Phylogenetics

MSCquartets 1.0: quartet methods for species trees and networks under the multispecies coalescent model in R

John A. Rhodes ^{1,*}, Hector Baños², Jonathan D. Mitchell^{1,3} and Elizabeth S. Allman¹

¹Department of Mathematics and Statistics, University of Alaska Fairbanks, Fairbanks, AK 99775-6660, USA, ²School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160, USA and ³Unité Bioinformatique Evolutive, C3BI USR 3756, Institut Pasteur & CNRS, Paris, France

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on May 1, 2020; revised on September 17, 2020; editorial decision on September 21, 2020; accepted on September 23, 2020

Abstract

Summary: MSCquartets is an R package for species tree hypothesis testing, inference of species trees and inference of species networks under the Multispecies Coalescent model of incomplete lineage sorting and its network analog. Input for these analyses are collections of metric or topological locus trees which are then summarized by the quartets displayed on them. Results of hypothesis tests at user-supplied levels are displayed in a simplex plot by color-coded points. The package implements the QDC and WQDC algorithms for topological and metric species tree inference, and the NANUQ algorithm for level-1 topological species network inference, all of which give statistically consistent estimators under the model.

Availability and implementation: MSCquartets is available through the Comprehensive R Archive Network: <https://CRAN.R-project.org/package=MSCquartets>.

Contact: j.rhodes@alaska.edu

1 Introduction

The advent of large multilocus datasets, often with discordant gene phylogenies, requires new statistical methods of species tree and network inference, with attention to attaining acceptable runtimes. The *Multispecies Coalescent* (MSC) and Network MSC models of incomplete lineage sorting (ILS), which is believed to be one of the most fundamental causes of gene tree discord, provides the theoretical framework for such work. Based on diverse methodologies, existing software for tree inference under these coalescent models include *Beast (Heled and Drummond, 2010), RevBayes (Höhna et al., 2016), BPP (Flouri et al., 2018), MP-EST (Liu et al., 2010), ASTRAL (Zhang et al., 2018b), ASTRID (Vachaspati and Warnow, 2015) and SVDQuartets (Kubatko and Chifman, 2019), and for network inference SNaQ (Solís-Lemus and Ané, 2016), PhyloNet (Yu and Nakhleh, 2015) and SpeciesNetwork (Zhang et al., 2018a).

MSCquartets enables investigations of fit of the MSC model to data and inference of species trees/networks through *quartet* summaries of gene tree collections. If n gene trees have been inferred on (subsets of) N taxa, the collection can be summarized by tabulating the counts of quartets displayed on the n trees for each of the $\binom{N}{4}$ 4-element subsets of the taxa. For example, if the four taxa $\{a, b, c, d\}$ occur on $m \leq n$ of the gene trees, and the trees are fully

resolved, then the counts $m_{ab|cd}$, $m_{ac|bd}$, $m_{ad|bc}$ of the three possible displayed topologies can be tabulated, where $xy|zw$ denotes the binary quartet tree separating xy from zw . The resultant *quartet count concordance factor* $qcCF = (m_{ab|cd}, m_{ac|bd}, m_{ad|bc})$ or its normalization $\widehat{CF} = \frac{1}{m} qcCF$, the *empirical quartet concordance factor* (Allman et al., 2020), are the data summaries of focus. The vector $\widehat{CF}$ estimates the expectation eCF under the MSC of quartet frequencies displayed on gene trees. As the number m of gene trees sampled from the MSC grows, $\widehat{CF} \rightarrow eCF$ in probability.

While basic MSCquartets functions compute concordance factor summaries of multilocus datasets, the main analysis tools perform:

1. Hypothesis testing of $qcCF$ s for fit of the MSC on a 4-taxon tree,
2. Inference of a large species tree from $qcCF$ s,
3. Inference of a large topological species network from $qcCF$ s.

Built on the ape and phangorn packages (Paradis and Schliep, 2019; Schliep, 2011), MSCquartets functions are designed for flexibility and easy export of output from R. Specialized graphics functions allow for display of hypothesis test results. Sample datasets, both simulated and empirical, are included, and example code provided in the documentation illustrates function calls to perform analyses.

2 Materials and methods

Data analysis with MSCquartets generally begins by tabulating qcCFs for a multilocus tree collection. This step's running time of $O(N^4n)$ dominates that of the more substantive analyses 1–3 above, both theoretically and practically, but need only be performed once. This running time allows for analyses of datasets with a moderate number N of taxa but a genomic-scale number n of loci. For instance, counting all displayed quartets in a mammalian dataset of 424 gene trees on 37 taxa required about half an hour on a laptop computer with a 2.9 GHz processor and 16 GB memory.

2.1 Hypothesis testing

Using new methodology developed by Mitchell *et al.* (2019), a qcCF can be tested for fit to the MSC on 4-taxon unrooted topological species trees.

Forms of the tests implemented in MSCquartets allow the null hypothesis H_0 : The qcCF arose from a species quartet tree of specified topology under the MSC, or H_0 : The qcCF arose from a species quartet tree of unspecified topology under the MSC. With exhaustive alternative hypotheses, rejection of the null indicates in the first case that the MSC on the specific species tree topology is rejected, and in the second that the MSC on any tree is rejected (perhaps due to hybridization). A third test, using more routine statistical methods, can test the MSC on a quartet star tree (a polytomy) against any alternative, as has also been implemented by Sayyari and Mirarab (2018).

The primary hypothesis tests use a novel approximating distribution, dependent on the sample size m , that performs better than the standard asymptotic distribution for models such as these with singularities or boundaries. A simplex plot, such as in Figure 1 for a *Heliconius* dataset with $N=16$ taxa (Martin *et al.*, 2013), displays test results for all $\binom{N}{4} = 1820$ subsets of 4 taxa. Since each $\widehat{CF}$ lies in the equilateral triangle formed by the intersection of the plane $x+y+z=1$ and the positive orthant $x,y,z \geq 1$, it can be represented by a point in the 2-dimensional probability simplex. For each 4-taxon set, a color-coded symbol is plotted at the $\widehat{CF}$, with the black lines indicating all possible eCFs under the MSC for the three 4-taxon species tree topologies. A blue circle or gold square, for example, indicates that a particular qcCF is in accord, at the specified level α , with the MSC on a species tree. Such points typically lie in closer proximity to the black model lines than do the red or green symbols indicating rejection of the null hypothesis. The simplex plot thus gives an easily comprehended visualization of fit of the full dataset to the MSC model.

A more extensive explanation of simplex plots, their interpretation and examples of use of these hypothesis tests on simulated and empirical datasets is given by Allman *et al.* (2020). By implementing a new approach for the evaluation of fit of the MSC to a dataset before either a species tree or network is inferred, MSCquartets allows for more complete analysis of genomic datasets than has been possible previously.

2.2 QDC and WQDC species tree estimators

Unrooted species trees, either topological or metric, can be inferred from topological quartets displayed on gene trees by first computing novel intertaxon distances introduced by Rhodes (2019) and Yourdkhani and Rhodes (2020). MSCquartets includes functions to find these distances from tabulated qcCFs, allowing for rapid statistically consistent inference under the MSC model, using standard tree building methods such as Neighbor Joining, which are already available in ape.

These methods offer an alternative to other quartet methods for inferring a species tree, such as heuristics for solving the Maximum Quartet Consistency problem. If quartets have already been tabulated in order to perform hypothesis testing to judge fit to the MSC, both running times and accuracy of QDC are competitive with the quartet-based ASTRAL.

Fig. 1. (Top) A simplex plot for *Heliconius* data shows results of two hypothesis tests. Each of the 1820 plotted symbols represents a $\widehat{CF}$ for one subset of 4 of the 16 taxa. Rejection levels are $\alpha = 10^{-9}$ for a test of an unspecified tree topology under the MSC, and $\beta = 10^{-4}$ for the polytomy test. With multiple individuals sampled for some taxa, failure to reject some polytomies is not surprising. (Bottom) The splits network constructed from the NANUQ distance for these data, presenting strong evidence for a 4-cycle in the species network

2.3 NANUQ species network topology estimator

The statistically consistent NANUQ algorithm for estimating a species network topology under the Network MSC was introduced by Allman *et al.* (2019), and used for analyzing both simulated and empirical datasets. With current Bayesian methods for this problem effective only for very small datasets, and even pseudolikelihood approaches relatively slow, NANUQ's distance-based approach is by far the fastest for moderate N and large n . This method takes as input a collection of gene trees summarized by their qcCFs, uses the outcome of the hypothesis tests discussed above to determine putative 4-taxon networks, and a modification to level-1 networks of the intertaxon quartet distance used by QDC to create a distance table. From the distance table, a splits graph is constructed after projecting to a circular split system with NeighborNet (Bryant and Moulton, 2003). Such a splits graph is shown for the *Heliconius* data in Figure 1. A last step involves interpretation of the splits graph using theoretically derived rules to replace certain splits graph substructures with cycles to obtain an inferred topological species network, omitting any 2- and 3-cycles. A full explanation is given by Allman *et al.* (2019). In MSCquartets, a single function call accomplishes all steps up to the use of NeighborNet, which can be performed with phangorn, or externally to R with SplitsTree (recommended) (Huson and Bryant, 2006). Running time of NANUQ for is typically measured in minutes for datasets that may require days by pseudolikelihood methods and not be analyzable at all by current Bayesian methodology. Figure 1 shows the NANUQ splits graph for the *Heliconius* data, displaying a reticulation that has been supported by other, more time-intensive analyses.

3 Conclusion

The R package MSCquartets implements a suite of fast and statistically consistent algorithms for species tree and network inference

under the MSC and Network MSC models. In addition, it implements novel hypothesis testing methods for the fit of multilocus datasets to the MSC model. All of these are based on recent theoretical advances on the relationship between the topology of a tree or network and expected quartet concordance factors, providing a common statistical framework for the package.

Funding

This work was supported by the National Institutes of Health [R01 GM117590], under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological Mathematical Sciences and [2P20GM103395], an NIGMS Institutional Development Award (IDeA).

Conflict of Interest: none declared.

References

- Allman, E. *et al.* (2019) NANUQ: a method for inferring species networks from gene trees under the coalescent model. *Algorithms Mol. Biol.*, **14**, 1–25.
- Allman, E. *et al.* (2020) Gene tree discord, simplex plots, and statistical tests under the coalescent. *bioRxiv*, doi:10.1101/2020.02.13.948083.
- Bryant, D. and Moulton, V. (2003) Neighbor-Net: an agglomerative method for the construction of phylogenetic networks. *Mol. Biol. Evol.*, **21**, 255–265.
- Flouri, T. *et al.* (2018) Species tree inference with BPP using genomic sequences and the multispecies coalescent. *Mol. Bio. Evol.*, **35**, 2585–2593.
- Heled, J. and Drummond, A.J. (2010) Bayesian inference of species trees from multilocus data. *Mol. Bio.. Evol.*, **27**, 570–580.
- Höhna, S. *et al.* (2016) RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Syst. Biol.*, **65**, 726–736.
- Huson, D. and Bryant, D. (2006) Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, **23**, 254–267.
- Kubatko, L. and Chifman, J. (2019) An invariants-based method for efficient identification of hybrid species from large-scale genomic data. *BMC Evol. Biol.*, **19**.
- Liu, L. *et al.* (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.*, **10**, 302.
- Martin, S. *et al.* (2013) Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. *Genome Res.*, **23**, 1817–1828.
- Mitchell, J. *et al.* (2019) Hypothesis testing near singularities and boundaries. *Electron. J. Stat.*, **13**, 2150–2193.
- Paradis, E. and Schliep, K. (2019) ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, **35**, 526–528.
- Rhodes, J. (2019) Topological metrizations of trees, and new quartet methods of tree inference. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, early access.
- Sayyari, E. and Mirarab, S. (2018) Testing for polytomies in phylogenetic species trees using quartet frequencies. *Genes*, **9**, E132.
- Schliep, K. (2011) phangorn: phylogenetic analysis in R. *Bioinformatics*, **27**, 592–593.
- Solís-Lemus, C. and Ané, C. (2016) Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.*, **12**, e1005896.
- Vachaspati, P. and Warnow, T. (2015) ASTRID: accurate Species TREes from Internode Distances. *BMC Genomics*, **16**, S3.
- Yourdkhani, S. and Rhodes, J. (2020) Inferring metric trees from weighted quartets via an intertaxon distance. *Bul. Math. Biol.*, **7**, 97.
- Yu, Y. and Nakhleh, L. (2015) A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics*, **16**, S10.
- Zhang, C. *et al.* (2018a) Bayesian inference of species networks from multilocus sequence data. *Mol. Biol. Evol.*, **35**, 504–517.
- Zhang, C. *et al.* (2018b) ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. *BMC Bioinformatics*, **19**, 15–30.