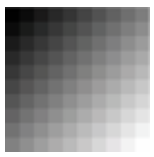

Phylogenetics

Treerecs: an integrated phylogenetic tool, from sequences to reconciliations

Nicolas Comte [1], Benoit Morel [3], Damir Hasic, Laurent Guéguen [2], Bastien Boussau [2,], Vincent Daubin [2], Simon Penel [2], Celine Scornavacca, Manolo Gouy [2], Alexandros Stamatakis [3,4], Eric Tannier [1,2,*] and David P. Parsons [1,*]

[1]Inria Grenoble Rhône-Alpes, Montbonnot 38334, France [2]Université de Lyon 1, Laboratoire de Biométrie et Biologie Évolutive, CNRS UMR5558 F-69622 Villeurbanne, France. [3]Scientific Computing Group, Heidelberg Institute for Theoretical Studies, Heidelberg, Germany [4]Institute of Theoretical Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Gene and species tree reconciliation methods can be used to root gene trees and correct uncertainties that are due to scarcity of signal in multiple sequence alignments. So far, reconciliation tools have not been integrated in standard phylogenetic software and they either lack of performance on certain functions, or usability for biologists.
Results: We present Treerecs, a phylogenetic software based on duplication-loss reconciliation. Treerecs is simple to install and to use, fast, versatile, with a graphic output, and can be used along with methods for phylogenetic inference on multiple alignments like PLL and Seaview.
Availability: Treerecs is open-source. Its source code (C++, AGPLv3) and manuals are available from https://project.inria.fr/treerecs/
Contact: eric.tannier@inria.fr or david.parsons@inria.fr
online.

1 Context

Phylogeny reconciliation methods are recognized to be powerful tools to understand the evolution of gene families (Szöllősi *et al.*, 2015), and host-parasite co-evolution (Bailly-Bechet *et al.*, 2017). They consist in annotating, rooting or improving gene trees by comparing them to species trees. Available tools (Bansal *et al.*, 2018; Stolzer *et al.*, 2012; Jacox *et al.*, 2016; Akerborg *et al.*, 2009; Szöllősi *et al.*, 2013), accomplish diverse reconciliation tasks: mainly annotating trees, rooting trees and improving trees. They take into account duplications, transfers, losses or incomplete lineage sorting, with binary or multifurcated gene or species trees, optimize according to a parsimony score or a likelihood function, or sample on a bayesian distribution.

The aim of this article is to cater the need for a reconciliation software that would be easy to install (for example several current pieces rely on external libraries making the installation tedious and difficult), easy to use (the formats for gene names and species names, or reconciled gene trees are often very specific, not flexible, and not compatible with one another – reconciled trees are difficult to vizualize), efficient on simple functions (no current software can efficiently correct and root highly multifurcated gene trees at the same time) and that could be applied together with standard phylogenetic software using multiple sequence alignments. We present Treerecs, a new reconciliation software that possesses all these qualities.

2 Usability

Treerecs is available on Linux, Mac OSX and Microsoft Windows, and does not require any external library to be used[1]. Debian and RPM packages are currently in preparation.

As do all reconciliation software, Treerecs requires 3 kinds of information: a rooted species tree, one or more gene trees (rooted or not) and a mapping between the leaves of gene trees and those of the species tree. One frequent difficulty with reconciliation software is the strict formats for the inputs. Treerecs accepts Newick, NHX and PhyloXML, and does not

[1] CMake and a C++14 capable compiler are required to compile Treerecs

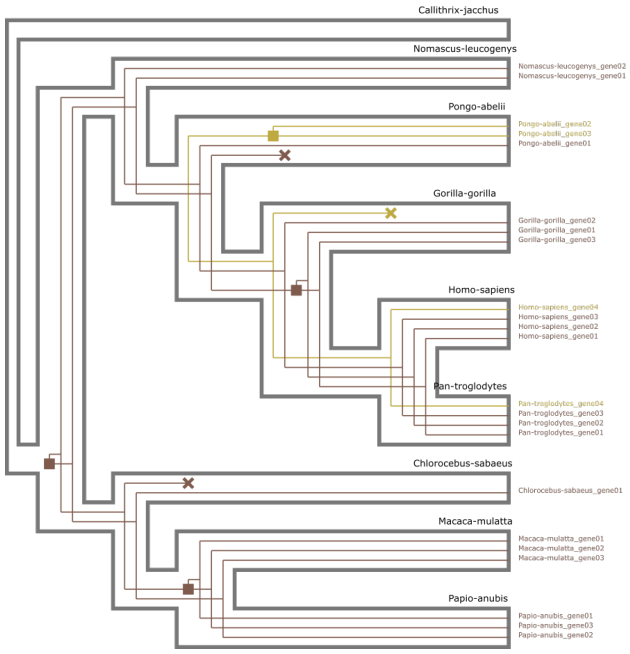

Fig. 1. Example of SVG output produced by Treerecs. Treerecs can draw several gene trees (two here) inside the associated species tree. Squares represent gene duplications events and crosses the losses.

require any special treatment for special characters frequently reserved for the format, as @,#,_,|. For the gene species mapping, it can be informed by the user in a separate file. In this case the column for genes and species does not matter. Alternatively, the mapped species can be specified in the gene names. In that case, the species name corresponding to a given gene is direclty sought in its name, with no requirement on its position nor on the separation character. Unless there is an ambiguity, *i.e.* if the gene name contains several species names, Treerecs is always able to infer the mapping automatically.

The output can also be given in a variety of formats including RecPhyloXML(Duchemin *et al.*, 2018), and SVG for display purposes (see Figure 1).

3 Efficiency

Basic functionalities of Treerecs include: (i) computing and showing reconciled gene phylogenies within the associated species tree with duplications and losses, (ii) rooting gene trees by searching a root minimizing the duplication and loss score (iii) resolving multifurcated nodes in gene trees, minimizing a duplication and loss cost, (iv) correcting gene trees by contracting weakly supported branches (according to a contraction threshold) and resolving the multifurcations minimizing the duplication and loss score. Resolution and correction are achieved using the ProfileNJ algorithm (Noutahi *et al.*, 2016). Note that the resolution and rooting can be done at the same time, a feature which no other current reconciliation software efficiently achieves. When there are several solutions for rooting, resolving or correcting, Treerecs outputs one or more random solutions from a uniform distribution. We provide metrics (likelihoods computed from site subtitution models and from gene content models) to give the possibility to choose among these solutions.

Figure 2 presents a comparison between Treerecs and three other reconciliation software: Ranger-DTL (Bansal *et al.*, 2018), Notung (Stolzer *et al.*, 2012) and ecceTERA (Jacox *et al.*, 2016) on these

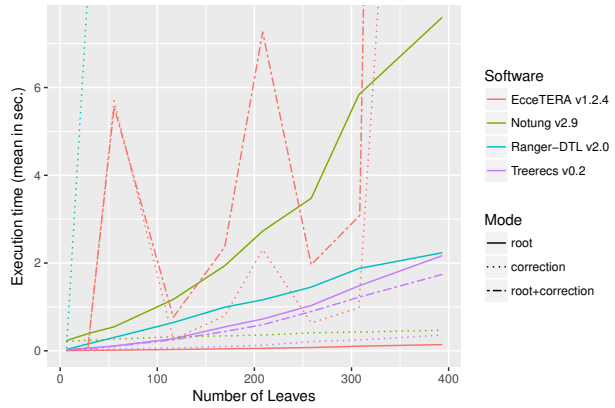

Fig. 2. Execution times (Intel® Core™ i7-6600U CPU @ 2.60GHz × 4) of different reconciliation software with similar functionnalities (average on 100 assays for each trial). Computation times are shown for 9 trees of different sizes from the Ensembl Compara database (V73) in three use cases. (i) root: searches a root that minimizes a reconciliation score, (ii) correction: creates polytomies by removing branches whose support is below a specified threshold (here the median of supports for each tree), (iii) root+correction: does both at the same time (note that only Treerecs and ecceTERA support this feature).

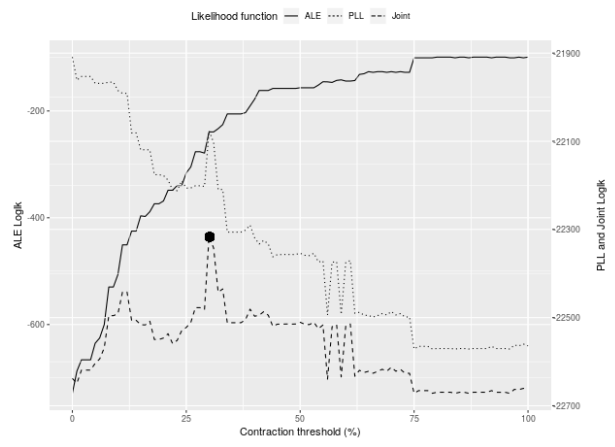

Fig. 3. Likelihoods of ALE and PLL. We are looking for a contraction threshold that can maximizing both the PLL function (on sequence information) and the ALE (on reconciliation). The dot shows the maximum of the joint likelihood (reached by contraction of branches with a support value below 31).

basic functions. Treerecs (purple) can be used in any situation with a reasonable execution time. With the exception of the rooting task, for which ecceTERA is notably the best, Treerecs is computationally more efficient than competing tools.

4 Integration

Treerecs offers more than just the basic functions presented above. In particular, it can compute the phylogenetic likelihood of a tree given a multiple sequence alignment, using the Phylogenetic Likelihood Library (PLL Flouri *et al.* (2014)). It also computes the reconciliation likelihood of a gene tree given the species tree with ALE (Szöllősi *et al.*, 2013). This allows, for instance, to explore the gene tree space, via a variation of the contraction threshold for branches with low support, scored by a joint likelihood. This procedure is illustrated in Figure 3.

Furthermore, Treerecs is integrated in the future version of Seaview (Gouy *et al.*, 2010, release 5.0 in preparation), already available from the Treerecs website. Thus, it can be used with a graphical interface along with a standard phylogenetic pipeline.

5 Evolvability

We are currently working on including Lateral Gene Transfers (LGT) in reconciliations, as well as on the possibility to provide unresolved species trees as input.

6 Acknowledgements

AS is funded by the Klaus Tschira Foundation. BM is funded via DFG Grant STA 860/6-1. We thank Tristan Lefébure and Marie Semon for their user feedback on beta versions.

References

Akerborg, O., Sennblad, B., Arvestad, L., and Lagergren, J. (2009). Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, **106**(14), 5714–5719.

Bailly-Bechet, M., Martins-SimÃµes, P., SzÃ¶llÅ‘si, G. J., Mialdea, G., Sagot, M.-F., and Charlat, S. (2017). How long does wolbachia remain on board? *Molecular Biology and Evolution*, **34**(5), 1183–1193.

Bansal, M. S., Kellis, M., Kordi, M., and Kundu, S. (2018). Ranger-dtl 2.0: rigorous reconstruction of gene-family evolution by duplication, transfer and loss. *Bioinformatics*, page bty314.

Duchemin, W., Gence, G., Chifolleau, A.-M. A., Arvestad, L., Bansal, M. S., Berry, V., Boussau, B., Chevenet, F., Comte, N., Davín, A. A., Dessimoz, C., Dylus, D., Hasic, D., Mallo, D., Planel, R., Posada, D., Scornavacca, C., Szöllősi, G., Zhang, L., Tannier, É., and Daubin, V. (2018). RecPhyloXML: a format for reconciled gene trees. *Bioinformatics*.

Flouri, T., Izquierdo-Carrasco, F., Darriba, D., Aberer, A., Nguyen, L.-T., Minh, B., Haeseler, A. V., and Stamatakis, A. (2014). The phylogenetic likelihood library. *Systematic Biology*, **64**(2), 356–362.

Gouy, M., Guindon, S., and Gascuel, O. (2010). Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution*, **27**(2), 221–224.

Jacox, E., Chauve, C., SzÃ¶llÅ‘si, G. J., Ponty, Y., and Scornavacca, C. (2016). eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, **32**(13), 2056–2058.

Noutahi, E., Semeria, M., Lafond, M., Seguin, J., Boussau, B., Guéguen, L., El-Mabrouk, N., and Tannier, E. (2016). Efficient gene tree correction guided by genome evolution. *PLOS ONE*, **11**(8), e0159559.

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

Szöllősi, G. J., Tannier, E., Daubin, V., and Boussau, B. (2015). The inference of gene trees with species trees. *Systematic Biology*, **64**(1), e42–e62.