Bioinformatics, 36(18), 2020, 4819–4821
doi: 10.1093/bioinformatics/btaa605
Advance Access Publication Date: 1 July 2020
Applications Note

Phylogenetics

PRANC: ML species tree estimation from the ranked gene trees under coalescence

Anastasiia Kim* and James H. Degnan

Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87106, USA

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on January 17, 2020; revised on June 16, 2020; editorial decision on July 17, 2020; accepted on June 23, 2020

Abstract

Summary: *PRANC* computes the Probabilities of RANked gene tree topologies under the multispecies coalescent. A ranked gene tree is a gene tree accounting for the temporal ordering of internal nodes. *PRANC* can also estimate the maximum likelihood (ML) species tree from a sample of ranked or unranked gene tree topologies. It estimates the ML tree with estimated branch lengths in coalescent units.

Availability and implementation: *PRANC* is written in C++ and freely available at github.com/anastasiiakim/PRANC.

Contact: anastasiia.kim@protonmail.com

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

1 Introduction

A species (respectively, gene) tree represents the evolutionary relationships among a set of species (respectively, genes). Discordance between species trees and gene trees is often modeled by the multispecies coalescent (MSC). The MSC is widely used to infer species trees directly from sequence data (Bryant *et al.*, 2012; Chifman and Kubatko, 2014; Heled and Drummond, 2010; Ronquist *et al.*, 2012; Yang, 2015), from unrooted gene tree topologies (Larget *et al.*, 2010; Liu and Yu, 2011; Mirarab *et al.*, 2014), or from rooted unranked gene tree topologies (Liu *et al.*, 2009; 2010; Pei and Wu, 2017; Wu, 2012), or from the gene trees with branch lengths (Kubatko *et al.*, 2009; Liu *et al.*, 2009).

We introduce *PRANC*, which computes the probabilities of ranked gene tree topologies under the MSC (Degnan *et al.*, 2012; Stadler and Degnan, 2012). We use ranked gene trees, which describe both the topological relationships among gene lineages and the order in which gene lineages coalesce. For example, the trees ((A:0.1, B:0.1):1.0,(C:1.0, D:1.0):0.1) and ((A:0.5, B:0.5):0.5,(C:0.1, D:0.1):0.9) are converted to ranked trees ((A, B)$_3$,(C, D)$_2$)$_1$ and ((A, B)$_2$,(C, D)$_3$)$_1$, respectively, where the subscripted number is the rank of the internal node starting from most ancient (the root) to most recent. Ranked trees preserve some branch length information, giving the potential to improve species tree inference. *PRANC* is the first method we are aware of that uses ranked gene trees to infer species trees.

PRANC takes a set of ranked gene tree topologies and searches for the maximum likelihood (ML) species tree. We evaluate the performance of *PRANC* in comparison with *STELLS2* and *ASTRAL* using a gibbon dataset (Carbone *et al.*, 2014; Shi and Yang, 2018).

2 Description

Let $\mathcal{T}$ be an n-taxon rooted species tree with branch lengths. Assuming that we have observed a collection of N independent ranked gene trees $\mathcal{G}_i, i = 1, 2, \ldots, N$, the ML species tree is

$$\mathcal{T}_{\mathrm{ML}} = \mathrm{argmax}_{\mathcal{T}}\, P[\mathcal{G}_1, \ldots, \mathcal{G}_N|\mathcal{T}] = \mathrm{argmax}_{\mathcal{T}} \prod_{i=1}^{N} P[\mathcal{G}_i|\mathcal{T}]. \qquad (1)$$

The probability of the ranked gene tree $P(\mathcal{G}|\mathcal{T})$ is described elsewhere (Degnan *et al.*, 2012; Kim *et al.*, 2020; Stadler and Degnan, 2012).

PRANC seeks to find a species tree with branch lengths in coalescent units $\mathcal{T}$ that maximizes the likelihood given by Equation (1). Supplementary Table S1 shows some available options.

To estimate the ML species tree, *PRANC* uses the following steps:

1. Process the initial species tree. If the tree has the branch lengths specified in coalescent units, treat the tree as a ranked tree. Find a set of speciation interval lengths that maximizes the likelihood. If the branch lengths are not specified in the tree, generate all possible rankings. Randomly select a subset of ranked trees (by default, all rankings will be considered but a subset of rankings can be considered, $2n$ rankings work well). Define the speciation interval length between the $(i-1)$th and ith speciation events as $t_i = s_{i-1} - s_i$, where s_i is the time of the interior node of rank i. For each of these trees, initialize each interval length t_i to 1.0 and find a set of speciation interval lengths that maximizes the likelihood. Pick the tree $\mathcal{T}$ with the highest likelihood.

2. Obtain all trees that are one nearest-neighbor interchange (NNI) away from $\mathcal{T}$. For each of these unranked trees, generate all possible ranked trees. Randomly select a subset of ranked trees (by default, $2n$ rankings). Find the speciation interval lengths that maximizes the likelihood of the ranked gene trees and pick the one with the highest likelihood. If this tree has a larger likelihood, then set $\mathcal{T}$ to this tree.
3. Repeat step 2 until convergence or until all trees within k (by default, $k = 5$) NNI steps are explored.
4. Calculate the branch lengths of the inferred tree. *PRANC* primarily estimates interval lengths and then calculates internal branch lengths from them. For convenience, the time of the most recent clade is set to 0.1 but could be set to any other value because it does not affect the probabilities.

The starting tree can be computed using *PRANC* using greedy consensus (i.e. extended majority rule) (Bryant, 2003) or some other options (see software). Faster species tree methods, such as *ASTRAL*, can also be used to supply the starting tree, which *PRANC* can then improve upon (see Supplementary Material). A list of starting trees can be provided by the user in a single file. *PRANC* then applies step 1 to each of these and finds the highest likelihood among the starting trees before proceeding to step 2.

PRANC optimizes the interval lengths using Brent's method (Brent, 1973) and L-BFGS: limited memory algorithm for bound constrained optimization method (Byrd *et al.*, 1995). We compute the initial likelihood for the tree obtained in step 1. Then, we optimize each length one at a time using Brent's method, fixing the other lengths. We randomly pick interval orders for optimization. After m rounds of such optimizations (by default, m is set to the number of taxa n), the optimal tree is reported. We allow the length to be in the interval $[0.001, 6]$. As an alternative, we propose to use L-BFGS method for the interval lengths optimization. It is well suited for the negative log likelihood minimization because it can minimize across multiple variables at the same time and the boundaries for the allowed values that parameters can take can be defined in L-BFGS method. We found that for the small-scale simulation L-BFGS method runs faster than Brent's method.

For balanced topologies, far more rankings exist than for less-balanced topologies. Computing the likelihood of gene trees for every possible ranked topology is not efficient for $n > 7$-taxon trees. We observed that in most cases, the values of likelihoods for different rankings of the same unranked species tree are close to each other. Therefore, in step 2, we compute likelihoods of a randomly chosen small subset of rankings (by default, n rankings). If at least one of the obtained n likelihoods is larger than the threshold, then *PRANC* computes the likelihoods for a larger subset of rankings (by default, up to $2n$ rankings). The user can change *PRANC*'s settings, such as the method for branch length optimization, the number of rankings to consider for each unranked species tree candidate, the number of NNI moves and the allowed length for each speciation interval.

3 Example

We used *PRANC* to infer the species tree for a genome-scale dataset consisting of 5 gibbons species with 12 413 non-coding loci, each of length 1000 bp. (Carbone *et al.*, 2014; Veeramah *et al.*, 2015). We used *IQ-TREE* (Nguyen *et al.*, 2015) to estimate 6-taxon unrooted gene trees from DNA sequences under the $GTR + \Gamma$ model. We rooted these trees on the outgroup, and then dropped it to get 5-taxon rooted gene trees and made them ultrametric. We obtained a subset of 10 706 5-taxon rooted ultrametric gene trees that passed a molecular clock test (Felsenstein, 2004) running all 12 413 trees through *Dnaml* and *Dnamlk* (Felsenstein, 2013).

BPP (Yang, 2015), *ASTRAL* (Mirarab *et al.*, 2014), *STELLS2* (Pei and Wu, 2017) and *PRANC* were used to estimate species trees from DNA sequences, a sample of unrooted, unranked and ranked gene tree topologies, respectively. We observed that all three methods converge as the number of genes increases to the species tree topology obtained by Shi and Yang (2018) using *BPP* (Fig. 1) using all loci. The results shown in Figure 1 are intuitive. Unrooted, unranked and ranked trees preserve increasing amounts of information about the rooted trees with specified branch lengths, respectively. As expected, for a small number of genes, the Bayesian *BPP* that calculates the posterior probabilities of different species trees from DNA sequences was more likely to estimate a tree that matched the species tree obtained from larger samples.

We also compared *PRANC*'s performance with *ASTRAL* and *STELLS2* on simulated data. We simulated 100 $n = 5-, 6-, 7-, 8-$taxon species trees under the Yule model with speciation rates $\lambda = 0.5$ and $\lambda = 1$. We used *TreeSim* (Stadler, 2011) to generate species trees and *hybrid-lambda* (Zhu *et al.*, 2015) to simulate 100, 500 and 1000 gene trees for each species tree. The greedy consensus tree and the trees estimated by *ASTRAL* and *STELLS2* were used as starting trees for *PRANC*. We ran *PRANC* under different settings. The simulation results are shown in Supplementary Figures S1 and S2, using true (not estimated) gene trees. In practice, users will need to first provide estimates of the gene trees, including either divergence times or ranking information. Under default settings described in Section 2, on average, *PRANC* can estimate an $n = 5$-, 6-, 7-taxon species tree from 100 and 1000 gene trees in seconds and in a few minutes, respectively. On average, *PRANC* can estimate an $n = 8$-taxon tree from 100 and 1000 gene trees in a few minutes and in 30–45 min, respectively. Both *ASTRAL* and *STELLS2* are much faster. *ASTRAL* runs in seconds to estimate an 8-taxon species tree from 1000 gene trees. It usually takes several minutes for *STELLS2* to estimate an 8-taxon species tree from 1000 gene trees.

To see how well *PRANC* can estimate branch lengths, we considered 100 estimated species trees by *PRANC*, *STELLS2* and *ASTRAL*. Note that, all 100 inferred trees had the same unranked topologies as their corresponding species trees. In particular, we generated 100 5- and 8-taxon trees under the Yule model with the birth rate $\lambda = 0.5$ and $\lambda = 1$. In each case, 100 or 1000 gene trees were generated from each species tree and were used to estimate the species tree internal branch lengths. For each inferred tree, we calculated the mean squared error between true and estimated internal branch lengths. On average, *PRANC* estimates branch lengths more accurately than *ASTRAL* and *STELLS2*. As expected, using 1000

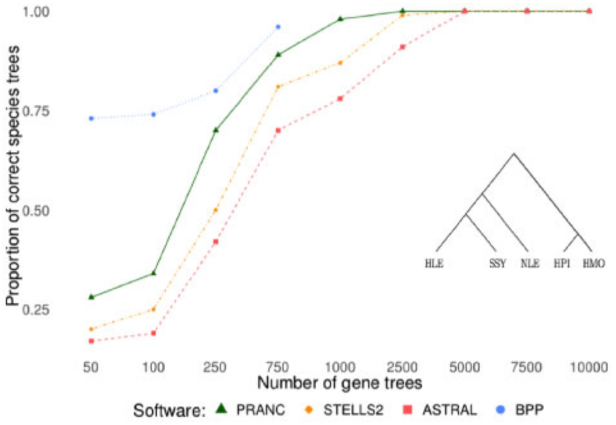

Fig. 1. The proportion of correct species trees in a gibbon dataset obtained by four different methods plotted against the number of gene trees. Note that, the true tree is unknown. We compared estimated unranked species tree topologies with that obtained by *BPP* (Yang, 2015). The gene trees were considered as ranked for *PRANC*, unranked for *STELLS2* and unrooted for *ASTRAL*. The greedy consensus tree was used as a starting tree for *PRANC*. To get an estimated rooted species tree from *ASTRAL*, we added an outgroup to the 5-taxon unranked gene trees. Then the estimated unrooted species tree by *ASTRAL* was rooted on the outgroup, and the outgroup was dropped to get a rooted 5-taxon tree. DNA sequences and no trees were used for *BPP*. The results for *BPP* were computed using up to 750 gene trees. In terms of approximate computational time, it took a few hours for *BPP* to estimate a 5-taxon species tree from 100 to 500 loci, whereas it took seconds to estimate a 5-taxon species tree from 100 to 500 gene trees with the other three programs

gene trees instead of 100 trees resulted in more accurate estimates for all three programs (Supplementary Table S2 and Figs S3–S5).

4 Conclusion

PRANC is a computational framework to work with the ranked gene trees. *PRANC* performs a heuristic search from the initial trees to find an ML species tree. There is a trade-off between *PRANC*'s estimation accuracy and its computational time. The speed of the program mainly depends on the choice of initial tree and the number of rankings considered for each unranked species tree candidate. We tested *PRANC*'s performance under different settings. In general, it is sufficient to consider $2n$ rankings for each unranked n-taxon species tree candidate. More rankings can be considered to improve accuracy at the expense of speed. In a gibbon dataset and in simulations on 5-8–taxon trees, *PRANC* outperformed *STELLS2* and *ASTRAL* (Supplementary Fig. S1). On average, *PRANC* estimated branch lengths more accurately than *ASTRAL* and *STELLS2* based on mean-squared error (Supplementary Table S2 and Figs S3–S5).

Acknowledgements

The authors thank the UNM Center for Advanced Research Computing, supported in part by the National Science Foundation, for providing the high-performance computing resources used in this work. They also thank Ziheng Yang and two anonymous reviewers for comments.

Funding

This work was supported by National Institutes of Health R01 [GM117590].

Conflict of Interest: none declared.

References

Brent,R. (1973) *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, New Jersey.

Bryant,D. (2003) A classification of consensus methods for phylogenetics. *DIMACS Ser. Discret. Math. Theor. Comput. Sci.*, **61**, 163–184.

Bryant,D. *et al.* (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, **29**, 1917–1932.

Byrd,R.H. *et al.* (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.*, **16**, 1190–1208.

Carbone,L. *et al.* (2014) Gibbon genome and the fast karyotype evolution of small apes. *Nature*, **513**, 195–201.

Chifman,J. and Kubatko,L. (2014) Quartet inference from SNP data under the coalescent model. *Bioinformatics*, **30**, 3317–3324.

Degnan,J.H. *et al.* (2012) The probability distribution of ranked gene trees on a species tree. *Math. Biosci.*, **235**, 45–55.

Felsenstein,J. (2004) *Inferring Phylogenies*. Sinauer associates, Sunderland, MA.

Felsenstein,J. (2013) *PHYLIP (Phylogeny Inference Package) Version 3.695*. Distributed by the Author. http://evolution.genetics.washington.edu/phylip.html.

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

Kim,A. *et al.* (2020) Probabilities of unranked and ranked anomaly zones under birth–death models. *Mol. Biol. Evol.*, **37**, 1480–1494.

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

Larget,B.R. *et al.* (2010) Bucky: gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics*, **26**, 2910–2911.

Liu,L. and Yu,L. (2011) Estimating species trees from unrooted gene trees. *Syst. Biol.*, **60**, 661–667.

Liu,L. *et al.* (2009) Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, **58**, 468–477.

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

Mirarab,S. *et al.* (2014) Astral: genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, i541–i548.

Nguyen,L.-T. *et al.* (2015) IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.

Pei,J. and Wu,Y. (2017) STELLS2: fast and accurate coalescent-based maximum likelihood inference of species trees from gene tree topologies. *Bioinformatics*, **33**, 1789–1797.

Ronquist,F. *et al.* (2012) MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, **61**, 539–542.

Shi,C.-M. and Yang,Z. (2018) Coalescent-based analyses of genomic sequence data provide a robust resolution of phylogenetic relationships among major groups of gibbons. *Mol. Biol. Evol.*, **35**, 159–179.

Stadler,T. (2011) Simulating trees on a fixed number of extant species. *Syst. Biol.*, **60**, 676–684.

Stadler,T. and Degnan,J.H. (2012) A polynomial time algorithm for calculating the probability of a ranked gene tree given a species tree. *Algorithm. Mol. Biol.*, **7**, 338–355.

Veeramah,K.R. *et al.* (2015) Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate Bayesian computation approach. *Genetics*, **200**, 295–308.

Wu,Y. (2012) Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evol. Int. J. Organic Evol.*, **66**, 763–775.

Yang,Z. (2015) The BPP program for species tree estimation and species delimitation. *Curr. Zool.*, **61**, 854–865.

Zhu,S. *et al.* (2015) Hybrid-Lambda: simulation of multiple merger and Kingman gene genealogies in species networks and species trees. *BMC Bioinformatics*, **16**, 292.