BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 3 2015, pages 432–433
doi:10.1093/bioinformatics/btu648

Phylogenetics Advance Access publication October 1, 2014

# MulRF: a software package for phylogenetic analysis using multi-copy gene trees

Ruchi Chaudhary[1,*], David Fernández-Baca[2] and John Gordon Burleigh[1]

[1]Department of Biology, University of Florida, Gainesville, FL 32611 and [2]Department of Computer Science, Iowa State University, Ames, IA 50011, USA

Associate Editor: David Posada

## ABSTRACT

**Summary:** MulRF is a platform-independent software package for phylogenetic analysis using multi-copy gene trees. It seeks the species tree that minimizes the Robinson–Foulds (RF) distance to the input trees using a generalization of the RF distance to multi-labeled trees. The underlying generic tree distance measure and fast running time make MulRF useful for inferring phylogenies from large collections of gene trees, in which multiple evolutionary processes as well as phylogenetic error may contribute to gene tree discord. MulRF implements several features for customizing the species tree search and assessing the results, and it provides a user-friendly graphical user interface (GUI) with tree visualization. The species tree search is implemented in C++ and the GUI in Java Swing.

**Availability:** MulRF's executable as well as sample datasets and manual are available at http://genome.cs.iastate.edu/CBL/MulRF/, and the source code is available at https://github.com/ruchiherself/MulRFRepo.

**Contact:** ruchic@ufl.edu

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

Received on June 5, 2014; revised on July 30, 2014; accepted on September 29, 2014

## 1 INTRODUCTION

Resolving phylogenetic trees from large genomic datasets remains a challenge for evolutionary biologists. One major problem is to reconcile the often conflicting phylogenetic signals from different genes. Although some of these conflicts can be explained by stochastic or systematic error, biological processes such as incomplete lineage sorting, gene duplication and loss and lateral transfer also contribute to gene tree discord (Maddison, 1997). An effective phylogenetic method must address the multiple sources of discord among gene trees while remaining computationally tractable for large genomic datasets.

Non-probabilistic methods like those implemented in GeneTree (Page, 1998), iGTP (Chaudhary *et al.*, 2010), PhyloNet (Yu *et al.*, 2011) and SPRSupertree (Whidden *et al.*, 2014), and probabilistic methods like those implemented in BEST (Liu and Pearl, 2007), BUCKy (Ané *et al.*, 2007), *BEAST (Heled and Drummond, 2010), STEM (Kubatko *et al.*, 2009), MP-EST (Liu *et al.*, 2010) and PHYLDOG (Boussau *et al.*, 2012) can infer species trees while accounting for incongruence among genes. Although these methods differ widely in their details, with the exception of BUCKy, their reconciliation models are based on a single biological cause of discordance among gene trees (e.g. deep coalescence). Also, some of these methods, like SPRSupertree and BUCKy, are designed only for singly labeled input trees.

Chaudhary *et al.* (2013) introduced an alternative approach for building unrooted species trees from collections of unrooted, multi-copy gene trees (i.e. trees with multiple leaves having the same label). This approach is based on the *MulRF distance*, a generalization of the Robinson–Foulds (RF) distance (Robinson and Foulds, 1981). The MulRF distance between a species tree and an input gene tree is obtained by first extending the species tree by adding additional copies of a leaf label in the star tree topology (Fig. 1). The RF distance is then calculated between the multi-copy gene tree and the extended species tree. The MulRF distance between a species tree and an input gene tree can be computed in time linear in the number of leaves of the multi-copy gene tree (Chaudhary *et al.*, 2013). A *MulRF supertree* is a species tree that minimizes the sum of the MulRF distances to the input gene trees. The MulRF cost function is not explicitly linked to any biological process, and its input gene trees are not limited to orthologous gene sequences.

Computing MulRF supertrees is NP-hard (Chaudhary *et al.*, 2012), but there is an efficient hill-climbing heuristic for estimating them based on the unrooted subtree prune and regraft (SPR) local search (Chaudhary *et al.*, 2013). Here we describe MulRF version 1.2, a software package that incorporates this heuristic, along with several features to improve the species tree search and assess the results. These features include the ability to: (i) weight the input gene trees, (ii) constrain the species tree topology in the tree search, (iii) automate multiple heuristic searches and (iv) evaluate the score of alternate species tree topologies. The software package also provides a graphical user interface with tree visualization to facilitate the evaluation of the results.

## 2 PROGRAM DESCRIPTION

The input for species tree search is a single file containing the collection of input gene trees, and the output is the best species tree found in the search. All the input and the output trees are in the Newick format. MulRF begins its tree-search heuristic from an initial species tree, which can either be supplied by the user or built using a step-wise leaf-adding algorithm. MulRF allows users to weight each input gene tree with a real number from 0 to 1. During species tree search, the MulRF distance from each

*To whom correspondence should be addressed.

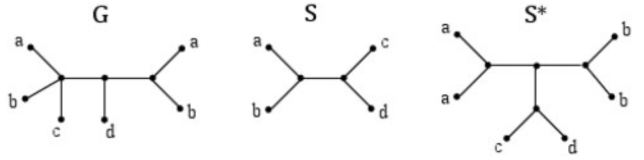

**Fig. 1.** Multi-copy gene tree *G* and the species tree *S*. The extension *S** of *S* relative to *G* is also shown. The MulRF distance between *G* and *S*, which is the RF distance between *G* and *S**, is 5

gene tree to the candidate species tree is multiplied by the gene tree weight, and the species tree is inferred based on the sum of the weighted MulRF distances from all gene trees. Users also can impose topological constraints on the inferred species trees. The MulRF heuristic is a randomized algorithm, and the tree search may get caught in a local minima. To explore the tree space more thoroughly, we recommend executing the tree-search heuristic multiple times on the same dataset. MulRF enables the user to automate this by running multiple replicates of the heuristic, each of which uses a different random seed.

The MulRF software package also enables users to determine the total weighted or un-weighted MulRF distance between a given species tree and a collection of gene trees, along with the MulRF distance of each gene tree. This allows users to compare alternate species tree topologies and identify the gene trees that most affect the resolution of a species tree. MulRF has an intuitive user interface. All available options are divided into four menu items: *File*, *Analysis*, *Options* and *Help*. The File menu allows opening and closing input gene tree and scoring files. The Analysis menu allows users to build and score species trees. The Options menu allows customizing the tree search. The Help menu provides descriptions of various menu items. MulRF visualizes the input gene trees and the inferred species tree using ATV (Zmasek and Eddy, 2001).

## 3 PERFORMANCE EVALUATION

MulRF has been extensively tested in two studies using gene tree simulations that included duplication and loss (Chaudhary *et al.*, 2014), and duplication, loss and lateral transfer (Chaudhary *et al.*, 2013). Both studies demonstrated that MulRF can easily run on datasets with hundreds of taxa and a thousand gene trees, and it often provides more accurate phylogenetic estimates than iGTP, SPRSupertree and PHYLDOG (Chaudhary *et al.*, 2013, 2014).

We executed MulRF on a set of 6966 gene trees from 36 mammalian species that were inferred using PhyML (Guindon *et al.*, 2010) by Boussau *et al.* (2012). Boussau *et al.* (2012) used PHYLDOG to estimate a phylogeny from these gene sequences. PHYLDOG and MulRF produced credible estimates of the phylogeny that were identical except for the placement of Chiroptera within Laurasiatheria (Boussau *et al.*, 2012, Fig. 5; Supplementary Fig. S1). PHYLDOG placed Chiroptera as sister to a clade containing Carnivora and Perissodactyla (Boussau *et al.* 2012, Fig. 5), a position supported by Nery *et al.* (2012). In contrast, MulRF placed Chiroptera as sister to a clade containing Carnivora, Perissodactyla and Cetartiodactyla (Supplementary Fig. S1), a position supported by McCormack *et al.* (2012). While the PHYLDOG analysis used 3000 processors running in parallel for 10 days to estimate the gene trees and species tree from the mammalian sequences, MulRF took only 1 h 47 min on Intel Core 2 Duo 3.16 GHz PC with 4 GB RAM to reconstruct the mammalian tree from the gene trees.

## 4 CONCLUSION

MulRF is a software package for phylogenetic inference using multi-copy gene trees based on a generalization of the RF distance. In experiments on biological or simulated datasets, MulRF has performed well compared with both probabilistic and non-probabilistic supertree approaches. MulRF's speed and performance on a variety of datasets suggests that the program is well suited for species tree reconstruction from large genomic datasets.

## ACKNOWLEDGEMENTS

The authors thank the reviewers for their thoughtful comments.

*Funding*: The National Science Foundation (DEB-1208428 and CCF-1017189).

*Conflict of interest*: none declared.

## REFERENCES

Ané,C. *et al.* (2007) Bayesian estimation of concordance among gene trees. *Mol. Biol. Evol.*, **24**, 1575.

Boussau,B. *et al.* (2012) Genome-scale coestimation of species and gene trees. *Genome Res.*, **23**, 323–330.

Chaudhary,R. *et al.* (2010) iGTP: A software package for large-scale gene tree parsimony analysis. *BMC Bioinformatics*, **11**, 574.

Chaudhary,R. *et al.* (2012) Fast local search for unrooted Robinson-Foulds supertrees. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **9**, 1004–1013.

Chaudhary,R. *et al.* (2013) Inferring species trees from incongruent multi-copy gene trees using the Robinson-Foulds distance. *Algorithms Mol. Biol.*, **28**, 8.

Chaudhary,R. *et al.* (2014) Assessing approaches for inferring species trees from multi-copy genes (under review).

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.

Heled,J. and Drummond,A.J. (2010) Bayesian inference of species trees from multilocus data. *J. Mol. Biol. Evol.*, **27**, 570–580.

Kubatko,L.S. *et al.* (2009) STEM: species tree estimation using maximum likelihood for gene trees under coalescence. *Bioinformatics*, **25**, 971–973.

Liu,L. and Pearl,D.K. (2007) Species trees from gene trees: reconstructing bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. *Syst. Biol.*, **56**, 504–514.

Liu,L. *et al.* (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.*, **10**, 302.

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

McCormack,J.E. *et al.* (2012) Ultraconserved elements are novel phylogenomic markers that resolve placental mammal phylogeny when combined with species-tree analysis. *Genome Res.*, **22**, 746–754.

Nery,M.F. *et al.* (2012) Resolution of the laurasiatherian phylogeny: evidence from genomic data. *Mol. Phylogenet. Evol.*, **64**, 685–689.

Page,R.D.M. (1998) GeneTree: comparing gene and species phylogenies using reconciled trees. *Bioinformatics*, **14**, 819–820.

Robinson,D.F. and Foulds,L.R. (1981) Comparison of phylogenetic trees. *Math. Biosci.*, **53**, 131–147.

Whidden,C. *et al.* (2014) Supertrees based on the subtree prune-and-regraft distance. *Syst. Biol.*, **63**, 566–581.

Yu,Y. *et al.* (2011) Algorithms for MDC-based multi-locus phylogeny inference. In: *RECOMB 2011*. Springer, pp. 531–545.

Zmasek,C.M. and Eddy,S.R. (2001) ATV: display and manipulation of annotated phylogenetic trees. *Bioinformatics*, **17**, 383–384.