Date of Submission: ADD DATE, 2019

Estimation of Speciation Times Under the Multispecies Coalescent

Jing Peng[1*], David Swofford[2], and Laura Kubatko[1,3]

[1]*Department of Statistics, The Ohio State University, Columbus, OH, USA*
[2]*Duke University* [3]*Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH, USA*

[*]*To whom correspondence should be addressed;*

E-mail: peng.650@osu.edu

Author to receive proofs:

Jing Peng

404 Cockins Hall

1958 Neil Ave.

Columbus, OH 43210

FAX: 614-292-2096

E-mail: peng.650@osu.edu

Abtract

Motivation

The coalescent model is now widely accepted as a necessary component for phylogenetic inference from genome-scale data. However, because model-based analysis under the coalescent is computationally prohibitive, a variety of inferential frameworks and corresponding algorithms have been proposed for estimation of species-level phylogenies and the associated parameters, including the speciation times and effective population sizes.

Results

We consider the problem of estimating the timing of speciation events along a phylogeny in a coalescent framework. We propose a pseudolikelihood method for estimation of these speciation times under a model of DNA sequence evolution for which exact site pattern probabilities can be computed. We demonstrate that the pseudolikelihood estimates are statistically consistent and asymptotically normally distributed, and we show how this result can be used to estimate their asymptotic variance. We also provide a more computationally efficient estimator of the asymptotic variance based on the nonparametric bootstrap. We evaluate the performance of our method using simulation and by application to an empirical dataset on gibbons.

Identifying the evolutionary relationships among species is a fundamental problem in evolutionary biology. At the species level, phylogenetic trees are essential to understanding the process of evolution, and thus to the interpretation of biological information in a collection of DNA sequences. As the typical attributes of a species tree, branch lengths, or speciation times, can help to date the evolutionary process of the formation of species. Therefore, estimating an accurate phylogenetic tree with branch lengths is a central problem in species-level phylogenetic inference.

Though numerous methods have been developed recently for species tree estimation, methods for estimating the associated speciation times are less well-developed. One possibility is to obtain maximum likelihood estimates of branch lengths from a concatenated alignment, but use of concatenation for species tree inference has been shown to estimate species trees inconsistently under the multi-species coalescent (e.g., Kubatko and Degnan (2007); Roch and Steel (2015)). Since incomplete lineage sorting is expected to cause incongruence between gene trees and the species tree resulting from the failure of lineages to coalesce in their most recent ancestral population, branch length estimates based on concatenation analyses are expected to be biased upward. Indeed, Kubatko et al. (2011) found that estimates of speciation times based on analysis of concatenated data for *Sistrurus* rattlesnakes differed by up to 70% from those obtained using a coalescent-based framework (see their Table 5).

Among the set of methods that estimate species trees from multi-locus data using estimated gene trees (called "summary statistic methods" or "summary methods"), ASTRAL (Sayyari and Mirarab, 2016) and MP-EST (Liu et al., 2010) also provide estimates of branch lengths. Under the coalescent model, ASTRAL uses the result for quartets of taxa to obtain an estimator based on all subsets of four taxa in the species tree. Specifically, for a given species tree with an arbitrary number of taxa, let Q be a branch and let N be the number of four-taxon subtrees that contain branch Q. Considering only unrooted trees, Sayyari and Mirarab (2016) represented Q as a quadripartition of all of the taxa (i.e., Q divides the taxa

into 4 non-overlapping subsets). The set A of N possible quartets that include branch Q is then built by enumerating every possibility that consists of choosing one leaf per subset in the quadripartition. For a given quartet in the set A, the multi-species coalescent model predicts that the probability of the topology that matches the species tree is $p_1 = 1 - \frac{2}{3}e^{-B}$, and the probabilities of the other two alternative topologies are $p_2 = p_3 = \frac{1}{3}e^{-B}$, where B is the length of branch Q in coalescent units (Nei, 1987). Therefore, for any quartet i in the set A, the frequencies for these three topologies across the input gene trees are modeled as a multinomial random variable $\boldsymbol{Z}_i = (z_{1i}, z_{2i}, z_{3i})$ with parameters $\boldsymbol{P} = (p_1, p_2, p_3)$, where $\sum_{j=1}^{3} z_{ji} = n$ is the number of gene trees. Considering all of the quartets in the set A, $\bar{z}_1$ is defined to be the average number of gene trees that agree with the species tree, i.e.,

$$\bar{z_1} = \frac{1}{N}\sum_{i=1}^{N} z_{1i} \tag{1}$$

Using the result above, a maximum likelihood (ML) estimator of the length of branch Q is given by

$$\hat{B} = -ln\left(\frac{3}{2}(1 - \frac{\bar{z_1}}{n})\right) \tag{2}$$

By applying this to all branches in the unrooted species tree of interest, ASTRAL gives estimates of the lengths of all of the internal branches.

Similarly, MP-EST (Liu et al., 2010) employs a pseudolikelihood approach to estimate the species tree topology and branch lengths. However, instead of using unrooted quartets of taxa as in ASTRAL, MP-EST uses rooted triples to develop an estimator of internal branch lengths based on the pseudolikelihood, defined as the product of marginal probabilities for rooted triplets across all subsets of three taxa in the species tree. Specifically, for a given s-taxon species tree with internal branches labeled $T_1, T_2, \cdots, T_{s-2}$, let the lengths of these branches be denoted $\boldsymbol{\tau} = (\tau_1, \tau_2, \cdots, \tau_{s-2})$, and let $S(\boldsymbol{\theta}, \boldsymbol{\tau})$ be the species tree with topology S, vector of branch lengths $\boldsymbol{\tau}$, and vector of effective population sizes $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_{s-2})$. Let A be the set that includes the $N = \binom{s}{3}$ possible subsets of three of the s taxa. For any three-taxon subset j in A, three different topologies can be observed in the gene trees with

corresponding probabilities predicted by the multi-species coalescent model, i.e., the probability of the topology that matches the species tree is $p_{1j} = 1 - \frac{2}{3}e^{-B_j}$, and the probabilities of the other two alternative topologies are $p_{2j} = p_{3j} = \frac{1}{3}e^{-B_j}$, where B_j is the internal branch length for the true species tree for the taxa in subset j, which is either equal to one of the elements in the vector $\boldsymbol{\tau}$ or is a linear combination of $\tau_1, \tau_2, \ldots, \tau_{s-2}$. As noted above, the multinomial distribution can be used to model the frequencies $\boldsymbol{Z}_j = (z_{1j}, z_{2j}, z_{3j})$ for these three topologies across the input gene trees, and thus the marginal probability for subset j is:

$$f(z_{1j}, z_{2j}, z_{3j}|S(\boldsymbol{\tau}, \boldsymbol{\theta})) = \frac{n!}{z_{1j}!z_{2j}!z_{3j}!}\left(1 - \frac{2}{3}e^{-B_j}\right)^{z_{1j}}\left(\frac{1}{3}e^{-B_j}\right)^{z_{2j}}\left(\frac{1}{3}e^{-B_j}\right)^{z_{3j}} \tag{3}$$

where $n = \sum_{k=1}^{3} z_{kj}$ is the number of genes. The marginal probabilities of triplets are multiplied to obtain the pseudolikelihood:

$$\begin{aligned} PL(\boldsymbol{\tau}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_N) &= \prod_{j=1}^{N} \frac{n!}{z_{1j}!z_{2j}!z_{3j}!}\left(1 - \frac{2}{3}e^{-B_j}\right)^{z_{1j}}\left(\frac{1}{3}e^{-B_j}\right)^{z_{2j}}\left(\frac{1}{3}e^{-B_j}\right)^{z_{3j}} \\ &= \prod_{j=1}^{N} \frac{n!}{z_{1j}!z_{2j}!z_{3j}!} p_{1j}(\boldsymbol{\tau})^{z_{1j}} p_{2j}(\boldsymbol{\tau})^{z_{2j}} p_{3j}(\boldsymbol{\tau})^{z_{3j}} \end{aligned} \tag{4}$$

Obviously, without considering the correlation structure of the triplets that include the same taxa, (4) is not the joint probability distribution of triplets, and thus Liu et al. (2010) called it this the pseudolikelihood. By maximizing (4), MP-EST gives the following estimator of $\boldsymbol{\tau}$:

$$\hat{\boldsymbol{\tau}} = \underset{\boldsymbol{\tau}}{\operatorname{argmax}}\, PL(\boldsymbol{\tau}|\boldsymbol{Z}_1, \boldsymbol{Z}_2, \cdots, \boldsymbol{Z}_N) \tag{5}$$

Though these methods have been claimed to be statistically consistent in branch length estimation (Sayyari and Mirarab, 2016), consistency only holds when an unbiased sample of true gene trees is given. For an empirical dataset, however, the gene trees are estimated from sequence data, and thus statistical consistency can only be achieved by using a statistically consistent method of gene tree estimation and allowing the gene length to go to infinity. In fact, both ASTRAL and MP-EST have been shown to underestimate branch lengths when input gene tree estimation error increases (Sayyari and Mirarab, 2016).

In contrast to summary-statistics methods such as ASTRAL and MP-EST, several approaches have employed a Bayesian framework to jointly estimate the species tree topology, speciation times, and effective population sizes using the full sequence data (i.e., without first estimating gene trees for each locus). Let D represent multi-locus data composed of individual sequence alignments $d_1, d_2, \ldots, d_n$ for a sample of n loci. Then the posterior distribution of the species tree given the data can be written as

$$P(S(\boldsymbol{\theta}, \boldsymbol{\tau})|D) \propto \int_G \left(\prod_{i=1}^{n} P(d_i|(g_i, \boldsymbol{t}_i)) P((g_i, \boldsymbol{t}_i)|S(\boldsymbol{\theta}, \boldsymbol{\tau})) \right) P(S(\boldsymbol{\theta}, \boldsymbol{\tau})) dG$$

where the integral is taken over G, the space of all possible gene trees, including both topologies g_i and branch lengths $\mathbf{t}_i$, for each of the $i = 1, 2, \ldots, n$ loci; $P(d_i|(g_i, \mathbf{t}_i))$ is the Felsenstein likelihood of the ith sequence alignment given a gene tree (Felsenstein, 1981); $P((g_i, \mathbf{t}_i)|S(\boldsymbol{\theta}, \boldsymbol{\tau}))$ is the gene tree density under the multi-species coalescent; and $P(S(\boldsymbol{\theta}, \boldsymbol{\tau}))$ is the prior distribution of the species tree $S(\boldsymbol{\theta}, \boldsymbol{\tau})$. To estimate the posterior distribution, Markov chain Monte Carlo (MCMC) is used, and an estimator of the branch lengths can be obtained from this estimated posterior distribution for a fixed species topology.

The most popular approaches for Bayesian species tree inference are *BEAST/StarBEAST2 (Heled and Drummond, 2010; Ogilvie et al., 2017) and BPP (Yang, 2015) for multi-locus sequence data, and SNAPP for biallelic SNP data (Bryant et al., 2012). StarBEAST2 and BPP differ in the prior distributions assumed for the species tree, and in the range of evolutionary models they employ. Unlike summary-statistics methods, Bayesian methods utilize a full-likelihood framework, which includes all of the information in the sequence data, and thus have an advantage over summary-based methods. However, the estimates of the branch lengths obtained by these methods will depend on the choice of prior distribution, especially when the species are closely related or the sequences are very similar. Moreover, because they rely on MCMC to approximate a posterior distribution, the computation is prohibitive for data sets with a large number of species and/or genes.

A third class of methods for inferring species trees includes the SVDQuartets method (Chifman and Kubatko, 2014), which infers trees directly from the sequence data (i.e., with-

out first inferring gene trees for each locus), but is more computationally efficient than the associated Bayesian methods. The SVDQuartets method does not at present include estimates of the branch lengths. Here we show how the work of Chifman and Kubatko (2015) that underlies the SVDQuartets method can be used to derive an estimator for branch lengths under the multi-species coalescent model (Kingman, 1982a,b; Tavaré, 1984) and the Jukes-Cantor model (JC69) (Jukes and Cantor, 1969), assuming the molecular clock. Our proposed estimator is distinct from the existing estimators described above, in that it uses pseudolikelihood to obtain computationally efficient estimators in a model-based framework. Because it uses the full data directly, rather than estimating gene trees explicitly, but does not require a Bayesian framework (and thus the computational expenses associated with MCMC), the method incorporates all of the variability in the data without losing computational efficiency. We show that this estimator is statistically consistent and asymptotically normally distributed. Though the uncertainty in the estimator can be quantified by the theoretical asymptotic variance predicted by our normality result, we find that use of the nonparametric bootstrap provides a more accurate estimate of the variance of the estimates. The performance and computational cost associated with our method of speciation time estimation is assessed using simulated datasets. A genome-scale dataset for gibbons (Carbone et al., 2014; Shi and Yang, 2018; Veeramah et al., 2015) is utilized to demonstrate the performance of our estimator for empirical data.

Methods

In a 4-taxon species tree, there are $4^4 = 256$ possible site patterns. Chifman and Kubatko (2015) show that each site pattern probability $p_{i_a i_b i_c i_d}$ on a 4-leaf species tree with species a, b, c, and d for a specific observation $i_a i_b i_c i_d$ with $i_j \in \{A, C, G, T\}$ at the tips of the tree can be written as a function of the effective population sizes $\boldsymbol{\theta}$ and the branch lengths $\boldsymbol{\tau}$ in the tree in coalescent units under the JC69 (Jukes and Cantor, 1969) model. Under this

model and the molecular clock assumption, the rooted symmetric 4-leaf species tree ((a, b), (c, d)) will have 9 distinct site patterns probabilities:

$$p_{xxxx}, p_{xxxy} = p_{xxyx}, p_{xyxx} = p_{yxxx}, p_{xyxy} = p_{yxxy}, p_{xxyy}, p_{xyxz} = p_{yxxz} = p_{xyzx} = p_{yxzx},$$
$$p_{xxyz}, p_{yzxx}, p_{xyzw},$$

while the rooted asymmetric 4-leaf species tree (a, (b, (c, d))) will have 11 distinct site patterns probabilities:

$$p_{xxxx}, p_{xxxy} = p_{xxyx}, p_{xyxx}, p_{yxxx}, p_{xyxy} = p_{yxxy}, p_{xxyy}, p_{xyxz} = p_{xyzx}, p_{yxxz} = p_{yxzx},$$
$$p_{xxyz}, p_{yzxx}, p_{xyzw}.$$

where x, y, z and w denote different nucleotide states. For example, p_{xxxx} includes the site patterns $p_{AAAA}, p_{CCCC}, p_{GGGG}$ and p_{TTTT}, which have identical probabilities under the model. As another example, p_{xxxy} includes the site patterns $p_{AAAC}, p_{AAAG}, p_{AAAT}, p_{CCCA}$, etc.

For simplicity, we use $\boldsymbol{p}_s = (p_{s_1}(\boldsymbol{\theta}, \boldsymbol{\tau}), p_{s_2}(\boldsymbol{\theta}, \boldsymbol{\tau}), \ldots, p_{s_9}(\boldsymbol{\theta}, \boldsymbol{\tau}))$ to denote the 9 different site pattern probabilities arising from the symmetric 4-taxon species tree, while $\boldsymbol{p}_a = (p_{a_1}(\boldsymbol{\theta}, \boldsymbol{\tau}), p_{a_2}(\boldsymbol{\theta}, \boldsymbol{\tau}), \ldots, p_{a_{11}}(\boldsymbol{\theta}, \boldsymbol{\tau}))$ are the 11 distinct site pattern probabilities from the asymmetric 4-taxon species tree, where we now use notation to indicate the dependence of each of the site pattern probabilities on the quantities $\boldsymbol{\theta}$ and $\boldsymbol{\tau}$. In an alignment of length M, assuming that the observed sites are independent, the site pattern frequencies for these classes can be modeled as a multinomial random variable, conditional on the species tree:

$$\boldsymbol{Z} \sim Multinomial(M, \boldsymbol{p}_s) \qquad \text{for a symmetric tree;}$$
$$\boldsymbol{Z} \sim Multinomial(M, \boldsymbol{p}_a) \qquad \text{for an asymmetric tree,}$$

where $\boldsymbol{Z}$ is the vector of site pattern frequencies for the 9 or 11 distinct classes.

Based on the results for a 4-leaf species tree, we can split a tree of arbitrary size into quartets, i.e., 4-leaf subtrees, and write the likelihood of the observed site pattern frequencies for each quartet. For the 5-leaf species tree in Figure 1, for example, we can consider all sets of 4 tips, to get $\binom{5}{4} = 5$ different quartets. For a quartet i, each site in an alignment of length M can be classified into one of the 9 or 11 distinct site patterns. For each site m, $m = 1, 2, \ldots, M$ and each quartet i, $i = 1, 2, \ldots, 5$, define $\boldsymbol{V_i^{(m)}}$ to be the random vector of length 9 or 11 that contains a 1 in the jth entry if site pattern j is observed at that site and 0 in all other entries, and let $\boldsymbol{v_i^{(m)}}$ be the corresponding observed data. Let $\boldsymbol{v_i} = (\boldsymbol{v_i^{(1)}}, \ldots, \boldsymbol{v_i^{(M)}})$ denote the observed data across all M sites, and v_{i_j} be the jth entry of the vector $\sum_{m=1}^{M} \boldsymbol{v_i^{(m)}}$, which counts the number of times site pattern j is observed. Then, the likelihood can be expressed as:

$$L_i(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{v}_i) = \frac{M!}{v_{i_1}!v_{i_2}!\cdots v_{i_{11}}!} p_{a_1}(\boldsymbol{\theta}, \boldsymbol{\tau})^{v_{i_1}} p_{a_2}(\boldsymbol{\theta}, \boldsymbol{\tau})^{v_{i_2}} \cdots p_{a_{11}}(\boldsymbol{\theta}, \boldsymbol{\tau})^{v_{i_{11}}} \qquad i = 1, 2; \tag{6}$$

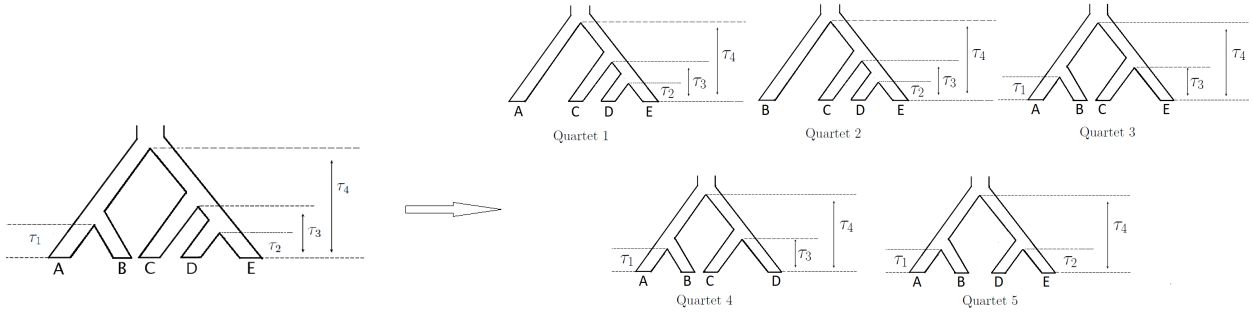
$$L_i(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{v}_i) = \frac{M!}{v_{i_1}!v_{i_2}!\cdots v_{i_9}!} p_{s_1}(\boldsymbol{\theta}, \boldsymbol{\tau})^{v_{i_1}} p_{s_2}(\boldsymbol{\theta}, \boldsymbol{\tau})^{v_{i_2}} \cdots p_{s_9}(\boldsymbol{\theta}, \boldsymbol{\tau})^{v_{i_9}} \qquad i = 3, 4, 5, \tag{7}$$

which are each functions of the population sizes $\boldsymbol{\theta}$ and the branch lengths $\boldsymbol{\tau}$. Clearly, these quartets are not independent, and thus we cannot compute a true likelihood without accounting for the correlation structure among them. Therefore, we use the so-called *pseudolikelihood* – the product of the individual likelihoods for all possible quartets. We note that the pseudolikelihood is also often referred to in the statistical and biology literature as the *composite likelihood* or the *approximate likelihood* (see, e.g., Varin et al. (2011) for a review of the history of composite likelihood methods).

After taking the logarithm, the pseudolikelihood is given by:

$$logPL(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{x}) = \sum_{i=1}^{Q} logL_i(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{v}_i) \tag{8}$$

where Q is the number of possible quartets, and $\boldsymbol{x} = (\boldsymbol{x^{(1)}}, \cdots, \boldsymbol{x^{(M)}})$ is defined similarly to the $\boldsymbol{v_i}$, but for the entire tree. Then each vector $\boldsymbol{x^{(m)}}$ records which of the possible distinct



(a) The 5-leaf species tree

(b) The 5 possible quartets arising from the tree in (a)

Figure 1: The 5-leaf species tree (left) can be split into 5 different 4-leaf subtrees (right), shown with speciation times marked.

site patterns on a tree of 5 tips is observed at site m. Maximizing this pseudolikelihood function gives us the maximum pseudolikelihood estimator (MPLE):

$$(\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\theta}}) = \underset{\boldsymbol{\tau},\boldsymbol{\theta}}{\operatorname{argmax}}\, logPL(\boldsymbol{\tau}, \boldsymbol{\theta}|\boldsymbol{x}) = \underset{\boldsymbol{\tau},\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{i=1}^{Q} logL_i(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{v}_i) \tag{9}$$

Statistical consistency

Arnold and Strauss (1991) consider the case of r independent random variables $\boldsymbol{X} = (X^{(1)}, \cdots, X^{(r)})$, each with joint density $f(\boldsymbol{x}; \boldsymbol{\beta})$ that undergo t transformations:

$$Z_i^{(j)} = g_i(X^{(j)}), \;\; i = 1, 2, \cdots, t; j = 1, 2, \cdots, r.$$

Rather than maximize the likelihood, Arnold and Strauss consider maximizing the following product of conditional densities of $\boldsymbol{Z_i}$'s:

$$\prod_{j=1}^{r} \left[\prod_{k} \prod_{k'} f^{\alpha_{kk'}}(\boldsymbol{z_k^{(j)}} | \boldsymbol{z_{k'}^{(j)}}; \boldsymbol{\beta}) \right] \tag{10}$$

where $\alpha_{kk'} > 0$. If we further define random variables $\boldsymbol{Y_i}$'s, where each observation $\boldsymbol{y_i}$ is of form $\boldsymbol{z_k}$ or $(\boldsymbol{z_k}, \boldsymbol{z_l})$, the ratios of unconditional densities of random variables $\boldsymbol{Y_i}$'s can be used to express the objective function (10). After taking the logarithm, the pseudolikelihood of the data is defined by:

$$logPL(\boldsymbol{\delta}, \boldsymbol{\beta}) = \sum_{i} \delta_i \sum_{j=1}^{r} log(f(\boldsymbol{y_i^{(j)}}; \boldsymbol{\beta})), \tag{11}$$

where δ_i is chosen to ensure that this pseudolikelihood arises from a product of positive likelihoods and conditional likelihoods. To show statistical consistency in our setting, we first recall the following theorem by Arnold and Strauss (1991).

Theorem (Arnold and Strauss, 1991)**.** *Let $X^{(1)}, \ldots, X^{(r)}$ be independent and identically distributed, and suppose that the $Y_i^{(j)'}s$ are such that their densities $f(\boldsymbol{y_i}; \boldsymbol{\beta})$ are differentiable with respect to $\boldsymbol{\beta}$ for almost every $\boldsymbol{y_i}$. Then with probability tending to 1 as $r \to \infty$, the score equation of the log pseudolikelihood has a root $\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$ such that $\tilde{\boldsymbol{\beta}}_{\boldsymbol{r}}$ converges in probability to the true parameter value.*

In our case, random variable X represents the site pattern counts from the entire tree, so $\boldsymbol{x^{(1)}}, \ldots, \boldsymbol{x^{(M)}}$ are the observations for each of the M sites. If we define Z_k as:

$$\boldsymbol{Z}_k = \begin{cases} 1 & \text{if } k = 1; \\ \boldsymbol{V}_{k-1} & \text{if } k = 2, \cdots, 6, \end{cases}$$

where $\boldsymbol{V}_{k-1} = (\boldsymbol{V}_{k-1}^1, \ldots, \boldsymbol{V}_{k-1}^M)$ is the collection of random vectors of site pattern indicators at each site as defined previously and $\alpha_{kk'} = 1$, the objective function (10) can be written as:

$$\prod_{j=1}^{M} \left[\prod_{k=2}^{6} f(\boldsymbol{z_k^{(j)}} | \boldsymbol{z_1^{(j)}}; \boldsymbol{\beta}) \right] = \prod_{j=1}^{M} \left[\prod_{k=1}^{5} f(\boldsymbol{v_k^{(j)}}) \right],$$

where $\boldsymbol{v_k^{(1)}}, \cdots, \boldsymbol{v_k^{(M)}}$ are the observations for the M sites from quartet k, defined as above. From Chifman and Kubatko (2015), the marginal density of the site pattern frequencies for the 4-leaf subtree can be expressed as (6) or (7). If we further define random variable $\boldsymbol{Y_i}$, $i = 1, 2, \ldots, Q$, where each observation $\boldsymbol{y_i^{(j)}}$ is of the form $\boldsymbol{v_k^{(j)}}$, then the logarithm of the pseudolikelihood is of the form (8). Using the theorem above, statistical consistency of the pseudolikelihood estimates of the speciation times is shown. Furthermore, the next theorem by Arnold and Strauss (1991) states that the pseudolikelihood estimator is asymptotically normal and gives an expression for its asymptotic variance.

Theorem (Arnold and Strauss, 1991)**.** *Suppose that $X^{(1)}, \cdots, X^{(r)}$ are independent and identically distributed. Assume that the standard regularity conditions hold. Then any con-*

sistent sequence $\tilde{\beta}_r = \beta_r(X^{(1)}, \cdots, X^{(r)})$ *of roots of the score equation of the log pseudolikelihood satisfies:*

$$\sqrt{r}(\tilde{\beta}_r - \beta) \to N\left(0, \frac{K(\beta)}{J^2(\beta)}\right)$$

where

$$K(\beta) = \sum_{i,i'} \mathrm{E}_\beta\left[\left\{\frac{\partial}{\partial\beta} logf(y_i;\beta)\right\}\left\{\frac{\partial}{\partial\beta} logf(y_{i'};\beta)\right\}\right]$$

and

$$J(\beta) = -\sum_i \mathrm{E}_\beta\left\{\frac{\partial^2}{\partial\beta^2} logf(y_i;\beta)\right\}$$

This theorem provides a way of quantifying the uncertainty of our estimator, by calculating the first and second derivatives of the log likelihoods for the individual quartets. For example, to compute the asymptotic variance of the MPLE $\tilde{\tau}_1$ for the 5-taxon species tree in Figure 1, note that quartets 3, 4, and 5 include branch length τ_1. Therefore, if we define $\boldsymbol{\tau}$ and $\boldsymbol{\theta}$ to be the vectors of the branch lengths and population sizes, respectively, and the relevant MPLEs are $\tilde{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{\theta}}$, then we can approximate $K(\tau_1)$ and $J(\tau_1)$ by $K^*(\tau_1)$ and $J^*(\tau_1)$, where

$$K^*(\tau_1) = \sum_{i,i'}\left[\left\{\frac{\partial}{\partial\tau_1} logL_i(\boldsymbol{\tau},\boldsymbol{\theta}|\boldsymbol{z_i})\Big|_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}\right\}\left\{\frac{\partial}{\partial\tau_1} logL_{i'}(\boldsymbol{\tau},\boldsymbol{\theta}|\boldsymbol{z_{i'}})\Big|_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}\right\}\right] \quad i = 3,4,5; i' = 3,4,5.$$

and

$$J^*(\tau_1) = \sum_i\left\{\frac{\partial^2}{\partial\tau_1^2} logL_i(\boldsymbol{\tau},\boldsymbol{\theta}|\boldsymbol{z_i})\Big|_{\tilde{\boldsymbol{\tau}},\tilde{\boldsymbol{\theta}}}\right\} \quad i = 3,4,5.$$

Then the asymptotic variance is calculated by

$$Var^*(\hat{\tau}_1) = \frac{K^*(\tau_1)}{J^*(\tau_1)^2}. \tag{12}$$

To calculate K^* and J^*, we plug in $\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\theta}}$ to estimate the 9 or 11 site pattern probabilities (rounding to 2 significant figures, which makes the variance estimates more stable).

In addition to the asymptotic variance, we can also use a bootstrapping approach to estimate the variance of the MPLE. In this approach, a bootstrap replicate is obtained by resampling the columns, i.e., site patterns in the original DNA sequences, by the following steps:

1. With replacement, one column is randomly selected from the original sequence alignment to be a new column in the resampled data. Repeating this M times, we get a new bootstrap replicate of the same size as the original data;

2. Repeat step 1 to get B bootstrap samples;

3. For each of the B of the bootstrap samples, re-do the analysis to compute the estimates $(\tilde{\boldsymbol{\theta}}_1, \tilde{\boldsymbol{\tau}}_1), \cdots, (\tilde{\boldsymbol{\theta}}_B, \tilde{\boldsymbol{\tau}}_B)$, and calculate the sample variance of the estimates, $\mathrm{Var}(\tilde{\boldsymbol{\theta}}_B)$ and $\mathrm{Var}(\tilde{\boldsymbol{\tau}}_B)$.

Simulation study

We first use simulation to assess the statistical consistency and asymptotic normality of the MPLE, and we compare the two methods of measuring uncertainty – calculation of the theoretical asymptotic variance and bootstrapping. We note that while many methods for inferring species-level phylogenies are based on multi-locus data, our method is designed for unlinked sites arising from the coalescent model, a data type that we call Coalescent Independent Sites (CIS). While multi-locus data is assumed to have evolved from the same genealogy within a locus, we assume the site patterns in the sequences are independent draws from the distribution characterized by the multi-species coalescence and the nucleotide substitution model (Chifman and Kubatko, 2014), conditional on the species tree. However, a straightforward argument can be made that methods developed for CIS data can also be applied to multi-locus data (see, e.g., Wascher and Kubatko (2019)) and thus we consider both data types here.

We wish to examine the properties of the pseudolikelihood estimates for both unlinked CIS data and multi-locus data, and thus we simulated two types of data: unlinked CIS data (each site has its own gene tree) and multi-locus data (a sequence of length l is simulated for each locus with the same underlying gene tree). The simulation is done by the following

procedures:

1. Use the program COAL (Degnan and Salter, 2005) to generate gene tree samples under the multi-species coalescent model based on a specified input species tree;

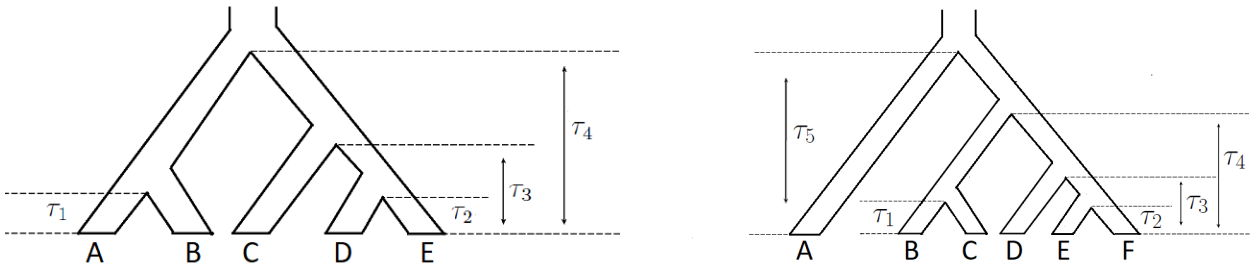
2. Use the program Seq-Gen (Rambaut and Grassly, 1997) to generate DNA sequences of length l for each gene tree under the JC69 model;

3. Compute the site pattern frequencies for all possible quartets, and compute the log pseudolikelihood. Maximize the log pseudolikelihood using the SANN algorithm (Bélisle, 1992), which is implemented in the R optim() function (R Core Team, 2018) to get the branch length estimates and estimate their theoretical asymptotic variances;

4. Resample the simulated sequences to get B bootstrap replicates, and compute the sample variance of the estimates from the bootstrap samples;

5. Repeat step 1–4 D times to obtain branch length estimates, as well as the associated theoretical asymptotic variance estimates and bootstrap variance estimates.

Given the above simulation algorithm, we can specify different simulation settings. In step 1, we set up two different model species trees: a 5-leaf tree and a 6-leaf tree (Figure 2). Here we measure time in coalescent units, which are the number of generations scaled by $2N$, where N is the effective population size. For simplicity, we fix the population size $\boldsymbol{\theta}$ to be the same for all the populations, which is $\theta = 4N\mu = 0.02$, where μ is the mutation rate. For speciation times, we use three schemes (0.5, 1.0, and 2.0) for all of the external and internal branch lengths on the two model trees, respectively. In step 2, the gene length l is chosen to be 1 in simulating unlinked CIS data (i.e., we simulate 100,000 genes and 1 site for each), while in the multi-locus data setting, we consider 200 genes, each of length l=500. Furthermore, we choose JC69 (Jukes and Cantor, 1969) to be the DNA substitution model. Specifically, we use the Seq-Gen options " -q -mHKY -s 0.01" to simulate under JC69. In steps 4 and 5, we choose B=100 and D=100, respectively.



(a) The 5-leaf model species tree

(b) The 6-leaf model species tree

Figure 2: Two different model species trees with speciation times as parameters for simulation process

Moreover, we argue that the above properties of the MPLE hold when multiple lineages per tip of the species tree are sampled. Therefore, we also carry out a simulation study for a 5-leaf model species tree with multiple lineages at the tips. Instead of COAL (Degnan and Salter, 2005) in step 1, ms (Hudson, 2002) is utilized to generate gene tree samples using the same 5-leaf model species tree and parameter settings as above, but with 2 lineages for species D and E. Following the simulation algorithm, we then redo the analysis for unlinked SNP data to get the branch lengths and their variance estimates.

Application to gibbon data

We also explore the performance of our MPLE in inferring speciation times for empirical data. We show the results of applying our method to a genome-scale dataset consisting of 5 species of gibbons : *Hylobates moloch* (Hm), *Hylobates pileatus* (Hp), *Nomascus leucogenys* (N), *Hoolock leuconedys* (B), and *Symphalangus syndactylus* (S). This dataset was generated by Carbone et al. (2014) and Veeramah et al. (2015), and consists of 11,323 coding loci, each 200 bp in length. Except for the outgroup (O), multiple lineages are included for each species: 2 for Hm, 2 for Hp, 4 for N, 4 for B, and 4 for S. This dataset has been previously analyzed by Shi and Yang (2018) and the branch lengths were estimated based on BPP and ASTRAL; details about the model assumptions can be found there.

Here, note that we base our analysis on the 6-leaf species tree topology in Figure 3. We enumerate all possible quartets by selecting one lineage per tip for this tree, and calculate the pseudolikelihood estimates for internal branch lengths ($\tau_{SB}, \tau_{NSB}, \tau_{HpHm}$) and population sizes ($\theta$s). For simplicity, we assume the population sizes for all of the groups are the same, and compare the internal branch length estimates with the results from BPP and ASTRAL without this assumption. As we did in the simulation study, 100 bootstrap samples will be used to measure the uncertainty of our estimates.

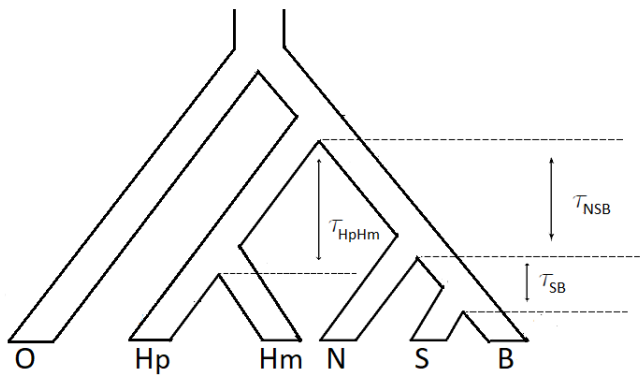


Figure 3: The species tree for 5 gibbon species and 1 outgroup (human)

Furthermore, to assess the reliability of the estimator, we use one of the simulated data sets generated by Shi and Yang (2018). This dataset is based on the same species tree in Figure 3 and the parameters (θs and τs) are the estimates under the multi-species coalescent model and JC69, which can be found in their paper. With the same lineage sampling scheme as in the real data, they simulate 10,000 loci each of 200 bp for this coding dataset. We estimate the internal branch lengths and compare with the true values used to generate the data. We also compare the performance of our method with BPP and ASTRAL.

Results

Simulation study

We plot histograms of the 100 pseudolikelihood estimates for branch lengths in the six 5-taxon and 6-taxon model trees (see supplementary files for figures). Figure 4 shows histograms of the 100 pseudolikelihood estimates of branch length τ_1 for the three 5-leaf model trees under our simulation conditions. From those plots, we see that the estimates are approximately normal and distributed around the true value. Thus it appears that our estimates are unbiased. Moreover, when we include multiple lineages per tip or analyze multi-locus data in the same way, the unbiasedness and asymptotic normality still hold (see supplemental files). If we increase the number of sites, we see these results even more clearly (see supplemental files).

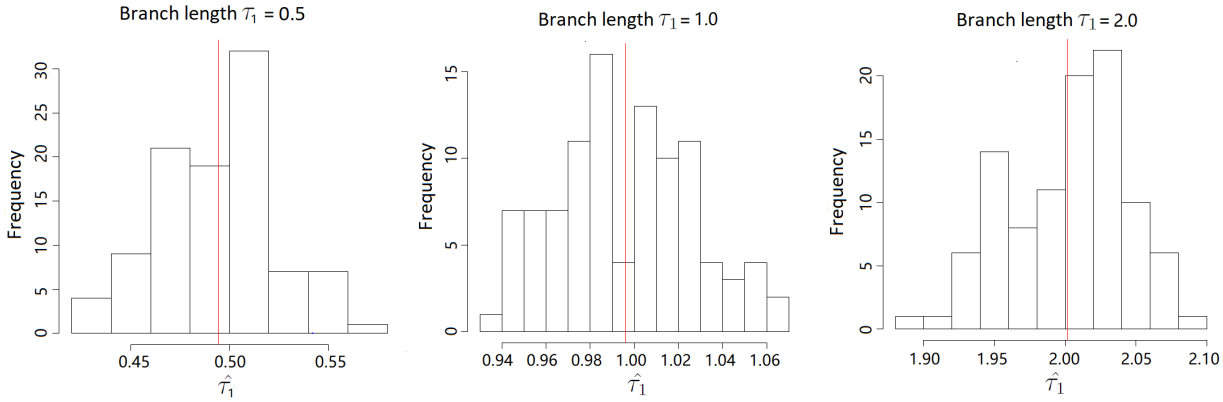
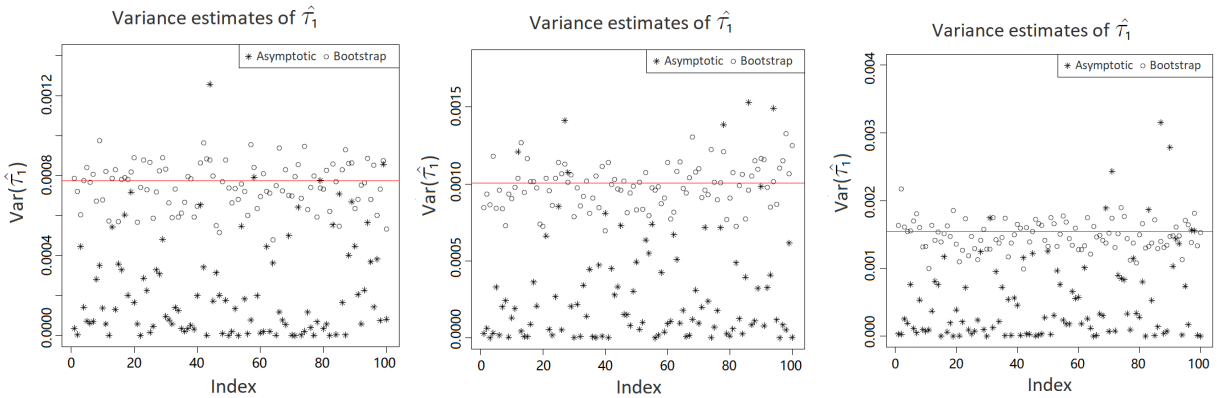


Figure 4: Histogram of the 100 pseudolikelihood estimates for branch length τ_1 using 100,000 unlinked CIS from the 5-leaf model trees with a single lineage per tip. The red line in each histogram is the sample mean of the 100 pseudolikelihood estimates. The true values are given in the figure titles.

To assess the performance of our method in estimating the uncertainty of the MPLE, Figure 5 shows plots of the 100 variance estimates of the pseudolikelihood estimates of branch length τ_1 for the three 5-leaf model trees under our simulation conditions. In the unlinked-DNA sequences, single-lineage-per-tip setting, it is immediately clear that in all



(a) model tree with $\tau_1 = 0.5$ (b) model tree with $\tau_1 = 1.0$ (c) model tree with $\tau_1 = 2.0$

Figure 5: Plots of the 100 variance estimates for branch length τ_1 using 100,000 unlinked CIS from the 5-leaf model trees with a single lineage per tip. Points denoted by $*$ are computed by the asymptotic variance formula in (12), while points denoted by o are obtained by bootstrapping. The x-axis is an index for the simulated samples. The red line in the plots is the sample variance of the 1000 pseudolikelihood estimates.

cases the bootstrap estimates are scattered around the sample variance from the 100 simulated datasets. This approximation is expected to be better when the number of sites increases. The asymptotic variance estimates calculated by formula (12), however, tend to underestimate the variance and to be unstable. We discuss this further in the Discussion, but note here that Varin et al. (2011) also pointed out that the bootstrap sometimes performs better in practice than the asymptotic variance estimate. Thus, the bootstrap estimator is recommended to measure the variance of the MPLE, though the asymptotic variance is theoretically reasonable. The results are similar when we use multi-locus data or include multiple lineages per tip in the 5-taxon and 6-taxon model trees (see supplementary files for details).

Application to gibbon data

Table 1 shows the estimates of the internal branch lengths in Figure 3 from ASTRAL, BPP and MPLE. We can see that the pseudolikelihood estimates based on the subset data (which

consist of randomly selecting one lineage per tip) are close to the estimates from the BPP analysis, though for the Hp-Hm branch, our estimate is smaller than the BPP and ASTRAL estimates. We also notice that our 95% confidence interval for Hp-Hm is wider than BPP's, indicating substantial uncertainty in our point estimate. We note that when more lineages are included, the MPLE estimates are not as close to the others from BPP or ASTRAL. This may be due to model misspecification, since our current method may be biased when the JC69 assumption is violated. On the other hand, we have assumed that the population sizes (θs) are the same for all ancestral and extant groups, which may be not true in this case.

Table 1: ASTRAL, BPP and MPLE estimates of internal branch lengths in coalescent units in the species tree (Figure 3) from the real coding data set. In the MPLE (subset) analysis, we randomly sample one lineage per tip and apply our method on the subset dataset. The MPLE estimates are simply computed based on the full real dataset. The 95% confidence intervals are calculated by the bootstrap method. We also estimate the population size $\theta = 0.00108$ with 95% confidence interval (0.00105, 0.00111). The JC69 model is assumed in all analyses.

	SB	SBN	Hp-Hm
ASTRAL (JC)	0.044	0.073	2.136
BPP	0.072	0.075	2.051
	(0.057,0.088)	(0.056,0.095)	(1.969,2.137)
MPLE (subset)	0.073	0.076	1.859
	(0.052, 0.095)	(0.064,0.088)	(1.538,2.180)
MPLE	0.060	0.108	1.849
	(0.029, 0.089)	(0.078,0.140)	(1.805,1.951)

Table 2: ASTRAL, BPP and MPLE estimates of internal branch lengths in coalescent units in the species tree in Figure 3 from the simulated coding data set. The dataset is simulated under JC69, and JC69 is assumed in all analysis. We also estimate the population size $\theta = 0.00099$ with 95% confidence interval (0.00096, 0.00103).

	SB	SBN	Hp-Hm
Truth	0.071	0.074	2.050
ASTRAL (JC)	0.083	0.084	2.295
BPP	0.109	0.099	2.062
	(0.089,0.128)	(0.073,0.126)	(1.974,2.155)
MPLE	0.097	0.104	2.090
	(0.062,0.128)	(0.075,0.151)	(2.024,2.184)

To remove the effect of model misspecification and check the performance of our estimator in the case where the true parameter values are known, we use the simulated coding dataset under the JC69 model. We can see that MPLE estimates are similar to the BPP estimates, but with wider confidence intervals, in general. The estimate for Hp-Hm is reasonable, but the estimates for SB and SBN are large. We argue that the assumption of the same population size in all groups may still cause this problem. We also note that BPP and ASTRAL also tend to overestimate speciation times for SB and Hp-Hm, respectively. Clearly, this is a challenging estimation problem for all of the methods. However, we note that for all three of the branch lengths, the true value is included (or very nearly included, in the case of SBN) in the 95% confidence interval obtained by MPLE, while the true value of the SB branch length is not included in the 95% confidence interval obtained using BPP.

Discussion

The stability of the asymptotic variance

We note that in Figure 5 the calculated asymptotic variance from formula (12) is unstable, and this is not expected to improve by increasing the number of sites. As indicated in formula (12), derivatives need to be computed to obtain the asymptotic variance, and the first derivative of (6) and (7) can be expressed as:

$$\frac{\partial log L_i(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{z}_i)}{\partial \boldsymbol{\tau}} = \sum_{j=1}^{11} \left[\frac{z_{ij}}{p_{aj}(\boldsymbol{\theta}, \boldsymbol{\tau})} \frac{\partial p_{aj}(\boldsymbol{\theta}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \right] \qquad i = 1, 2;$$

$$\frac{\partial log L_i(\boldsymbol{\theta}, \boldsymbol{\tau}|\boldsymbol{z}_i)}{\partial \boldsymbol{\tau}} = \sum_{j=1}^{9} \left[\frac{z_{ij}}{p_{sj}(\boldsymbol{\theta}, \boldsymbol{\tau})} \frac{\partial p_{sj}(\boldsymbol{\theta}, \boldsymbol{\tau})}{\partial \boldsymbol{\tau}} \right] \qquad i = 3, 4, 5.$$

Notice that in the formulas above, we have terms that include the reciprocal of site pattern probabilities, which are estimated by plugging in the MPLE estimates. Since most of these probabilities are very small (10^{-5}), even with estimation error on the order of 10^{-6}, the reciprocals vary by about 10^4. In fact, to get a good estimate, we need the relative error of the site pattern probabilities to be smaller than 0.001, but it is only around 0.01 in our case, even when the number of sites is very large. Therefore, though the asymptotic variance estimator is theoretically valid, it performs poorly in practice. We thus recommend using the bootstrap estimates to quantify the uncertainty of our estimator. This estimator is unbiased under the model assumptions, and it is computationally efficient.

The assumptions and performance of MPLE

Here we have assumed the JC69 model, which allows us to use the formulas of Chifman and Kubatko (2015) to compute the 9 or 11 distinct site pattern probabilities. Empirical data, however, may evolve under a nucleotide substitution model more complex than JC69, and preliminary simulations indicate that our method may not be robust when the nucleotide

substitution model is misspecificed (results not shown). However, for closely related species like gibbons, Shi and Yang (2018) argue that the JC69 model should be adequate, and we note that BPP also assumes the JC69 model. Thus, the differences in the estimates obtained by BPP and MPLE are not likely to be due to misspecification of the nucleotide substitution model. A second assumption in our empirical analysis is that the effective population sizes are equal for all populations in the species tree. However, the results shown in Table **??** indicate that our method may be sensitive to this assumption, which may explain some of the differences in our estimates in comparison to those obtained by BPP.

Therefore, we are still investigating plausible methods to extend our method to allow inference under more general models, such as the GTR model and its submodels. One idea is to generate Monte Carlo estimates of the site pattern probabilities by simulating the evolution of a large number of independent sites under the desired substitution model and the coalescent process with relevant parameters. The pseudolikelihood can then be maximized across the set of possible parameters to find the MPLE. However, for species trees of large size, the optimization may require evaluation of hundreds to thousands of points in the parameter space, with Monte Carlo sampling undertaken for each point under consideration, and thus this approach would be computationally expensive.

Though our method is at present limited to the JC69 model, the MPLE is more computationally efficient than estimators based in a Bayesian framework, such as BPP (which is also limited to the JC69 model). In comparison to ASTRAL and MP-EST, the MPLE does not rely on unbiased gene tree estimates, and we can quantify the uncertainty in the estimates that arises from all sources. Finally, we note that the guarantee of asymptotic normality ensures that the MPLE will have good statistical properties as dataset size grows in the number of sites.

Generally, to obtain the pseudolikelihood estimates of the speciation times in a species tree, we need to be able to do two things: compute the pseudolikelihood, and search the parameter space to find the values that optimize the pseudolikelihood. We argue that the first step can be done very efficiently for trees of arbitrary size. In Equation (8), we can see that the number of individual likelihoods for all possible quartets will grow with the size of the tree, but the summation is not very computationally expensive, even for a large tree. The second step, however, is more so, because we must find the global maximizer by optimizing the pseudolikelihood with respect to all of the parameters simultaneously. Thus, as the tree grows, we must evaluate thousands of points in our parameter space to find the optimal estimate, which may become expensive for large trees. We note, however, that this problem is faced by all methods that employ probabilistic methods for phylogenetic inference, with existing methods for inferring species-level phylogenies directly from multiple sequence alignments (e.g., BPP (Yang, 2015), *BEAST (Heled and Drummond, 2010), StarBEAST2 (Ogilvie et al., 2017)) known to be extremely computationally intensive. Thus, the MPLE is a useful addition to the collection available for inferring speciation times from genome-scale data under the coalescent model.

References

Arnold, B. C. and D. Strauss. 1991. Pseudolikelihood Estimation: Some Examples. Sankhyā: The Indian Journal of Statistics, Series B 53:233–243.

Bélisle, C. J. 1992. Convergence theorems for a class of simulated annealing algorithms on R^d. Journal of Applied Probability 29:885–895.

Bryant, D., R. Bouckaert, J. Felsenstein, N. Rosenberg, and A. RoyChoudhury. 2012. Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. Mol. Biol. Evol. 29:1917–1932.

Carbone, L., R. A. Harris, S. Gnerre, K. R. Veeramah, B. Lorente-Galdos, J. Huddleston, T. J. Meyer, J. Herrero, C. Roos, and B. Aken et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. Nature 513:195–201.

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

Chifman, J. and L. Kubatko. 2015. Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, site-specific rate variation, and invariable sites. Journal of Theoretical Biology 374:35–47.

Degnan, J. and L. Salter. 2005. Gene tree distributions under the coalescent process. Evolution 59:24–37.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–76.

Heled, J. and A. J. Drummond. 2010. Bayesian inference of species trees from multilocus data. Molecular Biology and Evolution 27:570–580.

Hudson, R. R. 2002. Generating samples under a Wright–Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Jukes, T. and C. R. Cantor. 1969. Evolution of protein molecules. Pages 21–123 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Kingman, J. F. C. 1982a. On the genealogy of large populations. J. Appl. Prob. 19A:27–43.

Kingman, J. F. C. 1982b. The coalescent. Stoch. Proc. Appl. 13:235–248.

Kubatko, L. and J. Degnan. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology 56:17–24.

Kubatko, L. S., H. L. Gibbs, and E. Bloomquist. 2011. Inferring species-level phylogenies using multi-locus data for a recent radiation of Sistrurus rattlesnakes. Syst. Biol. 60:393–409.

Liu, L., L. Yu, and S. V. Edwards. 2010. A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. BMC Evolutionary Biology 10:302.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press.

Ogilvie, H. A., R. R. Bouckaert, and A. J. Drummond. 2017. StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. Molecular Biology and Evolution 34:2101–2114.

R Core Team. 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria.

Rambaut, A. and N. Grassly. 1997. SeqGen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Computer Applications in Biosciences 13:235–238.

Roch, S. and M. Steel. 2015. Likelihood-based tree reconstruction on a concatenation of aligned sequence data sets can be statistically inconsistent. Theoretical Population Biology 100:56–62.

Sayyari, E. and S. Mirarab. 2016. Fast Coalescent-Based Computation of Local Branch Support from Quartet Frequencies. Molecular Biology and Evolution 33:1654–1668.

Shi, C.-M. and Z. Yang. 2018. Coalescent-Based Analyses of Genomic Sequence Data Provide a Robust Resolution of Phylogenetic Relationships among Major Groups of Gibbons. Molecular Biology and Evolution 35:159–179.

Tavaré, S. 1984. Line-of-descent and genealogical processes, and their applications in population genetics models. Theor. Popul. Biol. 26:119–164.

Varin, C., N. Reid, and D. Firth. 2011. An overview of composite likelihood methods. Statistica Sinica 21:5–42.

Veeramah, K. R., A. E. Woerner, L. Johnstone, I. Gut, M. Gut, T. Marques-Bonet, L. Carbone, J. D. Wall, and M. F. Hammer. 2015. Examining phylogenetic relationships among gibbon genera using whole genome sequence data using an approximate bayesian computation approach. Genetics 200:295–308.

Wascher, M. and L. Kubatko. 2019. Consistency of svdquartets and maximum likelihood for coalescent-based species tree estimation. Under review, available at https://doi.org/10.1101/523050 .

Yang, Z. 2015. The BPP program for species tree estimation and species delimitation. Current Zoology 61:854–865.