ecceTERA: Comprehensive gene tree-species tree reconciliation using parsimony

Edwin Jacox, Cedric Chauve, Gergely J Szöllösi, Yann Ponty, Celine Scornavacca

ecceTERA: Comprehensive gene tree-species tree reconciliation using parsimony

Edwin Jacox[1], Cedric Chauve[2], Gergely J. Szöllősi[3], Yann Ponty[4], 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] CNRS/Inria AMIB, Ecole Polytechnique, Palaiseau, France
[5] Institut de Biologie Computationnelle (IBC), Montpellier, France

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.

Supplementary Information: Supplementary data available online.

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

Contact: celine.scornavacca@umontpellier.fr

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 (David and Alm 2010; Szöllősi and Daubin 2012; Nguyen *et al.* 2012; Scornavacca *et al.* 2015; Bansal *et al.* 2015) to the reconstruction of ancestral genomes (Szöllősi *et al.* 2015; Chauve *et al.* 2013). 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 data set) to/from a sampled one. To the best of our knowledge, ecceTERA is the first parsimony

*to whom correspondence should be addressed

Program	**A** Handles $\mathbb{TL}$ events	**B** Handles $\mathbb{T}$ 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,U	○	●	○	○	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++

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. 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.

software to implement this last event, called "transfer from/to the dead" in Szöllősi *et al.* (2013b), where it was shown on a data set 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; 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. 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 *et al.* 2015), a software for visualizing and manipulating reconciliations (an example of visualization of an `ecceTERA` reconciliation is provided in Fig. S.4 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 data set composed of a dated species tree of 36 cyanobacteria and a set of $1,099$ 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 data sets – COG, CYANO and HOGENOM – described in detail in the Supplementary Material. The default parameters were used for both programs. For the COG and CYANO data sets the running times and memory requirements of the two programs were comparable while, on the HOGENOM data set (see Figure S.3), `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).

Acknowledgements. NSERC (Discovery Grant RGPIN-249834 to C.C.), ANR (grant ANR-10-BINF-01-02 to C.S. and E.J.), GJSz was supported by the FP7-PEOPLE-CIG grant "GENESTORY". We thank Jean-Francois Dufayard for providing the species tree for the HOG data set. CS and EJ benefited from the Montpellier Bioinformatics Biodiversity cluster computing platform.

REFERENCES

Bansal, M. S., Alm, E. J., and Kellis, M. 2012. Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, 28(12): i283–91.

Bansal, M. S., Wu, Y., Alm, E. J., and Kellis, M. 2015. Improved gene tree error correction in the presence of horizontal gene transfer. *Bioinformatics*, 31(8): 1211–1218.

Chauve, C., El-Mabrouk, N., Gueguen, L., Semeria, M., and Tannier, E. 2013. Duplication, rearrangement and reconciliation: A follow-up 13 years later. In C. Chauve, N. El-Mabrouk, and E. Tannier, editors, *Models and Algorithms for Genome Evolution*, pages 47–62. Springer.

Chevenet, F., Doyon, J.-P., Scornavacca, C., Jacox, E., Jousselin, E., and Berry, V. 2015. Sylvx: a viewer for phylogenetic tree reconciliations. *Bioinformatics*, page btv625.

Conow, C., Fielder, D., Ovadia, Y., and Libeskind-Hadas, R. 2010. Jane: A new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(1): 16.

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

Donati, B., Baudet, C., Sinaimeri, B., Crescenzi, P., and Sagot, M. 2015. EUCALYPT: efficient tree reconciliation enumerator. *Algorithms for Molecular Biology*, 10: 3.

Doyon, J., Scornavacca, C., Gorbunov, K. Y., Szöllosi, G. J., Ranwez, V., and Berry, V. 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*, pages 93–108. Springer.

Doyon, J., Ranwez, V., Daubin, V., and Berry, V. 2011. Models, algorithms and programs for phylogeny reconciliation. *Briefings in Bioinformatics*, 12(5): 392–400.

Durand, D., Halldórsson, B. V., and Vernot, B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology*, 13(2): 320–335.

Goodman, M., Czelusniak, J., Moore, G. W., Romero-Herrera, A. E., and Matsuda, G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, 28(2): 132–163.

Guéguen, L., Gaillard, S., Boussau, B., Gouy, M., Groussin, M., Rochette, N. C., Bigot, T., Fournier, D., Pouyet, F., Cahais, V., Bernard, A., Scornavacca, C., Nabholz, B., Haudry, A., Dachary, L., Galtier, N., Belkhir, K., and Dutheil, J. Y. 2013. Bio ++ : Efficient Extensible Libraries and Tools for Computational Molecular Evolution. *Molecular biology and evolution*, 30(8): 1745–1750.

Hallett, M., Largergren, J., and Tofigh, A. 2004. Simultaneous identification of duplications and lateral transfers. In *Proceedings of the Eight International Conference on Research in Computational Molecular Biology (RECOMB)*, pages 347–356. ACM Press.

Nguyen, T., Doyon, J.-P., Pointet, S., Arigon Chifolleau, A.-M., Ranwez, V., and Berry, V. 2012. Accounting for gene tree uncertainties improves gene trees and reconciliation inference. In B. Raphael and J. Tang, editors, *Algorithms in Bioinformatics*, volume 7534 of *Lecture Notes in Computer Science*, pages 123–134. Springer Berlin Heidelberg.

Nguyen, T.-H., Ranwez, V., Berry, V., and Scornavacca, C. 2013. Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods. *PloS ONE*, 8(10): e73667.

Ovadia, Y., Fielder, D., Conow, C., and Libeskind-Hadas, R. 2011. The cophylogeny reconstruction problem is np-complete. *Journal of Computational Biology*, 18(1): 59–65.

Page, R. D. 1994. Maps between trees and cladistic analysis of historical associations among genes, organisms, and areas. *Systematic Biology*, 43(1): 58–77.

Rusin, L. Y., Lyubetskaya, E. V., Gorbunov, K. Y., and Lyubetsky, V. A. 2014. Reconciliation of gene and species trees. *BioMed Research International*, 2014: 642089.

Scornavacca, C., Paprotny, W., Berry, V., and Ranwez, V. 2013. Representing a set of reconciliations in a compact way. *Journal of Bioinformatics and Computational Biology*, 11(2): 1250025.

Scornavacca, C., Jacox, E., and Szöllosi, G. J. 2015. Joint amalgamation of most parsimonious reconciled gene trees. *Bioinformatics*, 31(6): 841–848.

Stolzer, M., Lai, H., Xu, M., Sathaye, D., Vernot, B., and Durand, D. 2012. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. *Bioinformatics*, 28(18): i409–i415.

Szöllősi, G. J. and Daubin, V. 2012. Modeling gene family evolution and reconciling phylogenetic discord. *Methods in Molecular Biology*, 856: 29–51.

Szöllősi, G. J., Rosikiewicz, W., Boussau, B., Tannier, E., and Daubin, V. 2013a. Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 62(6): 901–912.

Szöllősi, G. J., Tannier, E., Lartillot, N., and Daubin, V. 2013b. Lateral gene transfer from the dead. *Systematic Biology*, 62(3): 386–397.

Szöllősi, G., Tannier, E., Daubin, V., and Boussau, B. 2015. The inference of gene trees with species trees. *Systematic Biology*, 64(1): e42–62.

To, T.-H., Jacox, E., Ranwez, V., and Sconavacca, C. 2015. A fast method for calculating reliable event supports in tree reconciliations via Pareto optimality. submitted.

Tofigh, A., Hallett, M. T., and Lagergren, J. 2011. Simultaneous identification of duplications and lateral gene transfers. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2): 517–535.