Bioinformatics, 31(6), 2015, 841–848
doi: 10.1093/bioinformatics/btu728
Advance Access Publication Date: 6 November 2014
Original Paper

OXFORD

Phylogenetics

Joint amalgamation of most parsimonious reconciled gene trees

Celine Scornavacca[1,2,*], Edwin Jacox[1] and Gergely J. Szöllősi[3,*]

[1]ISEM, UM2-CNRS-IRD, Place Eugène Bataillon 34095 Montpellier, France, [2]Institut de Biologie Computationnelle (IBC), 95 rue de la Galéra, 34095 Montpellier, France and [3]ELTE-MTA 'Lendület' Biophysics Research Group 1117 Bp., Pázmány P. stny. 1A., Budapest, Hungary

*To whom correspondence should be addressed.
Associate Editor: David Posada

Received on May 27, 2014; revised on October 28, 2014; accepted on October 29, 2014

Abstract

Motivation: Traditionally, gene phylogenies have been reconstructed solely on the basis of molecular sequences; this, however, often does not provide enough information to distinguish between statistically equivalent relationships. To address this problem, several recent methods have incorporated information on the species phylogeny in gene tree reconstruction, leading to dramatic improvements in accuracy. Although probabilistic methods are able to estimate all model parameters but are computationally expensive, parsimony methods—generally computationally more efficient—require a prior estimate of parameters and of the statistical support.

Results: Here, we present the Tree Estimation using Reconciliation (TERA) algorithm, a parsimony based, species tree aware method for gene tree reconstruction based on a scoring scheme combining duplication, transfer and loss costs with an estimate of the sequence likelihood. TERA explores all reconciled gene trees that can be amalgamated from a sample of gene trees. Using a large scale simulated dataset, we demonstrate that TERA achieves the same accuracy as the corresponding probabilistic method while being faster, and outperforms other parsimony-based methods in both accuracy and speed. Running TERA on a set of 1099 homologous gene families from complete cyanobacterial genomes, we find that incorporating knowledge of the species tree results in a two thirds reduction in the number of apparent transfer events.

Availability and implementation: The algorithm is implemented in our program *TERA*, which is freely available from http://mbb.univ-montp2.fr/MBB/download_sources/16__TERA.

Contact: celine.scornavacca@univ-montp2.fr, ssolo@angel.elte.hu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Molecular phylogenetics infers gene trees based on the information contained in molecular sequences. Unfortunately, individual sequences may contain limited signal, and, as a result, phylogenetic reconstruction often involves choosing between statistically equivalent or weakly distinguishable evolutionary relationships.

Although each homologous gene family has its own unique story, these are all related by a shared species history—which can be helpful for gene tree inference (Maddison, 1997; Szöllősi *et al.*, 2014). In the past decade, several methods have been developed that model the evolutionary processes that generate gene trees within the species tree (Akerborg *et al.*, 2009; Arvestad, 2003; Hallett and Lagergren, 2001; Rannala and Yang, 2003; Rasmussen and Kellis, 2007, 2012; Sjöstrand *et al.*, 2014; Suchard, 2005; Szöllősi *et al.*, 2012, 2013a; Than and Nakhleh, 2009). From an inference perspective, these methods attempt to find the optimal way to explain the phylogenetic signal in extant sequences—represented as a gene tree—given the species tree. They explore the set of possible *reconciliations*, i.e. different ways to draw the gene tree into the species tree given some combination of macro evolutionary events, such as

gene duplications, gene transfers, gene losses and incomplete lineage sorting. Studies that incorporate such events into gene tree inference have shown that information on the species phylogeny significantly improves the accuracy of gene tree inference (Akerborg *et al.*, 2009; Boussau *et al.*, 2013; Rasmussen and Kellis, 2010; Szöllősi *et al.*, 2013b). To design such *species tree aware methods* for reconstructing gene phylogenies, the space of reconciled gene trees must be explored using information from both a model of *sequence evolution* and a *reconciliation* model, in order to optimize a *joint* sequence-reconciliation score. Such exploration is computationally expensive with traditional optimization approaches that rely on the local search of the space of gene trees.

To circumvent this problem, David and Alm (2010) introduced the amalgamation algorithm, described in detail in Section 2.3 below and illustrated in Figure 1. Furthermore, Szöllősi *et al.* (2013b) recently developed an approach to exhaustively explore all reconciled gene trees that can be *amalgamated* from a sample of gene trees, i.e. obtainable by combining clades observed in the sample. Additionally, their method—ALE, for Amalgamated Likelihood Estimation—combines the amalgamation algorithm of David and Alm (2010) with conditional clade probabilities (CCPs) introduced by Höhna and Drummond (2012) and reconstructs the gene phylogenies by optimizing a joint sequence-reconciliation likelihood score, resulting in gene trees that are dramatically more accurate than those reconstructed using molecular sequences alone.

ALE overcomes a fundamental limitation of recent parsimony based methods that improve gene trees given a putative species tree (David and Alm, 2010; Nguyen *et al.*, 2012; Wu *et al.*, 2013). Unlike those methods, it does not require the user to specify a cost for each type of event or a threshold on statistical supports. However, ALE faces the drawbacks associated with probabilistic methods. In particular: (i) when computing the reconciliation score, ALE has an increased computational cost compared with a parsimony algorithm (e.g. Conow *et al.*, 2010; Doyon *et al.*, 2010), which is due to a potentially large constant factor resulting from the numerical integration of the likelihood; (ii) ALE's results are contingent on difficult to estimate time-like branch lengths of the species tree, while parsimony methods can reconcile gene trees relying only on the order of speciations in time (e.g. Doyon *et al.*, 2010), and even deal with undated species trees (e.g. Bansal *et al.*, 2012). Parsimony methods in general, despite lacking an explicit connection to a generative probabilistic model and relying on other heuristics, have been shown to be highly accurate, comparable to sophisticated probabilistic reconciliation methods, with reduced runtime (Wu *et al.*, 2013, 2014).

Here we present the TERA algorithm (Tree Estimation using Reconciliation and Amalgamation) that amalgamates the most parsimonious reconciled gene tree from a set of gene trees reconstructed from a unique gene alignment, according to a joint sequence-reconciliation score. Although TERA, like other parsimony based methods, requires the prior specification of the costs associated with duplication, transfer and loss (DTL) events, it does not require prior assumptions about a statistical support threshold, as it estimates a self-consistent support threshold from its input. Furthermore, TERA considers explicitly the possibility of transfer from extinct or unsampled branches of the species tree, which is expected to be the case for practically all transfers (Szöllősi *et al.*, 2013a). TERA does not, however, consider incomplete lineage sorting.

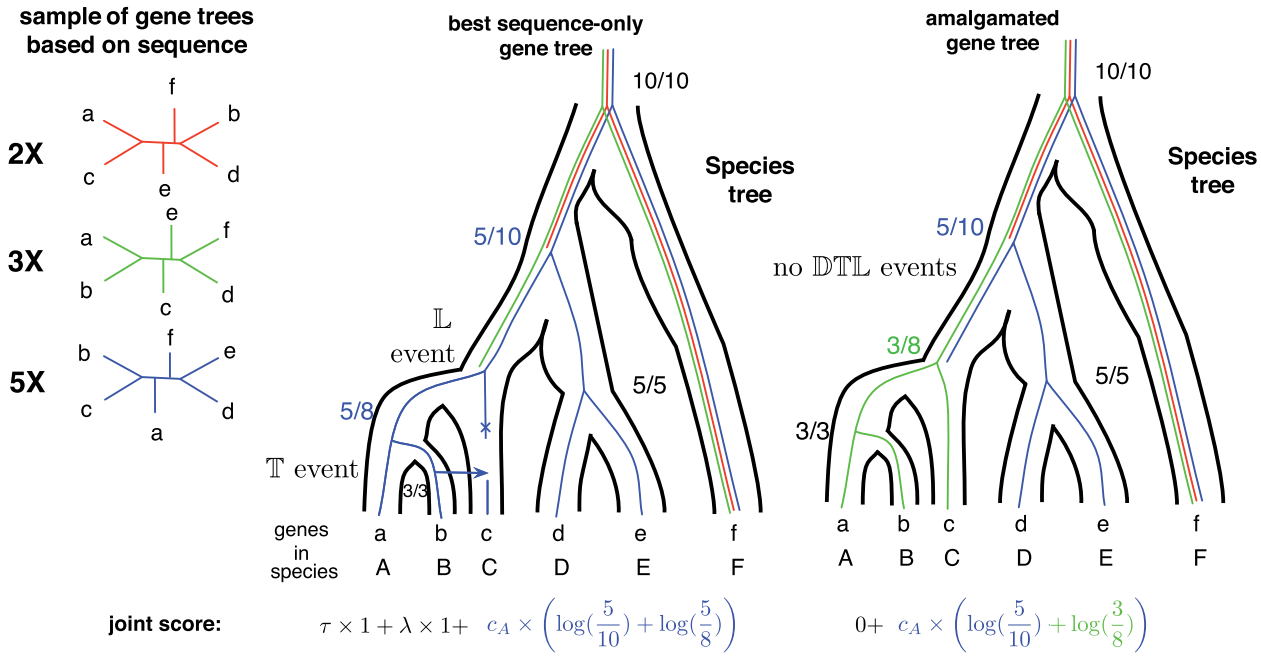

Fig. 1. CCPs can be used to estimate the posterior probability of any tree that can be *amalgamated* from clades present in a sample of gene trees (David and Alm, 2010; Höhna and Drummond, 2012). Conditional clade frequencies can be used to approximate CCPs and are computed as the proportion of occurrences of a particular split of a clade according to a tripartition π, e.g. (*abc* | *de*) among all trees in which the clade, e.g. (*abcde*), is found. Estimates based on the sample of trees on the left are shown as fractions for two different gene trees that can be amalgamated. The estimate for a gene tree is given by the sum of the reconciliation score and the logarithm of the tree CCPs. Based on the sample on the left, the tree with the highest posterior probability is the third tree (blue online). Reconciling it with the species tree requires one transfer and one loss event. It is, however, possible to combine clades present in the second (green online) and third (blue online) trees to produce a gene tree that is not present in the original sample but is identical to the species tree, *i.e.* it requires no events to draw it into the species tree. Depending on the costs of transfer and loss events, and the self-consistently estimated c_A parameter, the scenario without transfer might be optimal for the joint score

The self-consistent score estimation scheme used by TERA, introduced in Section 2.4, should be applicable to other parsimony methods, while amalgamation is in theory compatible with any reconciliation algorithm that assumes branches of the gene tree to be independent.

2 Materials and methods

2.1 Preliminaries

Given a binary rooted tree T, we respectively denote by $V(T)$, $E(T)$, $L(T)$ and $r(T)$, its node, edge, leaf node sets and root node. The label of each leaf u is denoted by $\mathcal{L}(u)$, while the set of labels of leaves of T is denoted by $\mathcal{L}(T)$. Given a node $u \in V(T)$, we denote respectively by u_p, u_s and $\{u_l, u_r\}$ the father, the sibling and the children of u (if they are defined). Note that in this article all trees are considered as unordered, so u_l and u_r are interchangeable. For a node u of T, T_u denotes the subtree of T rooted at u. Given two nodes u and v of T, $u \leq_T v$ ($u <_T v$, respectively) if and only if v is on the unique path from u to $r(T)$ (respectively, and $u \neq v$); in such a case, u is said to be a (strict) descendant of v. Given a node u of T, we define the *clade* associated with u, denoted by $C(u)$, as the set $\mathcal{L}(T_u)$. If u is an internal node, we define the *tripartition* associated with u, denoted by $\Pi(u)$, as the triple $(\mathcal{L}(T_u), \mathcal{L}(T_{u_l}), \mathcal{L}(T_{u_r}))$. For leaf nodes, the trivial tripartition $\Pi(u)$ is defined as the triple $(\mathcal{L}(u), \emptyset, \emptyset)$. Finally, the height of T is denoted by $h(T)$.

In this article, unless stated otherwise, we assume that gene and species trees are rooted, binary and uniquely leaf labeled, i.e. within each tree there is a bijection between leaves and labels. Due to this bijectivity we will refer to leaves and labels interchangeably.

We define a gene tree G as a tree where each leaf represents an extant gene. Similarly, a species tree S is defined as a tree in which each leaf represents a distinct extant species. Note that several leaves of a gene tree can be associated to the same species due to gene birth corresponding to duplication and transfer events. Formally, we indicate this by a surjective function $s : \mathcal{L}(G) \to \mathcal{L}(S)$, called the *species labeling* of G. The set of species labels of the leaves of G is denoted $\mathcal{S}(G)$.

A tree T is said to be *time ordered* when it is associated with a *time function* $\theta_T : V(T) \to \mathbb{R}^+$ that associates each of its nodes with a non-negative value so that, for any two nodes $x, x' \in V(T)$, if x' is a strict descendant of x then $\theta_T(x') < \theta_T(x)$. Moreover, $\forall x \in L(T)$, we have that $\theta_S(x) = 0$. A *subdivision* T' of a time-ordered tree T is the tree obtained from T by adding a new node y with $\theta_{T'}(y)$ on each edge $(x_p, x) \in E(T)$ such that there exists $z \in V(T)$ with $\theta_T(x) < \theta_T(z) < \theta_T(x_p)$. For nodes $x \in V(T')$ corresponding to nodes already present in T, we set $\theta_{T'}(x) = \theta_T(x)$.

2.2 Species tree-gene tree reconciliation

Here, we consider the problem of finding the most parsimonious reconciliation (MPR) when considering—as possible macro-events that result in the birth and death of gene copies—speciation, gene duplication, gene transfers and gene loss (Szöllősi and Daubin, 2012). The general problem of finding an MPR is known to be NP-complete, even for reconciling two binary trees (Tofigh *et al.*, 2011). The complexity of the problem is due to the difficulty of ensuring the *time consistency* of gene transfers, i.e. satisfying the chronological constraints among nodes of the species tree that are induced by transfer events. However, the problem becomes polynomially solvable when accepting a time-ordered species tree as input (among others Conow *et al.*, 2010; Doyon *et al.*, 2010 Tofigh, 2009, see Doyon *et al.*, 2011 for a review). In this article, we build upon the combinatorial reconciliation model introduced by Doyon *et al.* (2010), which can be used to solve this special case of the problem.

Some parsimony methods (e.g. Bansal *et al.*, 2012) do not need information on the order of speciations in time. This allows a more efficient recursion over reconciliations, but at the cost of considering reconciliations that contain transfer events that are not consistent with any ordering of the species tree (Tofigh *et al.*, 2011).

The $\mathbb{DTL}$ model of Doyon *et al.* (2010) can be used to reconcile a time-ordered binary species tree S with a binary gene tree G by constructing a mapping α that maps each node $u \in V(G)$ into an ordered list of nodes in $V(S')$, namely the ancestral and/or extant species in which the sequence corresponding to u evolved. This model takes into account four kinds of biological events: speciation, gene duplication, gene transfer and gene loss. The atomic events of this model are: a speciation ($\mathbb{S}$), a duplication ($\mathbb{D}$), a transfer ($\mathbb{T}$), a transfer followed immediately by the loss of the non-transferred child ($\mathbb{TL}$), a speciation followed by the loss of one of the two resulting children ($\mathbb{SL}$), and a contemporary event ($\mathbb{C}$) that associates an extant gene to its corresponding species. Finally, a null event ($\emptyset$), is used to model a gene lineage crossing a time boundary. Note that duplication-loss events and transfer followed by the loss of the transferred gene, unlike a transfer followed by the loss of the non-transferred gene and speciation-loss events, leave no trace and are therefore undetectable. This is why, in the $\mathbb{DTL}$ model, losses are never considered alone. We refer the reader to Doyon *et al.* (2010) for the formal definition of a $\mathbb{DTL}$ reconciliation.

Let θ, τ, λ be, respectively, the costs of a duplication, a transfer and a loss. Given a $\mathbb{DTL}$ reconciliation, we define the cost of α, denoted by $c(\alpha)$, as the sum $\theta d + \tau t + \lambda l$, where d, t and l are respectively the number of $\mathbb{D}$ events, of $\mathbb{T}$ and $\mathbb{TL}$ events, and of $\mathbb{SL}$ and $\mathbb{TL}$ events in α. In Doyon *et al.* (2010) the authors give an efficient algorithm to compute $c(G, S)$ for a time-ordered species tree S and a gene tree G, where $c(G, S)$ is defined as the minimum cost over all possible $\mathbb{DTL}$ reconciliations between G and S.

2.3 Choosing a reliable gene tree among several competing alternatives

Even though our aim is to reconstruct reliable gene trees from a multiple sequence alignment and a species phylogeny, our approach does not directly take sequence alignments as an input, but requires a sample of gene trees, typically produced from the alignment by either a Markov Chain Monte Carlo (MCMC) methods such as PhyloBayes (Lartillot *et al.*, 2009) and MrBayes (Ronquist *et al.*, 2012), or bootstrap resampling.

To find the optimal gene tree, clades found in the input sample of gene trees are combined using the *amalgamation* approach in order to recover an optimal tree with respect to our scoring scheme. The optimal tree recovered will only contain clades found in the input sample of gene trees, but it will not in general be found in the sample itself.

Figure 1 provides a schematic illustration of the amalgamation approach. Clades present in the sample of trees (the unrooted trees on the left) can be combined to obtain a tree such that each clade is found in the sample, but the tree itself is not. For example, one can produce a green-blue tree consisting of a green subtree with genes a, b and c, and a blue subtree with genes d, e and f. The sequence score of each tree is obtained using CCPs that depend on the number of times different trees are seen in the sample and is described in detail in the next section. The reconciliation score for each tree corresponds to the MPR of the gene tree with the species tree. The amalgamation algorithm itself is a joint dynamic programming recursion

over (i) all trees that can be produced from clades present in the input sample and (ii) all possible ways to reconcile each of these trees with the species tree, to recover a gene tree with the smallest joint sequence-reconciliation score. As shown in Figure 1 in the online appendix, amalgamation permits us to explore a vastly larger set of trees than those contained in the sample.

Conceptually, both ALE (Szöllősi *et al.*, 2013b) and TERA are based on the amalgamation approach of AnGST (David and Alm, 2010), and all three methods are—at the level of the dynamic programming recursion—closely related. TERA differs from AnGST in the underlying reconciliation model (Doyon *et al.*, 2010) and because it allows transfers going/coming from extinct or unsampled species (Szöllősi *et al.*, 2013a). Moreover, the AnGST scoring scheme is solely based on the reconciliation score. ALE differs from TERA in that it relies on a complex underlying probabilistic model; the results of which, in contrast to TERA, are contingent on time-like branch lengths of the species tree.

TERA's amalgamation algorithm can be regarded as a generalization of the gene tree reconciliation algorithm of Doyon *et al.* (2010), which iterates over reconciliations by mapping each node of a gene tree to branches of the species tree. In the joint recursion presented in this article, instead of nodes of a gene tree, the clades found in the input sample of gene trees are mapped into branches of the species tree.

More formally, assume we are given a set of (unrooted) gene trees $\mathcal{G}$ on the same leaf set reconstructed from a unique sequence alignment. We denote respectively by $\mathcal{A}(\mathcal{G})$ and $\Pi(\mathcal{G})$ the union of all the clades, and the union of all tripartitions in $\mathcal{G}$. For each tripartition π, we denote by $\pi[1]$ ($\pi[2]$ and $\pi[3]$, respectively) the first (second and third, respectively) element of π. If $\mathcal{G}$ contains unrooted trees we consider all possible rootings for each tree when computing $\mathcal{A}(\mathcal{G})$ and $\Pi(\mathcal{G})$. Furthermore, for a given clade C of $\mathcal{A}(\mathcal{G})$, we denote by $\Pi(C)$ the set of tripartitions $\pi \in \Pi(\mathcal{G})$ for which $\pi[1] = C$. When focusing only on the reconciliation score, the optimization problem consists of computing $c(\mathcal{G}, S) := \min_{G \in \mathcal{G}_{am}} c(G, S)$, where $\mathcal{G}_{am}$ is the set of gene trees such that $\mathcal{A}(G) \subseteq \mathcal{A}(\mathcal{G})$ for all $G \in \mathcal{G}_{am}$. The pseudocode is given in Algorithm 1 in the appendix. Roughly speaking, our algorithm starts by computing the subdivision S' of S, and the sets $\mathcal{A}(\mathcal{G})$ and $\Pi(\mathcal{G})$. Then, it performs a joint traversal of all gene tree clades and species tree branches wherein clades C in $\mathcal{A}(\mathcal{G})$ are considered in order of increasing size, and nodes x' of S in order of increasing height. For each pair (C, x') the algorithm computes the cost of reconciling clade C with x' by testing all possible tripartitions π in $\Pi(C)$. Because each non-trivial tripartition π can be seen as an internal node of an amalgamated tree, with children $\pi[2]$ and $\pi[3]$, the cost of reconciling a tripartition π with x' can be computed according to Algorithm 1 of Doyon *et al.* (2010). We refer the reader to Algorithm 1 of Doyon *et al.* (2010) for a better understanding of the pseudocode. The correctness of our approach is proven in the appendix.

Note that—for ease of writing—the pseudocode of the algorithm does not contain the transfers from the dead, i.e. the transfers going/coming from extinct or unsampled species (Szöllősi *et al.*, 2013a). However, Algorithm 1 can be easily modified to accommodate this kind of event by adding to the species tree S a sister group of the root clade such that, within this group, duplications and losses are free, speciations are not permitted, and transfers to this new group (formally corresponding to unrepresented speciations) cost zero—similar to what is done in the likelihood framework by Szöllősi *et al.* (2013a).

2.4 Taking into account the CCP

As described in the introduction, our goal is to create a species tree aware method for reconstructing gene phylogenies that uses information from both gene sequences and from the reconciliation with a species tree. That is, we wish to construct a method that optimizes a joint sequence-reconciliation score. In order to do this, we must find an efficient manner to incorporate a sequence based cost in addition to the reconciliation cost of Doyon *et al.* (2010) in the amalgamation scheme.

AnGST, the seminal algorithm of David and Alm (2010) that introduced the idea of amalgamation, does not distinguish between trees that can be amalgamated. The problem with this approach is that, as the number of input trees—and thus the amount of information given as input—increases, the set of possible trees that can be amalgamated also increases—until all possible tree topologies can be amalgamated. At this point, since all possible tree topologies can be amalgamated, the most parsimonious reconciled gene tree will only depend on the reconciliation score. In practice this introduces the problem that the topology of the amalgamated gene tree may vary significantly when adding only a few trees to the sample of trees (in the worst case only one tree).

In a probabilistic framework, conditional clade probabilities (CCPs, cf. Fig. 1) provide an accurate approximation of posterior probabilities for a very large number of tree topologies from a smaller MCMC sample (Höhna and Drummond, 2012; Larget, 2013; Szöllősi *et al.*, 2013b). The *CCP* of a rooted tree $G \in \mathcal{G}_{\text{am}}$ (Höhna and Drummond, 2012), denoted by $P_{\text{CCP}}(G)$, is defined as the product of the conditional probabilities of all partitions in $\Pi(G)$. The *conditional probability* of the partition of clade C according to the tripartition π is denoted $P_{\text{CCP}}(\pi)$ and is approximated by the ratio $f_{\mathcal{G}}(\pi)/f_{\mathcal{G}}(\pi[1])$, where for each clade $C \in \mathcal{A}(\mathcal{G})$ and for each tripartition $\pi \in \Pi(\mathcal{G})$, $f_{\mathcal{G}}(\pi)$ and $f_{\mathcal{G}}(C)$ is the frequency of C and π in $\mathcal{G}$.

Here, in order to construct a parsimony method that optimizes a joint sequence-reconciliation score, we choose to minimize the joint cost

$$c_{\text{joint}}(G, S) = c(G, S) + c_{\text{A}} N_{\text{A}} \tag{1}$$

over $G \in \mathcal{G}_{am}$, where the parameter c_A weights the contribution of the sequence alignment N_{A} to the cost, defined as

$$N_{\text{A}} = -\log\left(\frac{P_{\text{CCP}}(G)}{P_{\text{CCP}}(G_{\text{MAP}})}\right) \tag{2}$$

where $P_{\text{CCP}}(G_{\text{MAP}})$ corresponds to the posterior probability of the gene tree with the highest posterior probability according to the sequence alignment. The logarithm of the CCP provides an additive cost for deviation from the phylogeny preferred by the sequence alignment alone, similar to the additive cost for deviation from species phylogeny provided by the DTL event costs. The parameter c_A is analogous to a statistical support threshold, corresponding to a cost c_A for each point of log posterior probability difference between the log posterior probability of a given phylogeny and the gene tree with highest posterior probability.

As illustrated in Figure 1, Algorithm 1 in online appendix is easily modified by adding $c_A \cdot \log(P_{\text{CCP}}(\pi|C))$ to $c_{\mathbb{S}}$, $c_{\mathbb{D}}$ $c_{\mathbb{T}}$ while filling the dynamic programming matrix (on line 15, 17 and 18 of Algorithm 1, respectively). The term $+c_A \cdot \log(P_{\text{CCP}}(G_{\text{MAP}}))$, corresponding to the gene tree with the highest posterior probability, can be neglected during cost minimization as it simply corresponds to an additive constant.

Given estimates for the DTL costs (available for example in David and Alm, 2010; Nguyen *et al.*, 2013), the parameter c_A can be estimated in a self-consistent manner.

However, finding the proper weight between the disagreement with the species tree (increase in DTL events) and the disagreement

with the sequence alignment (decrease in log $(P_{CCP}(G))$) is difficult. Our estimation approach consists of looking for the set of costs that are the most self-consistent, i.e. the ratio of costs that best corresponds to the ratio of events.

We assume a simple model for how costs determine the number of events: each type of event, i.e. DTL events as well as the disagreement with the alignment counted by N_A, are considered to occur independently, such that events with smaller costs are expected to occur more frequently. In particular, the expected amount of disagreement with the species tree due to, respectively duplications, transfers and losses, is proportional to $\exp(-\delta)$, $\exp(-\tau)$ and $\exp(-\lambda)$, while the expected amount of disagreement with the sequence alignment is propotional to $\exp(-c_A)$. The observed amount of disagreement with the species tree is given by the sum of the number of DTL events, i.e. N_D, N_T and N_L, while the observed amount of disagreement with the sequence alignment is given by N_A. We then employ an expectation maximization like recursion equating at each step the observed frequencies with the expected frequencies:

$$c_A^{new} = -\log\left(\frac{N_A}{N_D + N_T + N_L + N_A}\right). \quad (3)$$

Algorithm 1 is then run until $|c_A^{new} - c_A|$ is larger then a threshold ϵ.

2.5 Implementation and validation

TERA is implemented in C++ and is freely available from http://mbb.univ-montp2.fr/MBB/download_sources/16__TERA.

Posterior samples of gene trees, for both simulated and real alignments, were downloaded together with the 'true' gene trees used to simulate alignments from the dryad data repository (doi:10.5061/dryad.pv6df) provided by Szöllősi *et al.* (2013b).

For all the parsimony-based species tree aware methods, including TERA, we used the DTL costs $\delta = 2$, $\tau = 3$, $\lambda = 1$ obtained by David and Alm (2010) using a criteria based on minimizing the change in ancestral genome sizes on a large biological dataset. We ran TreeFix-DTL with default parameters, JTT/GTR with a gamma distribution as models of evolution, and as a starting tree the PhyML tree. MowgliNNI was run with default parameters, a threshold of 50 for weak edges, and with the PhyML tree—with bootstrap values—as a starting gene tree. AnGST was run with default parameters using the dated species tree on samples of 1000 gene trees, whereas JPrIME-DLTRS was run with JTT with a gamma distribution as model of evolution, 100 000 iterations, a thinning factor of 10 and a time out of 10 h. Finally, we ran TERA with a starting c_A of 0.1 and for $\mathcal{G}$ samples from 10 up to 10 000 gene trees for each simulated alignment. The gene trees reconstructed by ALE were downloaded from the above mentioned data repository.

Note that, from a practical perspective, the DTL costs we use are the default parameters for all the parsimonious methods described in the article, and seem to work well for several parts of the Tree of Life. If the user suspects that these values are not suited for the analyses, these parameters should be estimated beforehand, e.g. using the ALE method.

3 Results

To test the accuracy of gene trees reconstructed using TERA we chose a dataset based on 1099 homologous gene families present in 36 cyanobacterial genomes. This dataset, published in Szöllősi *et al.* (2013b), was constructed using homologous families from the HOGENOM database (Penel *et al.*, 2009) and contains both real and simulated alignments as well as the gene trees used to simulate sequences. The mean number of genes per family in this dataset is 36.66, the largest family has 114 genes and the smallest 21 genes; the mean number of species in which a family is found is 31.49, with a minimum of 4 and a maximum of 36; the mean copy number per genome—counting as zero genomes in which a family is absent—is 1.012, with a minimum of 0.5833 and a maximum of 3.17.

We chose this dataset, because (i) it contains a diverse set of gene families from a reasonably large and divergent set of species, and (ii) the parametric bootstrap-like simulation procedure used attempts to retain as much of the complexity of the underlying biological dataset as possible (Szöllősi *et al.*, 2013b). Furthermore, to emulate the relative complexity of real data compared with available models of sequence evolution, we used a complex model of sequence evolution to simulate sequences—an LG model (Le and Gascuel, 2008) with across-site rate variation and invariant sites—and used PhyloBayes (Lartillot *et al.*, 2009) with a simpler model—a Poisson model (Felsenstein, 1981) with no rate variation—to produce the sample of gene trees used by both TERA and AnGST (see below for more details).

3.1 Validation on simulated data

For the simulated alignments, both the 'true' gene tree used to generate the sequences and the species tree—along which the gene trees evolved—are known. Consequently, it is possible to directly assess the accuracy of different reconstruction methods in recovering the correct gene tree.

As shown in Figure 1a in the online appendix, the number of possible amalgamations increases roughly exponentially with increasing sample size in the simulated dataset, but the median reconstruction accuracy achieved by TERA begins to saturate (Figure 1b in the online appendix). To compare the accuracy of our method to that of others, we reconstructed gene trees using six different 'species tree aware' methods: (i) the TERA algorithm described here, (ii) ALE (Szöllősi *et al.*, 2013b), (iii) TreeFix-DTL (Bansal *et al.*, 2014, submitted for publication, http://compbio.mit.edu/treefix-dtl/), (iv) MowgliNNI (Nguyen *et al.*, 2013), (v) AnGST (David and Alm, 2010) and (vi) JPrIME-DLTRS (Sjöstrand *et al.*, 2014) as well as the species tree unaware method, PhyML (Guindon *et al.*, 2010).

In Figure 2a, we plot the normalized Robinson-Foulds (defined as the Robinson-Foulds distance divided by its maximum possible value, and denoted as n-R-F in the following) distance of the reconstructed gene trees to the true tree. These results show that all of the species tree aware methods achieve better accuracy than the species tree unaware method PhyML, which is to be expected as they are given additional information in the form of the species tree. Among the species tree aware methods, with an input of 10 000 samples TERA's accuracy is statistically indistinguishable from the more complex maximum likelihood based results from ALE (paired Wilcoxon test $P > 0.1$) and is significantly more accurate than TreeFix-DTL (Bansal *et al.*, 2014, submitted for publication) (paired Wilcoxon test $P < 10^{-8}$) as well as the other species tree aware method MowgliNNI (Nguyen *et al.*, 2013). TERA also outperformed jPrIME-DLTRS, although the accuracy of the latter may have been limited by the available run time (recall that a time out of 10 h per each data set was given). For an input of 1000 samples, TERA is less accurate than either TERA or ALE with 10 000 input samples (paired Wilcoxon tests $P < 10^{-8}$), statistically indistinguishable from jPrIME-DLTRS, slightly more accurate then TreeFix-DTL (paired Wilcoxon test $P = 0.026$), and still significantly more accurate than MowgliNNI and AnGST (paired Wilcoxon tests $P < 10^{-8}$).

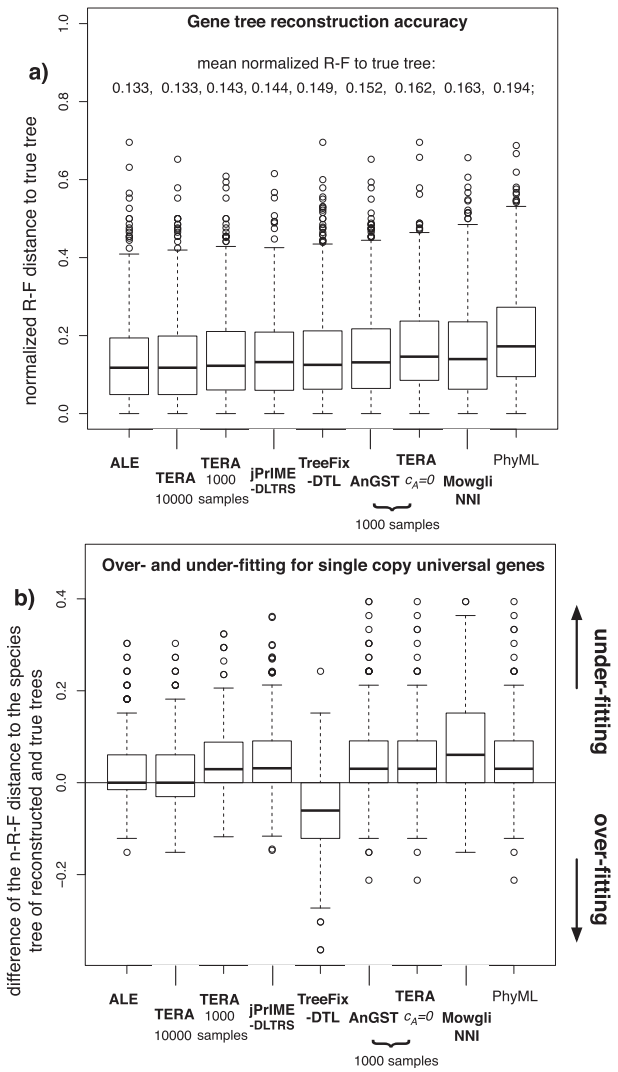

Fig. 2. **(a)** To compare the accuracy of TERA and other methods we used the simulated data set of Szöllősi *et al.* (2013b). We find that TERA achieves statistically equivalent accuracy to ALE and better accuracy than the other methods, see main text for details. **(b)** To test for over and underfitting of the species tree we examined the 431 gene families with exactly one copy in each of the 36 cyanobacterial species. For each family we plot the difference of the R-F distance of the true tree to the species tree and the R-F distance of the reconstructed gene tree from the species tree. Negative values for the difference indicate overfitting, while in the case of underfitting we expect a positive value

Results for AnGST are only shown for sample sizes of 1000 gene trees, due to the very large memory requirement of the AnGST implementation. To investigate the effect of using a joint sequence-reconciliation score we also ran TERA with $c_A = 0$, i.e. emulating AnGST in only optimizing the reconciliation score. We found that on a sample of 1000 trees AnGST was more accurate then TERA with $c_A = 0$ with an n-R-F of, 0.156 and 0.166, respectively. However, using TERA with only 1000 samples, but estimating c_A, resulted in a mean n-R-F of 0.146. The average c_A estimated by TERA was 0.49 while the average N_A was 6.57.

An important difference of TreeFix-DTL compared with the other methods considered here, is that it does not use information of the time order of speciation events in the species tree (note that AnGST can also run without information on time ordering). Therefore TreeFix-DTL uses less information, which may explain the difference in performance in comparison to TERA on the simulated dataset. Nonetheless, we ran TERA with 10 random time orderings of the species tree and this resulted in statistically identical n-R-F values when using the correct time order of speciations (Wilcoxon rank sum test $P = 0.6$).

Table 1. Mean runtimes in seconds for the methods discussed in the main text on a cluster of 2.1 GHz Intel Xeon processors with 24 GB of RAM with maximum runtime limited to 10 h per family

	Stand-alone [s]	Input [s]	
		PhyloBayes	
		1000 samples	10 000 samples
TERA	3.65	756.6	7566
AnGST	54.9	756.6	—
ALE	159.2	756.6	7566
		PhyML	
MowgliNNI	6.3	182.5	
TreeFix	5718.0	182.5	
		No input tree needed	
jPrIME	32 137.3	0	

The time required to compute inputs is given by the runtime of PhyloBayes for 1000 and 10 000 samples and for the time required for PhyML to compute an ML tree with SH branch supports. Stand-alone runtimes are given for 10 000 samples for TERA and ALE and 1000 samples for AnGST.

A potential concern regarding methods that optimize a joint reconciliation-sequence score is that we may overfit the species tree. If overfitting of the species tree occurs we expect the reconstructed gene trees to become too similar to the species tree. In the context of the simulated dataset used here, we expect that the reconstructed gene trees will become more similar to the species tree than the true trees used to simulate alignments. To test for such a signal of overfitting, we require a measure of similarity between gene trees and the species tree. The most straightforward solution is to restrict our analysis to gene families that have exactly one copy in each species. In this case, we can simply use the n-R-F distance between the species tree and each of the gene trees as our similarity measure. In Figure 2b, we show the results for the 431 single copy universal gene families in our simulated dataset. We measure the extent of over and underfitting as the difference in n-R-F distance between the species tree and the reconstructed gene tree and the n-R-F distance between the species tree and the true gene tree. We observe that the species tree unaware method, PhyML, as expected, reconstructs trees that are more distant to the species tree than the true tree. The results for the species tree aware methods are more variable: ALE, based on an explicit probabilistic approach, exhibits a median difference of zero and produces only a few examples of overfitting. TERA, which estimates the c_A parameter giving the relative weight of the sequence and reconciliation component of the joint score, also achieves a median difference of zero when 10 000 samples are given as input, but produces a somewhat larger number of slightly overfitted trees. When only 1000 samples are used, both TERA and AnGST underfit the species tree, similar to jPRIME-DLTRS, suggesting that it may be a lack of convergence of the sampling in all cases. TreeFix-DTL, which relies on a fixed support threshold, shows signs of more significant overfitting; while MowgliNNI substantially underfits the species tree, at least with the default parameters used here.

The runtimes for the methods discussed in this section are given in Table 1. We can see that TERA has the fastest stand-alone runtime. However, if the runtime necessary to generate the input tree(s)

are considered, Mowgli is the fastest method, but it is also the least accurate (cf. Fig. 2a). For an input size of 1000 samples TERA achieves comparable accuracy to TreeFix-DTL, but with a seven fold reduced mean runtime. For an input size of 10 000 samples TERA achieves similar accuracy to ALE and outperforms TreeFix-DTL, but considering the time to generate the required inputs TreeFix-DTL is 1.3 times faster on average.

3.2 Results on real data

In order to test TERA on biological data, we again used the dataset published in Szöllősi *et al.* (2013b), but focusing on real alignments. As inputs to TERA we used: (i) the tree samples obtained from real alignments and (ii) the ML species tree unaware gene trees obtained using PhyML from the same alignments. Similar to the results of ALE (Szöllősi *et al.*, 2013b), we find that the number of transfer and loss events (but not duplication events) in most parsimonious reconciled gene trees is substantially lower than those found in the most parsimonious reconciliations of PhyML trees: the mean and median number of transfers per family was 3.914 and 3 compared with 10.38 and 9, respectively; the mean and median number of losses per family was 5.088 and 4 compared with 7.542 and 7, respectively, while the mean and median number of duplications per family was 1.071 and 0 compared with 1.042 and 0, respectively.

4 Discussion

We have presented a detailed description of the TERA algorithm, a parsimony-based species tree aware method of gene tree reconstruction. We demonstrate that TERA reconstructs gene trees with nearly identical accuracy as the more complex ML based ALE method and, at least on the simulated datasets considered here, outperforms the other parsimony based species tree aware methods.

Examining a subset of single copy universal gene families we show that TERA does not overfit or underfit the species tree. This result lends credibility to TERA's results on biological data, whereby two thirds of apparent gene transfers in gene trees reconstructed without taking into consideration the species tree are not recovered given knowledge of the species phylogeny.

Although parsimony based methods are fundamentally limited in some aspects compared with model based probabilistic methods, in the case of species tree aware gene tree reconstruction our results indicate that parsimony based methods can closely approach their accuracy. A further advantage of TERA compared with the corresponding probabilistic method ALE is that it is faster (if only up to a constant factor), does not require explicit time-like branch lengths that are difficult to estimate, and due to its relative simplicity, in particular the lack of numerical integration, is more robust in practice. Compared with parsimony based methods that require prior assumptions about statistical support, TERA is distinguished by its ability to estimate a statistical support threshold from its input. In contrast to the methods considered here it does require more elaborate upstream analysis, taking as its input a sample of trees from e.g. an MCMC-based tree inference methods, while in contrast MowgliNNI requires a single tree with branch supports, and TreeFix-DTL a multiple sequence alignment.

Finally, while we have shown that it is possible to estimate the c_A parameter, we have been less successful in estimating all four costs $(\delta, \tau, \lambda, c_A)$ simultaneously, due to the tendency of the cost estimates to diverge toward a very low transfer cost, and a correspondingly large number of transfers. We expect that relaxing the, in general, unrealistic assumption of independence between events could ameliorate this problem.

Acknowledgements

The authors like to thank Eric Tannier, Bastien Boussau and Vincent Daubin for the constructive discussions.

Funding

C.S. and E.J. were supported by the French *Agence Nationale de la Recherche Investissements d'avenir/ioinformatique* (ANR-10-BINF-01-02, *Ancestrome*). G.J.Sz. was supported by the Marie Curie CIG 618438 'Genestory' and the Albert Szent-Györgyi Call-Home Researcher Scholarship A1-SZGYA-FOK-13-0005 supported by the European Union and the State of Hungary, co-financed by the European Social Fund in the framework of TÁMOP 4.2.4. A/1-11-1-2012-0001 'National Excellence Program'. This publication is contribution no. 2014-178 of the Institut des Sciences de l'Evolution de Montpellier [ISEM, UMR 5554].

Conflict of Interest: none declared.

References

Akerborg,O. *et al.* (2009). Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc. Nat. Acad. Sci.* USA, **106**, 5714–5719.

Arvestad,L. (2003). Bayesian gene/species tree reconciliation and orthology analysis using MCMC. *Bioinformatics*, **19**, 7i–15i.

Bansal,M.S. *et al.* (2012). Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**, i283–i291.

Boussau,B. *et al.* (2013). Genome-scale coestimation of species and gene trees. *Genome Res.* **23**, 323–330.

Conow,C. *et al.* (2010). Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms Mol. Biol.* **5**, 16.

David,L.A. and Alm,E.J. (2010). Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, **469**, 93–96.

Doyon,J. *et al.* (2011). Models, algorithms and programs for phylogeny reconciliation. *Brief Bioinform.* **12**, 392–400.

Doyon,J.-P. *et al.* (2010). An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: *Proceedings of the 2010 International Conference on Comparative Genomics, RECOMB-CG'10*, Springer-Verlag, Berlin, Heidelberg, pp. 93–108.

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

Guindon,S. *et al.* (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321.

Hallett,M.T. and Lagergren,J. (2001). Efficient algorithms for lateral gene transfer problems. In: *Proceedings of the Fifth Annual International Conference on Computational Biology,* ACM, New York, NY, pp. 149–156.

Höhna,S. and Drummond,A. (2012). Guided tree topology proposals for Bayesian phylogenetic inference. *Syst. Biol.* **61**, 1–11.

Larget, B. (2013). The estimation of tree posterior probabilities using conditional clade probability distributions. *Syst. Biol.* **62**, 501–511.

Lartillot, N. *et al.* (2009). PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics*, **25**, 2286.

Le,S.Q. and Gascuel,O. (2008). An improved general amino acid replacement matrix. *Mol. Biol. Evol.* **25**, 1307–1320.

Maddison,W.P. (1997). Gene trees in species trees. *Syst. Biol.* **46**, 523–536.

Nguyen, T. *et al.* (2012). Accounting for gene tree uncertainties improves gene trees and reconciliation inference. In: B.,Raphael and J Tang (eds.) *Algorithms in Bioinformatics, volume 7534 of Lecture Notes in Computer Science*. Springer, Berlin Heidelberg, pp. 123–134.

Nguyen,T.H. *et al*. (2013). Reconciliation and local gene tree rearrangement can be of mutual profit. *Algorithms Mol Biol,* **8**, 12.

Penel,S. *et al*. (2009). Databases of homologous gene families for comparative genomics. *BMC Bioinformatics,* **6**, S3.

Rannala,B. and Yang,Z. (2003). Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics,* **164**, 1645–1656.

Rasmussen,M.D. and Kellis,M. (2007). Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.* **17**, 1932–1942.

Rasmussen,M.D. and Kellis,M. (2010). A Bayesian Approach for Fast and Accurate Gene Tree Reconstruction. *Mol. Biol. Evol.* **28**, 273–290.

Rasmussen,M.D. and Kellis,M. (2012). Unified modeling of gene duplication, loss, and coalescence using a locus tree. *Genome Res.* **22**, 755–765.

Ronquist,F. *et al*. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.* **61**, 539–542.

Sjöstrand,J. *et al*. (2014). A bayesian method for analyzing lateral gene transfer. *Syst. Biol.* **63**, 409–420.

Suchard,M.A. (2005). Stochastic models for horizontal gene transfer: taking a random walk through tree space. *Genetics,* **170**, 419–431.

Szöllősi,G.J. and Daubin,V. (2012). Modeling gene family evolution and reconciling phylogenetic discord. *Methods Mol. Biol.* **856**, 29–51.

Szöllősi,G.J. *et al*. (2012). Phylogenetic modeling of lateral gene transfer reconstructs the pattern and relative timing of speciations. *Proc. Nat. Acad. Sci. USA,* **109**, 17513–17518.

Szöllősi,G.J. *et al*. (2013a). lateral gene transfer from the dead. *Syst. Biol.,* **62**, 386–397.

Szöllősi,G.J. *et al*. (2013b). Efficient exploration of the space of reconciled gene trees. *Syst. Biol.* **62**, 901–912.

Szöllősi,G. *et al*. (2014). The inference of gene trees with species trees. *Syst. Biol.* **64**, e42–e62.

Than,C. and Nakhleh,L. (2009). Species tree inference by minimizing deep coalescences. *PLoS Comput. Biol.* **5**, e1000501.

Tofigh,A. (2009). Using trees to capture reticulate evolution, lateral gene transfers and cancer progression. PhD Thesis, KTH Royal Institute of Technology, Sweden.

Tofigh,A. *et al*. (2011). Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **8**, 517–535.

Wu,Y.-C. *et al*. (2013). TreeFix: statistically informed gene tree error correction using species trees. *Syst. Biol.* **62**, 110–120.

Wu,Y.-C. *et al*. (2014). Most parsimonious reconciliation in the presence of gene duplication, loss, and deep coalescence using labeled coalescent trees. *Genome Res.* **24**, 475–486.