OXFORD

## Phylogenetics

# ecceTERA: comprehensive gene tree-species tree reconciliation using parsimony

Edwin Jacox[1], Cedric Chauve[2], Gergely J. Szöllősi[3], Yann Ponty[4] and Celine Scornavacca[1,5,*]

[1]ISE-M, Université Montpellier, CNRS, IRD, EPHE, Montpellier, France, [2]Department of Mathematics, Simon Fraser University, Burnaby, BC, Canada, [3]ELTE-MTA "Lendület" Biophysics Research Group, Budapest, Hungary, [4]LIX, CNRS/Inria AMIB, Ecole Polytechnique, Palaiseau, France and [5]Institut de Biologie Computationnelle (IBC), Montpellier, France

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso.

## Abstract

**Summary**: A gene tree-species tree reconciliation explains the evolution of a gene tree within the species tree given a model of gene-family evolution. We describe ecceTERA, a program that implements a generic parsimony reconciliation algorithm, which accounts for gene duplication, loss and transfer (DTL) as well as speciation, involving sampled and unsampled lineages, within undated, fully dated or partially dated species trees. The ecceTERA reconciliation model and algorithm generalize or improve upon most published DTL parsimony algorithms for binary species trees and binary gene trees. Moreover, ecceTERA can estimate accurate species-tree aware gene trees using amalgamation.

**Availability and implementation**: `ecceTERA` is freely available under http://mbb.univ-montp2.fr/MBB/download_sources/16__ecceTERA and can be run online at http://mbb.univ-montp2.fr/MBB/subsection/softExec.php?soft=eccetera.

**Contact**: celine.scornavacca@umontpellier.fr

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

## 1 Introduction

Reconciling gene trees with a species tree (Goodman *et al.*, 1979; Page, 1994) is a crucial step in many phylogenomics problems (Rusin *et al.*, 2014), from the reconstruction of gene trees (Bansal *et al.*, 2015; David and Alm, 2010; Nguyen *et al.*, 2012; Scornavacca *et al.*, 2015; Szöllősi and Daubin, 2012) to the reconstruction of ancestral genomes (Chauve et al., 2013; Szöllősi *et al.*, 2015). Internal gene tree nodes, representing ancestral genes, are assigned to species (extant or extinct) and evolutionary events such as gene duplications or horizontal transfers. This results in evolutionary histories (or *reconciliations*) for gene families explaining apparent discordances with the speciation history.

There exist several reconciliation software packages—such as Notung (Durand *et al.*, 2006; Stolzer *et al.*, 2012), RANGER-DTL

(Bansal *et al.*, 2012), Mowgli (Doyon *et al.*, 2010), Eucalypt (Donati *et al.*, 2015), AngST (David and Alm, 2010) and Jane (Conow *et al.*, 2010)—whose algorithms and models, differ, sometimes in subtle ways, in terms of the evolutionary events they consider (Doyon *et al.*, 2011). Furthermore, an important difference lies in the nature of the species tree. First, some software packages assume that the considered species tree is fully dated (i.e. speciation dates are provided) and ensure that horizontal transfers are time-consistent (i.e. that no transfer can occur between species that did not exist at the same time, see Doyon *et al.*, 2010), while others ensure only local time-consistency (Bansal *et al.*, 2012). But other software packages do consider undated species trees (i.e. the timing of speciation events is not known), and, as ensuring time-consistency with undated trees is an NP-hard problem (Hallett *et al.*, 2004;

Ovadia *et al.*, 2011; Tofigh *et al.*, 2011), these programs sometimes fail to find a consistent, parsimonious scenario and either output an inconsistent one or produce no output. This disparate body of parsimony models and algorithms makes it difficult to assess if discrepancies between the results provided by different software is due to differences in the model, the algorithm or the implementation.

## 2 Methods

We introduce ecceTERA, a program whose aim is to compute parsimonious reconciliations between species trees and gene trees under a comprehensive evolutionary model. The program ecceTERA improves upon existing parsimony reconciliation software in several aspects. In Table 1 we show how ecceTERA compares to the most commonly used parsimony algorithms/software that compute reconciliations using the Duplication-Transfer-Loss (DTL) model.

Most notably, its evolutionary model is comprehensive as it includes the following evolutionary events: speciation, speciation-loss (speciation followed by a loss of one gene copy), gene duplication, gene loss, gene transfer and transfer-loss (gene transfer with loss of the original gene) between two sampled species, and gene transfer and transfer-loss from/to an unsampled species (i.e. a species that is not represented in the dataset) to/from a sampled one. To the best of our knowledge, ecceTERA is the first parsimony software to implement this last event, called 'transfer from/to the dead' in Szöllősi *et al.* (2013b), where it was shown on a dataset of 36 cyanobacteria species that a significant number of gene transfers that occurred in the past could reasonably be transfers from/to unsampled species. A more detailed description of the evolutionary model of ecceTERA is provided in the Supplementary Material.

Also, ecceTERA, can take as input a species tree that is either undated, fully dated (speciations are totally ordered), or partially dated (the dating of a selected set of speciations is known), and computes parsimonious reconciliations that are consistent with the provided time information, if any. Thus, for fully dated species trees, ecceTERA provides consistent parsimonious reconciliations, while with partially dated or undated species tree, ecceTERA checks the consistency of the computed parsimonious reconciliations and indicates the inconsistent ones.

Finally, the ecceTERA software is based on a unified dynamic programming algorithm (described in the Supplementary Material) that builds upon the model of the Mowgli algorithm (Doyon *et al.*, 2010). The time and space complexity of the ecceTERA reconciliation algorithm is $O((k+1)|S||G|)$ to construct *one* parsimonious reconciliation, where $S$ is the species tree, $G$ the gene tree and $k$ the number of dated nodes of $S$. The use of a single algorithm ensures that, when comparing the results obtained on the same data with different models (for example to evaluate the impact of a speciation time), the observed differences can be attributed to differences in the models and not to model-specific algorithmic or implementation issues.

ecceTERA also includes several features of interest for the analysis of reconciled gene trees. For example, ecceTERA can compute a compact graph structure that represents the set of all parsimonious reconciliations for the chosen evolutionary model, as described in Scornavacca *et al.* (2013). It is also possible to associate support values with reconciliation events obtained by considering nearly-optimal or Pareto-optimal reconciliations, as described in Nguyen *et al.* (2013) and To *et al.* (2015).

Finally, an important feature of ecceTERA is its ability to reconstruct species-aware gene trees using the joint amalgamation method described in Scornavacca *et al.* (2015, the *TERA* algorithm)—note that AngST (David and Alm, 2010) also amalgamates gene trees, though using a different algorithm, and that MowgliNNI (Nguyen *et al.*, 2012), Notung (Stolzer *et al.*, 2012) and TreeFix-DTL (Bansal *et al.*, 2015) can modify gene trees using local rearrangements improving the reconciliation score.

ecceTERA is a command line software written in C++, using the Bio++ (Guéguen *et al.*, 2013) and BOOST C++ libraries, and is available as source code. It has been tested on the Mac and Linux platforms, for which executables are provided. It requires input trees in Newick format and outputs reconciliations in a custom format, described in the manual provided with the software. ecceTERA can also output reconciliations in a format readable by SylvX (Chevenet

**Table 1.** Comparison of selected features of the most commonly used parsimony reconciliation programs in the DTL model; all considered programs compute a parsimonious DTL reconciliation

| Program | A Handles TL events | B Handles T from the dead | C (Un) rooted gene trees? | D Performs Amalga- mation | E Rearranges gene trees | F Computes event supports | G Handles ILS | H (Un)dated species trees? | I Only feasible solutions (dated) | J Only feasible solutions (undated) | K All (feasible) co-optimal solutions | L GUI | M Source code |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| RANGER-DTL 1.0 | ○ | ○ | R,U | ○ | ○ | ○ | ○ | F,U | ○ | ○ | ○ | ○ | ○ C++ |
| Notung 2.8 | ○ | ○ | R,U | ○ | • | ○ | • | U | ○ | •/○ | • | • | ○ Java |
| Mowgli | • | •/○ | R | ○ | • | ○ | ○ | F | • | ○ | ○ | ○ | • C++ |
| EUCALYPT | ○ | ○ | R | ○ | ○ | ○ | ○ | U | ○ | •/○ | • | ○ | • Java |
| AngST | ○ | ○ | U | • | ○ | ○ | ○ | F,U | • | ○ | ○ | ○ | • Python |
| Jane 4 | ○ | ○ | R | ○ | ○ | • | ○ | F,U,P | ○ | ○ | ○ | • | • Java |
| ecceTERA | •/○ | •/○ | R,U | • | ○ | • | ○ | F,U,P | • | •/○ | • | ○ | • C++ |

The symbols • and ○ indicate whether or not a method implements a given feature. The symbol •/○ is used for features that can be turned on/off. Column **A.** indicates support for the transfer-loss atomic operation. Col. **B.** indicates whether transfers from/to an unsampled species are considered by the model. Col. **C.** specifies which algorithms support Rooted (R) or Unrooted (U) gene trees. Col. **D.** and **E.** specify which algorithms can perform gene tree reconstruction, either through almagamation of clades (**D.**) or local rearrangements (**E.**). Col. **F.** indicates whether the program provides support values for the events of the computed reconciliations. Col. **G.** indicates which programs account for Incomplete Lineage Sorting (ILS). Col. **H.** specifies whether Fully dated (F), Undated (U), or Partially dated (P) species trees are supported by the program. Columns **I.** and **J.** indicate whether the program only reports optimal solutions that are feasible (i.e. consistent with both the topology of the species tree and the speciation dates if available). For Col. **J.**, a •/○ value means that no feasible solution is provided when all optimal solutions are infeasible. Col. **K.** shows whether the program can enumerate all optimal (feasible) solutions. Col. **L.** indicates the availability of a Graphical User Interface (all programs have a command-line mode). Note that ecceTERA can be run online at http://mbb.univ-montp2.fr/MBB/subsection/softExec.php?soft=eccetera. Col. **M.** indicates the availability of the source code and the associated programming language that is used. Programs without a source code distribution can nevertheless be executed on all major operating systems.

et al., 2016), a software for visualizing and manipulating reconciliations (an example of visualization of an `ecceTERA` reconciliation is provided in Fig. S4 of the Supplementary Material).

## 3 Case study

We applied `ecceTERA` to the reconstruction of species-tree aware gene trees using the joint amalgamation method presented in Scornavacca *et al.* (2015), under several evolutionary models. The tests were performed on a dataset composed of a dated species tree of 36 cyanobacteria and a set of 1099 gene trees obtained from simulated alignments generated as described in Szöllősi *et al.* (2013a, Supplementary Material). We found that using a dated species tree and a full evolutionary model, including transfer-loss and transfer from/to extinct or unsampled species, achieves the highest accuracy, with a mean Robinson–Foulds (RF) distance of 9.42 to the true trees; this corresponds to the model described in Scornavacca *et al.* (2015). Excluding transfer from/to extinct or unsampled species led to a larger mean RF distance of 9.78, while considering the species as undated provided the worst results, with mean RF distances slightly above 10. These experiments are described in more detail in the Supplementary Material.

To assess its computational efficiency, we compared `ecceTERA` to `RANGER-DTL` (Bansal *et al.*, 2012), the most efficient parsimony software available to date, using three large datasets—COG, CYANO and HOGENOM—described in detail in the Supplementary Material. The default parameters were used for both programs. For the COG and CYANO datasets the running times and memory requirements of the two programs were comparable while, on the HOGENOM dataset (see Supplementary Fig. S3), `ecceTERA` was faster, by a mean factor of 3.29×, for 73% of the gene families, and `RANGER-DTL-U` performed better, by a mean factor of 1.28×, on very large gene families (indeed, most of the remaining 27% of the gene families where `RANGER-DTL`-U was faster contained more than 800 genes). We refer the reader to the Supplementary Material for a precise comparison.

## 4 Conclusion

`ecceTERA` supports a comprehensive set of reconciliation models, for which parsimonious reconciliations are computed using an efficient implementation of a unified dynamic programming algorithm. Moreover, it offers additional features that were not available in a single software, such as species-tree aware gene tree reconstruction and the computation of all parsimonious reconciliations. Future extensions will include the reconciliation of rooted non-binary gene trees (manuscript in preparation, partial support for this feature is provided in the current release), the sampling of optimal and suboptimal reconciliation scenarios, and provide a model of incomplete lineage sorting (ILS) (Chan *et al.*, in preparation), in the spirit of Stolzer *et al.* (2012).

## References

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

Bansal,M.S. *et al.* (2015) Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, **31**, 1211–1218.

Chauve,C. *et al.* (2013). Duplication, rearrangement and reconciliation: a follow-up 13 years later. In: Chauve, C. *et al* (eds) *Models and Algorithms for Genome Evolution*. Springer, London, pp. 47–62.

Chevenet,F. *et al.* (2016) Sylvx: a viewer for phylogenetic tree reconciliations. *Bioinformatics*, **32**, 608–610.

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.

Donati,B. *et al.* (2015) EUCALYPT: efficient tree reconciliation enumerator. *Algorithms Mol. Biol.*, **10**, 3.

Doyon,J. *et al.* (2010) An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. In: *Comparative Genomics – International Workshop, RECOMB-CG 2010, Ottawa, Canada, October 9–11, 2010. Proceedings*, volume 6398 of *Lecture Notes in Computer Science*, pp. 93–108. Springer.

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

Durand,D. *et al.* (2006) A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.*, **13**, 320–335.

Goodman,M. *et al.* (1979) Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Syst. Zool.*, **28**, 132–163.

Guéguen,L. *et al.* (2013) Bio++: efficient extensible libraries and tools for computational molecular evolution. *Mol. Biol. Evol.*, **30**, 1745–1750.

Hallett,M. *et al.* (2004) Simultaneous identification of duplications and lateral transfers. In: *Proceedings of the Eight International Conference on Research in Computational Molecular Biology (RECOMB)*. ACM Press, pp. 347–356.

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

Nguyen,T.H. *et al.* (2013) Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. *PloS ONE*, **8**, e73667.

Ovadia,Y. *et al.* (2011) The cophylogeny reconstruction problem is np-complete. *J. Comput. Biol.*, **18**, 59–65.

Page,R.D. (1994) Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Syst. Biol.*, **43**, 58–77.

Rusin,L.Y. *et al.* (2014) Reconciliation of gene and species trees. *BioMed Res. Int.*, **2014**, 642089.

Scornavacca,C. *et al.* (2013) Representing a set of reconciliations in a compact way. *J. Bioinf. Comput. Biol.*, **11**, 1250025.

Scornavacca,C. *et al.* (2015) Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, **31**, 841–848.

Stolzer,M. *et al.* (2012) Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, **28**, i409–i415.

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.* (2013a) Efficient exploration of the space of reconciled gene trees. *Syst. Biol.*, **62**, 901–912.

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

Szöllősi,G. *et al.* (2015) The inference of gene trees with species trees. *Syst. Biol.*, **64**, e42–e62.

To,T.H. *et al.* (2015) A fast method for calculating reliable event supports in tree reconciliations via Pareto optimality. *BMC Bioinf.*, **16**, 384.

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