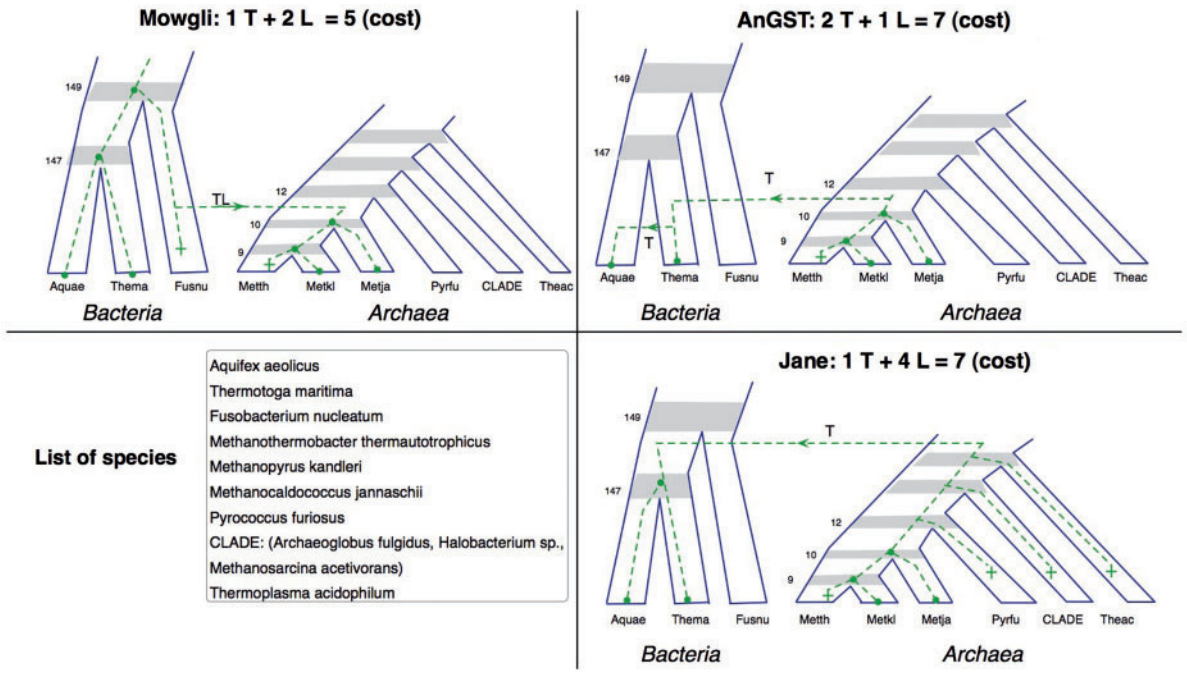

Figure 6: Importance of considering TL events. A gene family where *Mowgli* outperforms *AnGST* and *Jane* (COG 1542). The cost of a Duplication, a Transfer, resp. a Loss is 2, 3, resp. 1 [18]. Depending on whether TL events are considered or not, transfer happen from Bacteria to Archaea or in the opposite direction.

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

1. Bansal M, Eulenstein O. The multiple gene duplication problem revisited. *Bioinformatics* 2008;**24**(13): i132–8.
2. Treangen T, Rocha E. Horizontal transfer, not duplication, drives the expansion of protein families in Prokaryotes. *PLoS Genet* 2011;**7**(1):e1001284.
3. Boussau B, Daubin V. Genomes as documents of evolutionary history. *Trends Ecol Evol* 2010;**25**:224–32.
4. Goodman M, Czelusniak J, Moore G, *et al*. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst Zool* 1979;**28**(2):132–63.
5. Guigo R, Muchnik I, Smith T. Reconstruction of ancient molecular phylogeny. *Mol Phylogenet Evol* 1996;**6**(2): 189–213.
6. Page R. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst Biol* 1994;**43**(1):58.
7. Abby S, Tannier E, Gouy M, *et al*. Detecting lateral gene transfers by statistical reconciliation of phylogenetic forests. *BMC Bioinformatics* 2010;**11**(1):324.
8. Beiko R, Hamilton N. Phylogenetic identification of lateral genetic transfer events. *BMC Evol Biol* 2006;**6**(1):15.
9. Nakhleh L, Ruths F, Wang LS. RIATA-HGT: A Fast and Accurate Heuristic for Reconstructing Horizontal Gene Transfer. *Proc 11th Int Comput Combinatorics Conf* 2005. (LNCS 3595): 84–93.
10. Dufayard J-F, Duret L, Penel S, *et al*. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* 2005;**21**(11):2596–603.
11. Storm C, Sonnhammer E. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics* 2002;**18**(1):92.
12. Van Der Heijden R, Snel B, Van Noort V, *et al*. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics* 2007;**8**(1):83.
13. Wapinski I, Pfeffer A, Friedman N, *et al*. Natural history and evolutionary principles of gene duplication in fungi. *Nature* 2007;**449**(7158):54–61.
14. Zmasek C, Eddy S. RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 2002;**3**(1):14.
15. Chen F, Mackey A, Vermunt J, *et al*. Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE* 2007;**2**(4):e383.
16. Altenhoff AM, Dessimoz C. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol* 2009;**5**(1):e1000262.
17. Rivera A, Pankey M, Plachetzki D, *et al*. Gene duplication and the origins of morphological complexity in pancrustacean eyes, a genomic approach. *BMC Evol Biol* 2010; **10**(1):123.
18. David LA, Alm EJ. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature* 2011;**469**(7328): 93–6.
19. Akerborg O, Sennblad B, Arvestad L, *et al*. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA* 2009;**106**(14):5714–9.
20. Durand D, Halldórsson B, Vernot B. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J Comput Biol* 2006;**13**(2):320–35.
21. Rasmussen M, Kellis M. A bayesian approach for fast and accurate gene tree reconstruction. *Mol Biol Evol* 2011;**28**(1): 273–90.
22. Wapinski I, Pfeffer A, Friedman N, *et al*. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 2007;**23**(13):i549.
23. Sanderson M, McMahon M. Inferring angiosperm phylogeny from EST data with widespread gene duplication. *BMC Evol Biol* 2007;**7(Suppl 1)**:S3.
24. Page R, Charleston M. Trees within trees: phylogeny and historical associations. *Trends Ecol Evol* 1998; **13**(9):356–9.
25. Nieberding C, Jousselin E, Desdevises Y. The use of co-phylogeographic patterns to predict the nature of interactions, and vice-versa. In: Morand S, Krasnov B, (eds). *The Geography of Host-parasite Interactions*. New York: Oxford University Press, 2010.
26. Brooks D, Ferrao A. The historical biogeography of co-evolution: emerging infectious diseases are evolutionary accidents waiting to happen. *J Biogeography* 2005;**32**(8): 1291–9.
27. Bonizzoni P, Vedova G, Dondi R. Reconciling a gene tree to a species tree under the duplication cost model. *Theor Comput Sci* 2005;**347**(1-2):36–53.
28. Górecki P, Tiuryn J. DLS-trees: a model of evolutionary scenarios. *Theor Comput Sci* 2006;**359**(1-3):378–99.
29. Zhang L. On a Mirkin-Muchnik-Smith conjecture for comparing molecular phylogenies. *J Comput Biol* 1997;**4**(2): 177–87.
30. Mirkin B, Muchnik I, Smith T. A biologically consistent model for comparing molecular phylogenies. *J Comput Biol* 1995;**2**(4):493–507.
31. Zmasek C, Eddy S. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics* 2001;**17**(9):821.
32. Bender MA, Farach-Colton M. The LCA Problem Revisited. In: Gonnet H. Gaston, Daniel Panario, Alfredo Viola (eds). *Proceedings of the 4th Latin American Symposium on Theoretical Informatics* (LATIN '00). London, UK: Springer-Verlag, 2000;88–94.
33. Arvestad L, Berglund A, Lagergren J, *et al*. Gene tree reconstruction and orthology analysis based on an integrated model for duplications and sequence evolution. *Proc Eighth annu Int Conf Res Comput Mol Biol* 2004;326–35.
34. Doyon JP, Chauve C, Hamel S. Space of gene/species trees reconciliations and parsimonious models. *J Comput Biol* 2009;**16**(10):1399–418.
35. Chauve C, El-Mabrouk N. New perspectives on gene family evolution: losses in reconciliation and a link with supertrees. *Res Comput Mol Biol* 2009;**5541**:46–58.
36. Tofigh A. Using trees to capture reticulate evolution, lateral gene transfers and cancer progression. PhD thesis. Sweden: KTH Royal Institute of Technology, 2009.
37. Hahn M. Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution. *Genome Biol* 2007;**8**(7):R141.
38. Berglund-Sonnhammer AC, Steffansson P, Betts M, *et al*. Optimal gene trees from sequences and species trees using

a soft interpretation of parsimony. *J Mol Evol* 2006;**63**(2): 240–50.

39. Chang W, Eulenstein O. Reconciling gene trees with apparent polytomies. Proc COCOON 2006 (LNCS 4112): p. 235–44.
40. Vernot B, Stolzer M, Goldman A, *et al*. Reconciliation with non-binary species trees. *J Comput Biol* 2008;**15**(8): 981–1006.
41. Wapinski I, Pfeffer A, Friedman N, *et al*. Automatic genome-wide reconstruction of phylogenetic gene trees. *Bioinformatics* 2007;**23**(13):i549.
42. Arvestad L, Berglund AC, Lagergren J, *et al*. Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics* 2003;**19(Suppl 1)**:i7–15.
43. Novozhilov A, Karev G, Koonin E. Biological applications of the theory of birth-and-death processes. *Brief Bioinform* 2006;**7**(1):70.
44. Arvestad L, Lagergren J, Sennblad B. The gene evolution model and computing its associated probabilities. *J ACM* 2009;**56**(2):1–44.
45. Sennblad B, Lagergren J. Probabilistic orthology analysis. *Syst Biol* 2009;**58**(4):411.
46. Sennblad B, Lagergren J. PrimeGEM - probabilitic orthology analysis. 2009. http://prime.sbc.su.se/primeGEM (8 September 2011, date last accessed).
47. Doyon JP, Hamel S, Chauve C. An Efficient Method for Exploring the Space of Gene Tree/Species Tree Reconciliations in a Probabilistic Framework. *IEEE/ACM Trans Comput Biol Bioinformat* 2011;**99** (PrePrints).
48. Górecki P, Burleigh GJ, Eulenstein O. Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinformatics* 2011; **12(Suppl 1)**:S15.
49. Górecki P, Eulenstein O. DrML - Maximum Likelihood Estimation in the Duplication Loss Model 2010. http://bioputer.mimuw.edu.pl/~gorecki/drml (8 September 2011, date last accessed).
50. Csűrös M, Miklós I. Streamlining and large ancestral genomes in Archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol* 2009;**26**(9):2087.
51. Kumar S. Molecular clocks: four decades of evolution. *Nat Rev Genet* 2005;**6**(8):654–62.
52. Rasmussen M, Kellis M. Accurate gene-tree reconstruction by learning gene-and species-specific substitution rates across multiple complete genomes. *Genome Res* 2007; **17**(12):1932.
53. De Bie T, Cristianini N, Demuth J, *et al*. CAFE: a computational tool for the study of gene family evolution. *Bioinformatics* 2006;**22**(10):1269–71.
54. Ovadia Y, Fielder D, Conow C, *et al*. The cophylogeny reconstruction problem is NP-complete. *J Comput Biol* 2011;**18**:59–65.
55. Tofigh A, Hallett M, Lagergren J. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans Comput Biol Bioinform* 2011;**8**:517–35.
56. Charleston M, Perkins S. Traversing the tangle: algorithms and applications for cophylogenetic studies. *J Biomed Inform* 2006;**39**(1):62–71.
57. Merkle D, Middendorf M. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory Biosci* 2005;**123**(4):277–99.
58. Libeskind-Hadas R, Charleston M. On the computational complexity of the reticulate cophylogeny reconstruction problem. *J Comput Biol* 2009;**16**(1):105–17.
59. Doyon JP, Scornavacca C, Szöllősi GJ, *et al*. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *Proc 14th Int Conf Res Comput Mol Biol (RECOMB-CG) 2011. Volume 6398 of LNCS*; 93–108.
60. Merkle D, Middendorf M, Wieseke N. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics* 2010;**11(Suppl 1)**:S60.
61. Charleston M. Jungles: a new solution to the host/parasite phylogeny reconciliation problem. *Mathl Biosci* 1998;**149**(2): 191–223.
62. Rasmussen M. SPIMAP - Species Informed Maximum A Posteriori Gene Tree Reconstruction. 2011. http://compbio.mit.edu/spimap (8 September 2011, date last accessed).
63. Akerborg O, Sennblad B, Arvestad L, *et al*. PrIME-GSR: a Bayesian integrated model for genes, sequences, and rates. 2008. http://prime.sbc.su.se/primeGSR (8 September 2011, date last accessed).
64. Akerborg O, Sennblad B, Lagergren J. Birth-death prior on phylogeny and speed dating. *BMC Evol Biol* 2008;**8**:77.
65. Charleston MA, Page RDM. TreeMap 3 program. 2002. http://sydney.edu.au/engineering/it/~mcharles/software/treemap/treemap3.html (8 September 2011, date last accessed).
66. Zmasek C, Eddy S. 2010. http://www.phylosoft.org/forester/applications/sdi (8 September 2011, date last accessed).
67. Doyon JP, Hamel S, Chauve C. Korak - Exploring the Space of Reconciliations and Probabilistic Framework. http://www.lirmm.fr/~doyon/Korak (8 September 2011, date last accessed).
68. Lartillot N, Lepage T, Blanquart S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 2009;**25**:2286–8.
69. Sanderson M. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 2003;**19**(2):301.
70. Conow C, Fielder D, Ovadia Y, *et al*. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol Biol* 2010;**5**:16.
71. Doyon JP, Scornavacca C, Szöllő si GJ, *et al*. Mowgli program. 2011. http://www.atgc-montpellier.fr/Mowgli/ (8 September 2011, date last accessed).