Date of Submission: December 21, 2018

Consistency of SVDQuartets and Maximum Likelihood for Coalescent-based Species Tree Estimation

Matthew Wascher[1*] and Laura Kubatko[1,2]

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

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

E-mail: wascher.1@osu.edu

Author to receive proofs:

Matthew Wascher

1958 Neil Ave.

Columbus, OH 43210

FAX: 614-292-3648

E-mail: wascher.1@osu.edu

Abtract– Numerous methods for inferring species-level phylogenies under the coalescent model have been proposed within the last 20 years, and debates continue about the relative strengths and weaknesses of these methods. The approaches taken by these methods can be classified into three groups. Summary methods carry out inference for multilocus data by first computing a summary statistic for each locus, and then forming an estimate of the species phylogeny using these single-locus summaries as input. Such methods are often criticized for failing to account for uncertainty in the estimates computed for each locus in the second stage of the analysis. The second class of methods have been referred to as "full data methods" because they utilize the sequence data directly by carrying out estimation in a Bayesian framework, with specific implementations designed to handle either multilocus or SNP data. Because such methods require both computation of the likelihood and Markov chain Monte Carlo for inference, they are limited in the size of the data sets that can be analyzed. The third class of methods uses the information available in the observed site pattern frequencies to estimate the species phylogeny by inferring the phylogenetic relationships among quartets of taxa and then using quartet assembly methods to construct the overall species tree estimate. Although designed for coalescent independent sites data, methods based on site pattern frequencies can be easily modified to accommodate both SNP and multilocus data. Recently, Shi and Yang (2018) hypothesized that site pattern-based methods are less efficient than likelihood-based methods. Here, we show that the site pattern-based method SVDQuartets is statistically consistent and has the same order of convergence as maximum likelihood, which we verify is also statistically consistent. We provide theoretical justification for the use of SVDQuartets with multilocus data, and discuss the impact of the number of sites and number of loci considered in multilocus studies. Finally, we compare the performance of maximum likelihood and SDVQuartets using simulation, and show that although both have the same order of convergence, under certain conditions either may show better performance.

Advances in sequencing technology over the last 20 years have led to widespread availability of large-scale sequence data sets from multiple loci for which the goal is to obtain an estimate of the species-level phylogenetic relationships among the taxa under consideration. Analysis of such data has presented significant computational challenges, however, because inference methods must include models that capture variation at two distinct scales. First, a model for the process by which the phylogenetic histories of individual loci vary given the overall species tree must be developed. The coalescent process (Kingman, 1982b,c,a; Tavare, 1986) is usually used for this purpose. Second, the mutation process arising along the locus-specific phylogenies, typically called gene trees, must be modeled. This is usually accomplished using standard nucleotide substitution models (Goldman, 1993). Together, these two model components are often referred to as the multispecies coalescent (MSC).

Inference of the species phylogeny under the MSC is challenging because the gene trees are not directly observed, and must therefore be integrated over when computing probabilities associated with the DNA sequence data (see Equation (1) below). Current algorithms for species tree inference handle this complication in one of three different ways. First, summary methods (Liu et al., 2009) break the inference problem into two distinct steps. In the first step, gene trees are estimated for each locus, and these gene trees are used as input to the second step of the algorithm which estimates the species tree using information contained in the gene trees. Several of these methods are fairly widely used at present (e.g., ASTRAL (Mirarab et al., 2014), MP-EST (Liu et al., 2010)), though they are sometimes criticized for failing to account for uncertainty in the gene tree estimates when they are used as input to the second step. The second group of methods for estimation are often referred to as "full data methods" because they operate directly on the sequence data, rather than requiring a summary to be computed for each locus. These methods employ a Bayesian framework for inference and thus use Markov chain Monte Carlo (MCMC) to approximate the posterior probability distribution of the species tree. Algorithms have been developed for both multilocus data (e.g., StarBEAST2 (Heled and Drummond, 2010; Ogilvie et al., 2017),

BEST (Liu and Pearl, 2007), BPP (Yang, 2015)) and for biallelic SNP data (e.g., SNAPP (Bryant et al., 2012)), and these methods are widely used. However, an important limitation of these methods is that they are computationally intensive, and can be prohibitively so when the data are large, either in terms of the number of taxa or the number of loci.

The final class of methods for species tree inference includes those based on site pattern probabilities, such as SVDQuartets (Chifman and Kubatko, 2014). These methods are appealing because although they are fully model-based, they avoid explicit computation of the site pattern probabilities and instead rely on assessing the fit of the observed site pattern probabilities (which are easy to compute) to expected relationships among the true site pattern probabilities predicted by the model. Shi and Yang (2018) hypothesized that SVDQuartets is less statistically efficient than maximum likelihood. Here we show that in fact both SVDQuartets and maximum likelihood are statistically consistent methods for estimating species trees under the model, and that they have the same order of convergence. We first consider coalescent independent sites data, and then discuss multilocus data. In the case of coalescent independent sites, simulations will be used to compare their performance over a range of conditions. We begin by defining notation.

Consider a species tree with M species labeled $1, 2, \ldots, M$, and suppose that m_j individuals are sampled within each species j. Thus, $\mathcal{M} = \sum_{j=1}^{M} m_j$ is the total number of sequences in the data set. Using the framework of the MSC, we denote the probability density of gene tree history h and associated vector of coalescent times $\mathbf{t}_h$, conditional on species tree topology S and vector of speciation times τ, by $f_{(h,\mathbf{t}_h)|(S,\tau)}$ (see Rannala and Yang (2003) for a description of how to compute this density). We further define a site pattern to be an assignment of states $i_1 i_2 \cdots i_{\mathcal{M}}$ to the $\mathcal{M}$ tips of the tree, such that $i_k \in \{A, C, G, T\}$ for $k = 1, 2, \ldots, \mathcal{M}$, and we denote the probability of site pattern $p^h = i_1 i_2 \cdots i_{\mathcal{M}}$ arising from *gene tree history* $(h, \mathbf{t}_h)$ by $p^h_{(i_1 i_2 \cdots i_{\mathcal{M}})|(h,\mathbf{t}_h)}$. This probability is the usual phylogenetic likelihood along a gene tree, computed assuming one of the standard nucleotide substitution models. The probability of observing site pattern $p = i_1 i_2 \cdots i_{\mathcal{M}}$ from the *species tree* is then

given by

$$p_{i_1 i_2 \cdots i_{\mathcal{M}}|(S,\tau)} = \sum_{h \in \mathcal{H}} \int_{\mathbf{t}_h} p^h_{(i_1 i_2 \cdots i_n)|(h,\mathbf{t}_h)} f_{(h,\mathbf{t}_h)|(S,\tau)} d\mathbf{t}_h \tag{1}$$

where the sum is taken over all gene tree histories $\mathcal{H}$ with corresponding branch lengths $\mathbf{t}_h$ appropriately integrated out. See Chifman and Kubatko (2015) for full details of the calculations.

Note that Equation (1) implies that each site in the sequence alignment is an independent observation from the model; that is, each site represents a draw from the distribution of gene trees given the species tree as specified by the MSC, with subsequent mutation along the sampled gene tree according to one of the standard nucleotide substitution models. We refer to such data as *coalescent independent sites (CIS)* to distinguish them from SNP data, which do not usually include invariable sites. Under this model, a sample of N CIS can be viewed as a sample from the multinomial distribution, where the number of categories is the number of possible sites patterns, $4^{\mathcal{M}}$, and the category probabilities are given by the site pattern probabilities. Thus, assuming that the sites are independent conditional on the species tree, the log likelihood of species tree (S, τ) is given by

$$\mathcal{L}\Big((S,\tau)\Big) = \sum_{q=1}^{4^{\mathcal{M}}} x_q \log(p_q) \tag{2}$$

where x_q is the observed number of sites with pattern q, p_q is the probability of site pattern q under the model, $q = 1, 2, \ldots, 4^{\mathcal{M}}$, and $\sum_{q=1}^{4^{\mathcal{M}}} x_q = N$.

To study the convergence properties of SVDQuartets and of maximum likelihood (ML), we consider the case in which $M = 4$ and $m_j = 1$ for all j, that is, we consider four-taxon trees with one sequence sampled in each species. In this case, there are $4^4 = 256$ possible site patterns, 15 rooted species trees, and 3 unrooted species trees. To find the ML estimate of the species tree, one needs to be able to compute the true site pattern probabilities for each possible species tree. Formulas for these site pattern probabilities were given by Chifman and Kubatko (2015) for simple substitution models (e.g., JC69 (Jukes and Cantor, 1969)). Under the JC69 model and using these formulas, we can find the ML estimate of the species

tree by considering each of the 15 rooted species trees and finding the set of speciation times that maximize the likelihood for each. The tree with the highest likelihood is the ML estimate.

To obtain an estimate of the four-taxon species tree for SVDQuartets, let L denote the set of four taxa under consideration, and suppose that L is partitioned into two sets, L_1 and L_2, such that $|L_1| = |L_2| = 2$. We say that $L_1|L_2$ is a *split*. The split $L_1|L_2$ is *valid* for tree S if the subtrees containing the taxa in L_1 and in L_2 do not intersect; otherwise the split is not valid. For example, consider the tree $((1,2),(3,4))$. The split $12|34$ is valid, while the splits $13|24$ and $14|23$ are not valid.

For each of the three possible splits, the 256 possible site patterns can be arranged into a 16×16 matrix in which the rows of the matrix correspond to possible states for the taxa in L_1 and the columns correspond to possible states for the taxa in L_2. Such a matrix is called a *flattening matrix*, and is denoted $Flat_{L_1|L_2}$. For an empirical data set, the entries of the matrix are the observed frequencies of the site pattern that corresponds to the row and column indices, i.e.,

$$Flat_{12|34} = \begin{pmatrix} & [AA] & [AC] & [AG] & [AT] & [CA] & \cdots & [TT] \\ [AA] & p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots & p_{AATT} \\ [AC] & p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots & p_{ACTT} \\ [AG] & p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots & p_{AGTT} \\ [AT] & p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots & p_{ATTT} \\ [CA] & p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots & p_{CATT} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ [TT] & p_{TTAA} & p_{TTAC} & p_{TTAG} & p_{TTAT} & p_{TTCA} & \cdots & p_{TTTT} \end{pmatrix}$$

For example, the $(3,2)$ entry, p_{AGAC}, is the probability of observing nucleotide A for taxon 1, G for taxon 2, A for taxon 3, and C for taxon 4. When the rows and columns of the matrix correspond to a valid split, the matrix will have rank 10 for data observed perfectly from the model. When the rows and columns correspond to a split that is not valid, the matrix will be rank 16. The SVDQuartets method constructs three matrices (one for each

of the three possible splits for four taxa), and computes the *SVD score* for each matrix,

$$SVD(L_1|L_2) = \sqrt{\sum_{k=11}^{16} \hat{\sigma}_k^2} \tag{3}$$

where $\hat{\sigma}_k$ is the kth singular value computed for the matrix of observed site pattern frequencies. For observed data, the magnitudes of the 11th through 16th singular values are expected to be small when the matrix corresponds to the valid split, and thus the split $L_1|L_2$ with the lowest $SVD(L_1|L_2)$ is selected. Note that in the case of four taxa, identifying the valid split is equivalent to inferring the unrooted species tree.

Under either criterion for estimation, we denote the estimator of the species tree by S^* and the true species tree by S. Intuitively, consistency means that as more data are used to form the species tree estimate, the probability that $S^* = S$ goes to 1. SVDQuartets has been assumed to be statistically consistent, but a formal proof has not been provided. ML is known to be consistent when used to estimate gene trees, but consistency of ML has not been formally examined in the species tree case, since ML cannot be used for species tree inference for more than four taxa in that case. In the next section, we show that both methods are statistically consistent, and compare their performance using simulation.

Coalescent Independent Sites

Consistency and Order of Convergence for Maximum Likelihood

While the literature contains numerous proofs of consistency of ML for estimation of gene trees, no such proofs have been given for the case of ML estimation of the species tree, in part because it is not computationally feasible to use ML for species tree estimation under the coalescent model for trees of arbitrary size. Some recent attention has also been given to evaluating the consistency of methods other than ML for estimating the species tree, but such work has focused primarily on the case in which multilocus data are collected and summary

statistics methods are used to form estimators (Roch et al., 2018) or on the concatenation method (Roch and Steel, 2015). In this section, we establish that ML estimation of the species tree under the multispecies coalescent described above is statistically consistent for four-taxon trees. We follow the proof of Truszkowski and Goldman (2016) for the case of gene trees, as most of their proof generalizes directly to the species tree case and their proof corrects the omissions of earlier proofs. We refer the reader to Truszkowski and Goldman (2016) for many of the details.

Early proofs of the consistency of ML estimation for gene trees were given by Yang (1994) and Rogers (1997), but more recent examinations by RoyChoudhury et al. (2015) and Truszkowski and Goldman (2016) have found these proofs are incomplete. RoyChoudhury et al. (2015) explains the problems with these proofs succinctly; we outline his argument here. First, note that the following proposition results from a straightforward application of the Strong Law of Large Numbers.

Proposition 0.1. Suppose T_0 is the true tree and T_j is any other tree. Then there exists N such that for all $n \geq N$

$$L(T_0) > L(T_j) \;\; a.s. \tag{4}$$

However, Proposition (0.1) is not sufficient to conclude that ML estimation is consistent. To see this, consider the typical definition of consistency of the maximum likelihood estimator (MLE) that states that if $\hat{T}$ is the MLE then

$$\hat{T} \xrightarrow{P} T_0 \tag{5}$$

under a metric $D(\cdot, \cdot)$, where $\xrightarrow{p}$ denotes convergence in probability. In order to guarantee that (5) holds, we either need to show that for any $\epsilon > 0$ there exists some constant $C_\epsilon > 0$ such that

$$\sup_{T_j : D(T_j, T_0) > \epsilon} \{L(T_0) - L(T_j)\} \geq C_\epsilon, \tag{6}$$

so that we are assured there cannot be trees of arbitrarily high likelihood far away from the true tree, or that the parameter space is compact. Under any reasonable metric, it is easy to

see that the parameter space is not compact because it does not include trees with branches of length 0. Truszkowski and Goldman (2016) provide a corrected proof by defining the following metric and showing that (6) holds for this metric (see Lemma 3 of Truszkowski and Goldman (2016)).

Definition 0.2 (Distance between two trees)**.** For two taxa a and b in tree S, define their distance, $d_S(a,b)$, to be the sum of the lengths of all edges on the path from a to b. Further, define the distance between two trees S_1 and S_2 to be $D(S_1,S_2) = \max_{a,b\in L} |d_{S_1}(a,b) - d_{S_2}(a,b)|$. Note that $D(\cdot,\cdot)$ is a metric as long as all branch lengths are positive.

We now state and prove a modified version of the consistency result of Truszkowski and Goldman (2016) for a sample of CIS obtained under the multispecies coalescent.

Theorem 0.3 (Consistency of the ML estimator of the species tree)**.** *Let S_N^* denote the MLE of species tree S for a sample of N CIS obtained under the multispecies coalescent. Then $D(S_N^*, S) \to 0$ with probability 1 as $N \to \infty$.*

Proof. Because the site pattern counts for a random sample of N CIS follow a multinomial distribution with probabilities given in Equation (1) above, the likelihood function for the ML estimate of the species tree is similar in form to that in the case of a gene tree. Thus Proposition (0.1) and most steps in the consistency proof given by Truszkowski and Goldman (2016) can be verified in a straightforward manner. The only non-trivial condition to be verified in the species tree case is that Lemma 3 of Truszkowski and Goldman (2016) still holds for the particular site pattern probabilities that arise in the species tree setting. This lemma involves some conditions on the pairwise site pattern probabilities, which we define below.

Following the notation of Truszkowski and Goldman (2016), let f_{xy}^{ab} denote the frequency with which taxon a is observed to have state x and taxon b is observed to have state y, where $x, y \in \{A, C, G, T\}$. Let $p_{xy}^{d'}$ denote the probability that taxon a has state x and taxon b has state y, where $x, y \in \{A, C, G, T\}$, when $d_S(a,b) = d'$. To verify Lemma 3 of Truszkowski

and Goldman (2016) it is sufficient to verify that the function

$$\sum_{x,y\in\{A,C,G,T\}} f_{xy}^{ab} \log(p_{xy}^{d'}) \tag{7}$$

is concave in d'. Under the JC69 model (Jukes and Cantor, 1969) and the multspecies coalescent model, Chifman and Kubatko (2015) (see their Supplement A) gave explicit formulas for the site patterns probabilities on four-taxon trees. Using these with $\mu = 4/3$ as specified by the JC69 model and θ, the effective population size parameter, set to 0.01, we can sum over pairs of taxa to find

$$p_{xy}^{d'} = \begin{cases} \frac{1}{4} + \frac{225}{304}e^{-4d'/3}, & x = y \\ \frac{3}{4} - \frac{225}{304}e^{-4d'/3}, & x \neq y \end{cases} \tag{8}$$

Ruskino (2018, personal communication) has derived more general expressions as a function the θ parameter that follow the same general form. Using these expressions, it is straightforward to verify that the expression in Equation (7) is concave in d' and thus that Lemma 3 of Truszkowski and Goldman (2016) holds. This establishes Theorem 0.3.

□

Consistency and Order of Convergence for SVDQuartets

Recall that for the SVDQuartets method, we choose the tree with valid split $\text{argmin}_{L_1|L_2} SVD(L_1|L_2)$ as our estimate of the species tree. In this section, we prove that this estimator is consistent in the following sense and give its rate of convergence.

Theorem 0.4 (Consistency)**.** *Suppose that the conditions of the model proposed by Chifman and Kubatko (2015) are satisfied, and $L_1|L_2^*$ is the true valid split among splits with $|L_1| = |L_2| = 2$. Fix $\epsilon > 0$. Then $\exists N$ such that $\forall n \geq N$, $\mathbb{P}\left(argmin_{L_1|L_2} SVD(L_1|L_2) \neq L_1|L_2^*\right) < \epsilon$.*

We give the details of the proof of Theorem 0.4 in the remainder of this section. The result follows from consistency of the $\hat{p}_{ij}$ in the flattening matrix and the fact that singular values of a matrix satisfy a Lipschitz condition with respect to perturbations of the matrix (see Golub and Van Loan (2013)).

Lemma 0.5 (Corollary 8.6.2 of Golub and Van Loan (2013))**.** *Let $A, E \in \mathbb{R}^{m\times n}$ with $m \geq n$, and let σ_i, $i \in \{1, \dots n\}$ denote the singular values in descending order. Then for $i \in \{1, \dots n\}$, $|\sigma_i(A+E) - \sigma_i(A)| \leq ||E||_2 = \sigma_1(E)$.*

We first establish the following.

Lemma 0.6. *For any split $L_1|L_2$, $SVD(L_1|L_2) \xrightarrow{p} \sqrt{\sum_{i=11}^{16} \sigma_i^2}$ where σ_i are the descending ordered singular values of the 16×16 matrix $Flat_{L_1|L_2}(P)$.*

Proof. Because $SVD(L_1|L_2)$ is a continuous function of the vector $(\sigma_1, \dots, \sigma_{16})$, it suffices to show that $\hat{\sigma}_i \xrightarrow{p} \sigma_i$ uniformly in i. Fix $\epsilon, \delta > 0$. We will show $\exists N$ such that $\forall n \geq N$, $\mathbb{P}\left(\sup_i |\hat{\sigma}_i - \sigma_i| > \delta\right) < \epsilon$.

First note that for the entries $\{\hat{p}_{ij}\}$ of $Flat_{L_1|L_2}(\hat{P})$, the central limit theorem implies that $e_{ij} := \hat{p}_{ij} - p_{ij} \sim N(0, \tau^2)$ for large N, where $\tau^2 \leq \frac{1}{4n}$. Let $X \sim N(0, \frac{1}{4n})$ and choose N large so that, when $n \geq N$, $\mathbb{P}\left(|X| > \frac{\delta}{64}\right) < \frac{\epsilon}{512}$, and for all (i,j), $\mathbb{P}\left(|e_{ij}| > \frac{\delta}{64}\right) - \mathbb{P}\left(|X| > \frac{\delta}{64}\right) < \frac{\epsilon}{512}$. Then using the union bound,

$$\begin{aligned}
\mathbb{P}\left(\max_{i,j}|e_{ij}| > \frac{\delta}{64}\right) &\leq \sum_{i,j} \mathbb{P}\left(|e_{ij}| > \frac{\delta}{64}\right) \\
&= \sum_{i,j} \left\{ \mathbb{P}\left(|X| > \frac{\delta}{64}\right) + \left[\mathbb{P}\left(|e_{ij}| > \frac{\delta}{64}\right) - \mathbb{P}\left(|X| > \frac{\delta}{64}\right)\right] \right\} \\
&< \epsilon
\end{aligned}$$

Now choose E in Lemma 1 to be $E = \{e_{ij}\}$. Then $\sup_i |\hat{\sigma}_i - \sigma_i| \leq ||E||_2$. It is well-known that for any matrix $E \in \mathbb{R}^{k\times k}$, $||E||_2 \leq \sqrt{k}||E||_1 \leq k\sqrt{k}|\max_{i,j}(e_{ij})|$. Applying this fact with $k = 16$, we have $\mathbb{P}\left(\sup_i |\hat{\sigma}_i - \sigma_i| > \delta\right) < \epsilon$, as desired. □

This also provides the rate of convergence.

Corollary 0.7. $|SVD(L_1|L_2) - \sqrt{\sum_{i=11}^{16} \sigma_i^2}| = O_p(n^{-1/2})$

Proof. Lemma 0.6 establishes that $|\hat{\sigma}_i - \sigma_i| = O_p(\max_{i,j}|e_{ij}|)$ for all $i = 1, \dots, 16$. The central limit theorem gives that $|e_{ij}| = O_p(n^{-1/2})$, and so $\max_{i,j}|e_{ij}| \leq \sum_{i,j} |e_{i,j}| = O_p(n^{-1/2})$. It follows that $|SVD(L_1|L_2) - \sqrt{\sum_{i=11}^{16} \sigma_i^2}| = O_p(|\hat{\sigma}_i - \sigma_i|) = O_p(n^{-1/2})$. □

We can now prove Theorem 0.4:

Proof. Theorem 1 of Chifman and Kubatko (2015) implies that $\sqrt{\sum_{i=11}^{16} \sigma_i^2} = 0$ if and only if we choose the split $L_1|L_2^*$. Then because we have finitely many (3) splits to choose from, we can find some $c > 0$ such that $\sqrt{\sum_{i=11}^{16} \sigma_i^2} > c$ for any split $L_1|L_2 \neq L_1|L_2^*$. Fix $\epsilon > 0$. Choose $\epsilon^* = \frac{\epsilon}{3}$ and $\delta = \frac{c}{2}$ in Lemma 0.6. Then for the N that satisfies Lemma 0.6 using ϵ^* and δ, for $n \geq N$, we will have $\mathbb{P}\left(SVD(L_1|L_2^*) > c/2\right) < \epsilon/3$ and $\mathbb{P}\left(SVD(L_1|L_2) < c/2\right) < \epsilon/3$ for every $L_1|L_2 \neq L_1|L_2^*$. Then, using the union bound,

$$\begin{aligned} & \mathbb{P}\left(\text{argmin}_{L_1|L_2} SVD(L_1|L_2) \neq L_1|L_2^*\right) \\ \leq \; & \mathbb{P}\left(SVD(L_1|L_2^*) > c/2\right) + \sum_{L_1|L_2 \neq L_1|L_2^*} \mathbb{P}\left(SVD(L_1|L_2) < c/2\right) \\ < \; & \epsilon \end{aligned}$$

which completes the proof. □

Note that this proof depends on the structure of a four-taxon species tree only insofar as Theorem 1 of Chifman and Kubatko (2015) has only been proven for trees of four taxa. Should that result be extended to trees with a larger number of taxa, our arguments above imply that the estimator based on the SVD score in such cases would also be consistent and have rate of convergence $n^{-1/2}$.

Comparison via Simulation

We conducted a simulation study to comparatively evaluate the performance of ML and SVDQuartets. CIS data were simulated along the four-taxon symmetric and asymmetric species trees by first simulating gene trees using the package COAL (Degnan and Salter, 2005) and then simulating sequence data under the JC69 model (Jukes and Cantor, 1969) using Seq-Gen (Rambaut and Grassly, 1997). The JC69 model was used because Chifman and Kubatko (2015) provided explicit formulas for the site pattern probabilities for four-taxon trees under the coalescent for this model, allowing us to implement the maximum

likelihood method in this case. We considered species tree branch lengths of 0.5, 1.0, and 2.0 in coalescent units with the effective population size parameter $\theta = 4N\mu$ set at 0.005, 0.02, 0.1, and 0.2. We examined the proportion of times out of 100 replicates that each of the methods correctly estimated the unrooted species tree when the total number of sites ranged from 200 to 20,000. Because both methods were less accurate for larger sample sizes when $\theta = 0.2$ and branch lengths were 2.0 (a setting that corresponds to extremely high mutation rates), we extended the number of CIS simulated to 100,000 in this case.

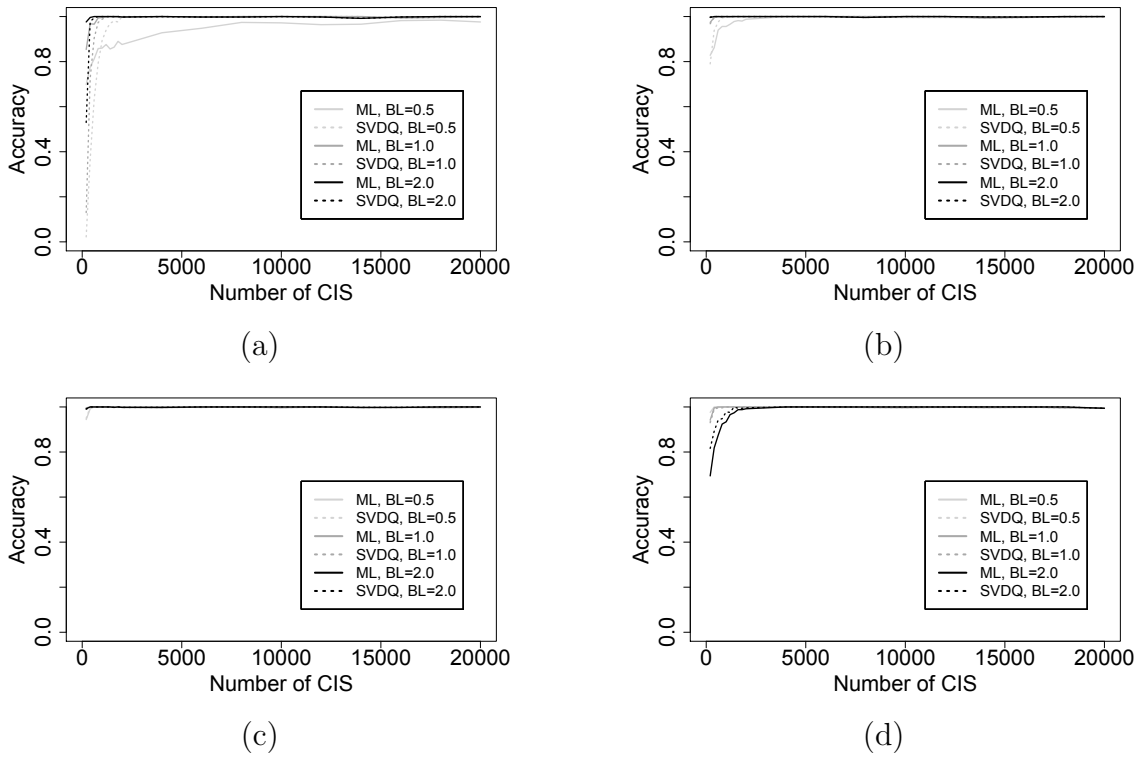

Figure 1: Results of the simulation study for the symmetric species tree. The x-axis shows the number of CIS, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method. (a) $\theta = 0.005$; (b) $\theta = 0.02$; (c) $\theta = 0.1$; (d) $\theta = 0.2$.

Figure 1 shows the results of the simulations for the symmetric species tree, and Figure 2 shows the results for the asymmetric species tree. In general, both methods are able to accurately infer the unrooted four-taxon species trees once the amount of data is sufficient. When the model species tree is symmetric (Figure 1), both methods have nearly 100% accuracy once 5,000 or more CIS are sampled and $\theta \geq 0.02$. However, both methods perform

more poorly when data are simulated from the asymmetric tree (Figure 2). For example, when $\theta = 0.02$ and branch lengths are 0.5, more than 10,000 CIS are needed for the accuracy of ML to reach accuracy above 95%, and ML is even less accurate when $\theta = 0.005$. However, SVDQuartets shows very slow improvement in accuracy with increased sampled size when $\theta = 0.2$ and branch length were 2.0, with only 93.4% accuracy when 100,000 CIS are sampled (we note that these settings correspond to a very high rate of mutation). Comparing across all simulation conditions, we observe that under certain conditions one method may perform better than the other, but overall, both methods are very accurate and appear to show similar rates of convergence.

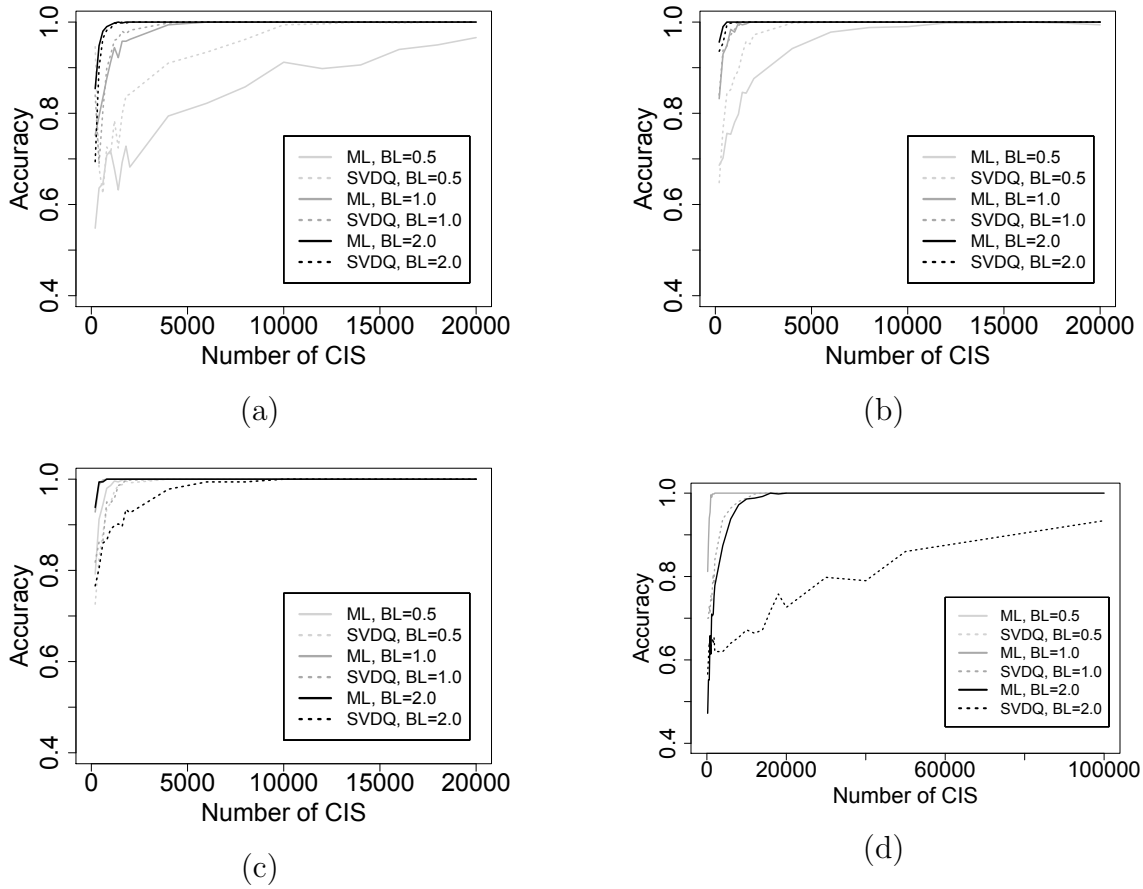

Figure 2: Results of the simulation study for the asymmetric species tree. The x-axis shows the number of CIS, and the y-axis shows the proportion of correctly-estimated unrooted species trees for each method. (a) $\theta = 0.005$; (b) $\theta = 0.02$; (c) $\theta = 0.1$; (d) $\theta = 0.2$. Note that the x-axis has been extended in subfigure (d).

Multilocus Data

SVDQuartets was originally formulated for SNP data. However, in many cases, multilocus data have been sequenced and are already available. While one could, of course, simply discard data at random and use only one nucleotide per gene, we show that SVDQuartets can be adapted for multilocus data in such a way as to gain some statistical benefit from the additional data. In this section, we formalize a model that we assume generates multilocus data, and we consider using SNPs as well as three reasonable estimators based on the full multilocus data to estimate the vector of site pattern probabilities, $\mathbf{p}$. Recall that in our proof of the consistency of SVDQuartets, we showed that $|SVD(L_1|L_2) - \sqrt{\sum_{k=11}^{16} \sigma_k^2}|$ satisfies a Lipschitz condition in $\sup_i |p_i - \hat{p}_i|$. Thus, it suffices to consider the bias and variance of our estimator $\hat{\mathbf{p}}$ to assess the performance of SVDQuartets.

We assume that multilocus data are generated from the following statistical model:

1. Population and genome sizes are large enough that the fact that genes and sites are sampled without replacement can be ignored.

2. Define $\mathbf{p}$ to be a vector of multinomial probabilities such that if we select a gene at random and sample one nucleotide at random from that gene, the unconditional site pattern distribution $\mathbf{X} \sim Multinomial(1, \mathbf{p})$.

3. Define $\{\mathbf{D}\}_{i=1}^{N} \overset{iid}{\sim} F$ such that $E(\mathbf{D_i}) = \mathbf{0}$ and if we select N genes at random, each of the N genes will have multinomial site pattern probabilities $\mathbf{p_i}$ where $\mathbf{p_i} \overset{d}{=} \mathbf{p} + \mathbf{D_i}$.

4. Conditional on $\{\mathbf{D}\}_{i=1}^{N}$, if $\mathbf{X_i}$ are the observed site pattern counts for a sample of n_i nucleotides from gene i, then $\mathbf{X_i} \sim Multinomial(n_i, \mathbf{p_i})$ and the collection $\{\mathbf{X_i}\}_{i=1}^{N}$ is independent.

Based on varying sampling schemes included in this model, four reasonable estimators of $\mathbf{p}$ are:

1. If sampling N genes with one nucleotide per gene: $\hat{\mathbf{p}}_1 = \mathbf{X}/N$.

2. If sampling N genes with n nucleotides per gene: $\hat{\mathbf{p}}_2 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i/n$

3. If sampling N genes with n_i nucleotides per gene: $\hat{\mathbf{p}}_3 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i/n_i$.

4. If sampling N genes with n_i nucleotides per gene: $\hat{\mathbf{p}}_4 = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\mathbf{X}_i$.

Note that we assume that the number of sites sampled per locus are fixed constants (i.e., not random variables). We briefly discuss the case in which the number of sites sampled per locus is random in the Discussion.

In the next section we consider the bias and variance for each of these estimators. We show that for general multilocus data, method 3 should be preferred, and we characterize the statistical benefit of using multilocus data versus SNPs.

Modified Estimators and Consistency

We first consider sampling N genes, 1 nucleotide per gene, and using $\hat{\mathbf{p}}_1 = \mathbf{X}/N$, which is the SNP setting. In this case, based on assumptions 1-2, if $\mathbf{X}$ gives the counts of site patterns in our sample of N genes, then $\mathbf{X} \sim Multinomial(N, \mathbf{p})$. In this case using the standard estimator $\hat{\mathbf{p}}_1 = \mathbf{X}/N$, we can see that

1. $E(\hat{\mathbf{p}}_1) = \mathbf{p}$.

2. $Var(\hat{\mathbf{p}}_1) = \frac{1}{N^2}Var(\mathbf{X})$.

3. If X_i are the elements of $\mathbf{X}$ and $p_{1,i}$ the elements of $\mathbf{p}$, then $Var(\hat{p}_{1,i}) = \frac{p_{1,i}(1-p_{1,i})}{N} \leq \frac{1}{4N}$ for all i.

We next consider sampling N genes with n nucleotides per gene, and using $\hat{\mathbf{p}}_2 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i/n$, which is the case in which we have multilocus data with the same number of sites sampled in each gene. In this case, based on assumptions 1-4, if $\mathbf{X}_i$ are the site pattern counts for

the sample of n nucleotides from gene i, then $\mathbf{X}_i|\mathbf{D}_i \sim Multinomial(n, \mathbf{p}+\mathbf{D}_i)$, and we can estimate $\mathbf{p}$ by $\hat{\mathbf{p}}_2 = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{X}_i}{n}$. Then

$$\begin{aligned} E(\hat{\mathbf{p}}_2) &= E(E(\hat{\mathbf{p}}_2|\{\mathbf{D}\}_{i=1}^N)) \\ &= \frac{1}{N}\sum_{i=1}^{N} E(\frac{n(\mathbf{p}+\mathbf{D}_i)}{n}) \\ &= \mathbf{p} \end{aligned}$$

so this estimator is unbiased. The variance is given by

$$Var(\hat{\mathbf{p}}_2) = Var(E(\hat{\mathbf{p}}_2|\{\mathbf{D}\}_{i=1}^N) + E(Var(\hat{\mathbf{p}}_2|\{\mathbf{D}\}_{i=1}^N))$$

For the first term

$$Var(E(\hat{\mathbf{p}}_2|\{\mathbf{D}\}_{i=1}^N) = Var(\frac{1}{N}\sum_{i=1}^{N}\frac{n(\mathbf{p}+\mathbf{D_i})}{n}) = \frac{1}{N}Var(\mathbf{D_1})$$

For the second term

$$E(Var(\hat{\mathbf{p}}_2|\{\mathbf{D}\}_{i=1}^N)) = \frac{1}{N^2}\frac{1}{n^2}\sum_{i=1}^{N} E(\mathbf{\Sigma_i})$$

where $\mathbf{\Sigma_i} = Var(X_i|\{\mathbf{D}\}_{i=1}^N)$. Thus,

$$Var(\hat{\mathbf{p}}_2) = \frac{1}{N}Var(\mathbf{D_1}) + \frac{1}{N^2}\frac{1}{n^2}\sum_{i=1}^{N} E(\mathbf{\Sigma_i}),$$

and so, in this case, the variance of $\hat{\mathbf{p}}_2$ can be decomposed into two parts, and sampling multiple nucleotides from each gene is only able to reduce the latter part. We note that the diagonal entries of the variance-covariance matrix, i.e., $Var(\hat{p}_{2,i})$, are of interest here, and that the magnitude of these values is determined by a tradeoff between the size of the diagonal entries of $Var(\mathbf{D_1})$ and $E(\mathbf{\Sigma_i})$. Putting more of the mass of the distribution of $\mathbf{D}_1$ toward the edges of the parameter space increases the diagonal entries of $Var(\mathbf{D_1})$ but also means you are more likely to draw extreme $\mathbf{p_i}$'s, which reduce the diagonal entries of $E(\mathbf{\Sigma_i})$. Generally, the reduction in $Var(\hat{\mathbf{p}}_2)$ from sampling n nucleotides per gene instead of one depends on the distribution of $\mathbf{D}_1$.

We next consider sampling N genes with n_i nucleotides per gene, and using $\hat{\mathbf{p}}_3 = \frac{1}{N}\sum_{i=1}^{N}\mathbf{X}_i/n_i$. This corresponds to the realistic situation in which multilocus data with varying lengths for each locus are sampled, and our estimator of the site pattern probabilities is formed by averaging the estimates obtained for each of the individual loci. In this case, based on assumptions 1-4, if $\mathbf{X}_i$ are the site pattern counts for the sample of n_i nucleotides from gene i, then $\mathbf{X}_i|\mathbf{D}_i \sim Multinomial(n_i, \mathbf{p}+\mathbf{D}_i)$, and we can estimate $\mathbf{p}$ by $\hat{\mathbf{p}}_3 = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathbf{X}_i}{n_i}$. Then

$$\begin{aligned} E(\hat{\mathbf{p}}_3) &= E(E(\hat{\mathbf{p}}|\{\mathbf{D}\}_{i=1}^{N})) \\ &= \frac{1}{N}\sum_{i=1}^{N} E(\frac{n_i(\mathbf{p}+\mathbf{D}_i)}{n_i}) \\ &= \mathbf{p} \end{aligned}$$

so this estimator is also unbiased. The variance is given by

$$Var(\hat{\mathbf{p}}_3) = Var(E(\hat{\mathbf{p}}_3|\{\mathbf{D}\}_{i=1}^{N}) + E(Var(\hat{\mathbf{p}}_3|\{\mathbf{D}\}_{i=1}^{N}))$$

For the first term

$$Var(E(\hat{\mathbf{p}}_3|\{\mathbf{D}\}_{i=1}^{N}) = Var(\frac{1}{N}\sum_{i=1}^{N}\frac{n_i(\mathbf{p}+\mathbf{D_i})}{n_i}) = \frac{1}{N}Var(\mathbf{D_1})$$

For the second term

$$E(Var(\hat{\mathbf{p}}_3|\{\mathbf{D}\}_{i=1}^{N})) = \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{n_i^2}E(\mathbf{\Sigma_i})$$

where $\mathbf{\Sigma_i} = Var(X_i|\{\mathbf{D}\}_{i=1}^{N})$. Thus,

$$Var(\hat{\mathbf{p}}_3) = \frac{1}{N}Var(\mathbf{D_1}) + \frac{1}{N^2}\sum_{i=1}^{N}\frac{1}{n_i^2}E(\mathbf{\Sigma_i}),$$

and this case is largely analogous to the previous case. If we had some prior information about which $\mathbf{p_i}$ were closer to $\frac{1}{2}$ it would be possible to use that information to choose sample sizes to draw more from genes with higher internal nucleotide entropy, but in practice this is not the case.

Finally, we consider sampling N genes with n_i nucleotides per gene, and using $\hat{\mathbf{p}}_4 = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N}\mathbf{X}_i$. This corresponds to the situation in which multilocus data with varying

lengths for each locus are sampled, and our estimator of the site pattern probabilities is formed by computing the overall frequency with which each site pattern is observed. In this case, based on assumptions 1-4, if $\mathbf{X}_i$ are the site pattern counts for the sample of n_i nucleotides from gene i, then $\mathbf{X}_i|\mathbf{D}_i \sim Multinomial(n_i, \mathbf{p}+\mathbf{D}_i)$, and we estimate $\mathbf{p}$ by $\hat{\mathbf{p}}_4 = \frac{1}{\sum_{i=1}^N n_i}\sum_{i=1}^N \mathbf{X}_i$. Then

$$E(\hat{\mathbf{p}}_4) = E(E(\hat{\mathbf{p}}_4|\{\mathbf{D}\}_{i=1}^N))$$
$$= \frac{1}{\sum_{i=1}^N n_i}\sum_{i=1}^N E(n_i(\mathbf{p}+\mathbf{D_i})) = \mathbf{p}$$

so this sampling scheme is still unbiased under the assumptions we have made. The variance is given by

$$Var(\hat{\mathbf{p}}_4) = Var(E(\hat{\mathbf{p}}_4|\{\mathbf{D}\}_{i=1}^N) + E(Var(\hat{\mathbf{p}}_4|\{\mathbf{D}\}_{i=1}^N))$$

For the first term

$$Var(E(\hat{\mathbf{p}}_4|\{\mathbf{D}\}_{i=1}^N) = Var(\frac{1}{\sum_{i=1}^N n_i}\sum_{i=1}^N n_i(\mathbf{p}+\mathbf{D_i})) = \frac{\sum_{i=1}^N n_i^2}{(\sum_{i=1}^N n_i)^2}Var(\mathbf{D_1})$$

For the second term

$$E(Var(\hat{\mathbf{p}}_4|\{\mathbf{D}\}_{i=1}^N)) = \frac{1}{N^2}\sum_{i=1}^N \frac{1}{n_i^2}E(\mathbf{\Sigma_i})$$

where $\mathbf{\Sigma_i} = Var(X_i|\{\mathbf{D}\}_{i=1}^N)$. Thus,

$$Var(\hat{\mathbf{p}}_4) = \frac{\sum_{i=1}^N n_i^2}{(\sum_{i=1}^N n_i)^2}Var(\mathbf{D_1}) + \frac{1}{N^2}\sum_{i=1}^N \frac{1}{n_i^2}E(\mathbf{\Sigma_i})$$

Since $\frac{\sum_{i=1}^N n_i^2}{(\sum_{i=1}^N n_i)^2} \geq \frac{1}{N}$ with strict equality only when the n_i are all equal, so $Var(\hat{\mathbf{p}}_4) \geq Var(\hat{\mathbf{p}}_3)$. Thus, we recommend $\hat{\mathbf{p}}_3$ be used rather than $\hat{\mathbf{p}}_4$.

Discussion

Our work gives the first consistency results for four-taxon species tree inference under the coalescent model for both maximum likelihood and SVDQuartets. Previous consistency results for maximum likelihood were only derived in the case of gene trees. In addition,

since both the maximum likelihood estimator and the SVDQuartets estimator converge at rate $O(n^{-1/2})$, the asymptotic relative efficiency of the two is a constant. This constant probably depends on the structure of the tree being estimated, but our simulations show that it does not appear to be particularly large in a variety of reasonable scenarios. Thus, even if the MLE is asymptotically efficient, as is often the case, SVDQuartets is only slightly less efficient, contrary to the suggestion of Shi and Yang (2018), and it has the advantage of a much lower computational cost.

We show mathematically that estimation using multilocus data, if available, is better than estimation using only SNPs from the sampled loci if we wish to minimize the mean squared error (MSE) of the estimates of the site pattern probabilities. But where, exactly, does this benefit come from? Consider the following scenario. We have N genes, and we are interested in the quantity $\mathbf{p}$ for each gene, a vector of probabilities that a randomly chosen nucleotide in that gene takes on each of its possible combinations of values at the four gene tree tips. Clearly, if we knew $\mathbf{p}$ for all genes, or at least a large number of genes, then we could estimate the entries of the flattening matrix well. However, if we sample only a single SNP $\mathbf{X}$, where $\mathbf{X}$ is a length 256 vector of 0's and 1's that counts the number of times each site pattern in $\{A, C, G, T\}^4$ occurs, from each gene and let $\hat{\mathbf{p}} = \mathbf{X}$, then even though $\mathbf{X}$ is unbiased we can only observe a single value in the sample space $\{A, C, G, T\}^4$, so it is likely that $\mathbf{X}$ will be far from the true value of $\mathbf{p}$. If instead we sample n_i sites from each gene and let $\hat{\mathbf{p}} = \frac{1}{n_i}\sum_{j=1}^{n_i} \mathbf{X_j}$, then we are more likely to obtain a $\hat{\mathbf{p}}$ close to $\mathbf{p}$. This in turn means we are more likely to get a better estimate of the flattening matrix, and thus more likely to pick the correct tree, assuming that the number of genes N remains fixed. In other words, using multilocus data reduces the estimation error that results from failing to properly quantify within-gene variation.

Considering the two multilocus data estimators of the site pattern probabilities $\hat{\mathbf{p}}_3 = \frac{1}{N}\sum_{i=1}^{N} \mathbf{X}_i/n_i$ and $\hat{\mathbf{p}}_4 = \frac{1}{\sum_{i=1}^{N} n_i}\sum_{i=1}^{N} \mathbf{X}_i$, we show that $\hat{\mathbf{p}}_3$ should be preferred.While both are unbiased under our assumptions, $\hat{\mathbf{p}}_3$ has lower variance and thus lower MSE. However,

there is another important reason to prefer $\hat{\mathbf{p}}_3$. In our discussion of multilocus data, we assumed the number of loci being sampled from each gene, n_i, is a fixed constant known in advance and thus independent of any characteristics of the genes. In practice it is not known whether this is true. It could be, for example, that n_i tends to be larger for longer genes and that longer genes are structurally different from shorter genes. If this is in fact the case, then $\hat{\mathbf{p}}_4$ is no longer unbiased, while $\hat{\mathbf{p}}_3$ remains unbiased. The currently available software package for implementing SVDQuartets uses $\hat{\mathbf{p}}_4$ as its default estimator for multilocus data. In light of our results, we suggest that practitioners who wish to use SVDQuartets for estimation use $\hat{\mathbf{p}}_3$ instead.

References

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.

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.

Goldman, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

Golub, G. H. and C. F. Van Loan. 2013. Matrix Computations. AMS Johns Hopkins University Press.

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

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. Exchangeability and the evolution of large populations. Pp. 97-112 *in* G. Koch and F. Spizzichino, eds. Exchangeability in probability and statistics. North-Holland: Amsterdam.

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

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

Liu, L. and D. Pearl. 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., 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.

Liu, L., L. Yu, L. Kubatko, D. K. Pearl, and S. V. Edwards. 2009. Coalescent methods for estimating multilocus phylogenetic trees. Molecular Phylogenetics and Evolution 53:320–328.

Mirarab, S., R. Reaz, M. S. Bayzid, T. Zimmermann, M. S. Swenson, and T. Warnow. 2014. ASTRAL: Genome-scale coalescent-based species tree estimation. Bioinformatics 30:i541–i548.

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.

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.

Rannala, B. and Z. Yang. 2003. Likelihood and Bayes estimation of ancestral population sizes in hominoids using data from multiple loci. Genetics 164:1645–1656.

Roch, S., M. Nute, and T. Warnow. 2018. Long-branch attraction in species tree estimation: inconsistency of partitioned likelihood and topology-based summary methods. arXiv:1803.02800 .

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.

Rogers, J. 1997. On the consistency of maximum likelihood estimation of phylogenetic trees from nucleotide sequences. Systematic Biology 46:354–357.

RoyChoudhury, A., A. Willis, and J. Bunge. 2015. Consistency of a phylogenetic tree maximum likelihood estimator. Journal of Statistical Planning and Inference 161:73–80.

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.

Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences (American Mathematical Society) 17:57–86.

Truszkowski, J. and N. Goldman. 2016. Maximum likelihood phylogenetic inference is consistent on multiple sequence alignments, with or without gaps. Systematic Biology 65:328–333.

Yang, Z. 1994. Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. Systematic Biology 43:329–342.

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