EVALUATING VARIATIONS ON THE STAR ALGORITHM FOR RELATIVE EFFICIENCY AND SAMPLE SIZES NEEDED TO RECONSTRUCT SPECIES TREES

JAMES H. DEGNAN

Department of Mathematics and Statistics, University of Canterbury, Christchurch, 8140, New Zealand
** E-mail: j.degnan@math.canterbury.ac.nz*
www.canterbury.ac.nz

Many methods for inferring species trees from gene trees have been developed when incongruence among gene trees is due to incomplete lineage sorting. A method called STAR (Liu et al, 2009), assigns values to nodes in gene trees based only on topological information and uses the average value of the most recent common ancestor node for each pair of taxa to construct a distance matrix which is then used for clustering taxa into a tree. This method is very efficient computationally, scaling linearly in the number of loci and quadratically in the number of taxa, and in simulations has shown to be highly accurate for moderate to large numbers of loci as well as robust to molecular clock violations and misestimation of gene trees from sequence data. The method is based on a particular choice of numbering nodes in the gene trees; however, other choices for numbering nodes in gene trees can also lead to consistent inference of the species tree. Here, expected values and variances for average pairwise distances and differences between average pairwise distances in the distance matrix constructed by the STAR algorithm are used to analytically evaluate efficiency of different numbering schemes that are variations on the original STAR numbering for small trees.

Keywords: Statistical consistency, phylogenetics, multispecies coalescent, incomplete lineage sorting, sample size

1. Introduction

Numerous methods have been developed in recent years for inferring species trees (trees describing the history of speciation events for a set of species) from gene trees (trees on which DNA sequences evolve).[1–5] Methods that explicitly model the multispecies coalescent and account for uncertainty in the gene trees due to the mutation process can be the most accurate when gene tree discordance is due to incomplete lineage sorting, but can also be computationally very slow, particularly in the number of genes. In practice researchers sometimes have difficulty with convergence of the MCMC algorithms for these methods due to the relatively large number of genes.[6] With whole genome sequencing becoming increasingly common, this problem with the methods being able to keep up with the data is likely to increase in the future and motivates the need for computationally more efficient methods that will still be powerful enough to make accurate inferences. Methods that do not explicitly model the multispecies coalescent (e.g., rooted triple consensus,[7] R*,[8] STEAC and STAR,[9,10] the quartet version of BUCKY,[11] and triplet MRP[12] can still be robust under the model and can have the advantages of performing well under model violations and being computationally efficient enough to handle genomic levels of data.

A particularly promising method in simulations has been STAR,[9] which stands for Species Tree inference using Average Ranks. The method assigns a value to each node in an input gene

tree. The pairwise distance between two leaves of the tree is interpreted as twice the value of the node of their most recent common ancestor (MRCA) in the gene tree, and the pairwise distances for every pair of species is averaged over all loci. The resulting distance matrix can then be used to construct a tree using any clustering algorithm, for instance UPGMA or neighbor joining.

A key issue for the algorithm to work is how to assign the node values. The original STAR algorithm assigns a value of n to the root node, ρ, and the value of a node k is n minus the number of edges separating the node from the root. These node values are called "ranks" in Liu et al. (2009), where a higher rank means fewer edges separate the node from the root. (This usage of "rank" is slightly different from the usage of ranked trees elsewhere, where real-valued divergence times are sorted and their relative order is used to determine the rank of a node[13,14]) The node numbering used by STAR can also be interpreted as replacing all branch lengths on the gene trees with length 1 (extending external branch lengths as necessary to make trees ultrametric), and computing the average distance for each pair of species on these transformed gene trees. This numbering scheme leads to statistically consistent estimation of the species tree topology in the sense that as more independent loci (gene trees) are used, the probability that the method returns the correct species tree topology approaches 1.

Although the original numbering scheme used in STAR is statistically consistent, other numbering schemes also lead to consistent inference, as is shown in.[15] This naturally raises the question of whether other numbering schemes could be better or worse than STAR, and whether there is an optimal numbering scheme? This paper addresses this question by analytically determining expected values and variances of average distances between species in the distance matrix constructed by generalized versions of STAR for 4-taxon trees. An additional application of this approach is that sample sizes (numbers of independent loci) needed to confidently reconstruct certain inequalities in pairwise distances between taxa can be estimated.

2. Generalized STAR

To generalize STAR, let the value assigned to an internal node of a gene tree be a_j, where j is the number of edges separating the node from the root, ρ. Thus, the root node gets value a_0, the two daughter nodes of the root get value a_1 (assuming neither is a leaf), etc. There are at most $n-1$ distinct "ranks" in a gene tree, and each is only used if the gene tree is completely unbalanced (a *caterpillar* topology in which only one internal node has two leaves as its immediate descendants). Thus, a balanced four-taxon tree only uses a_0 for the root and a_1 for the two internal nodes. Thus a numbering scheme can be specified as an $(n-1)-tuple$, $(a_0, a_1, \cdots, a_{n-2})$. For the standard STAR algorithm, $a_0 = n$ and $a_i = a_{i-1} - 1$, $1 \leq i \leq n-2$. We define a *generalized STAR numbering scheme* for an n-taxon species tree to be any $(n-1)$-tuple $(a_0, \ldots, a_{n-2})$ satisfying $a_0 \geq a_1 \geq \cdots, \geq a_{n-2}$, where at least one of the inequalities is strict. The same numbering scheme is applied to each gene tree at each locus, and we assume that all gene trees have the same taxa, although these assumptions can be relaxed somewhat (see Allman et al. (2012)).

In the notation used in this paper, the STAR algorithm works by creating a distances matrix, where the (i,j)th entry is the average distance between taxa i and j, $\overline{D}_{ij}$. Letting $D_{ij}^{(\ell)}$

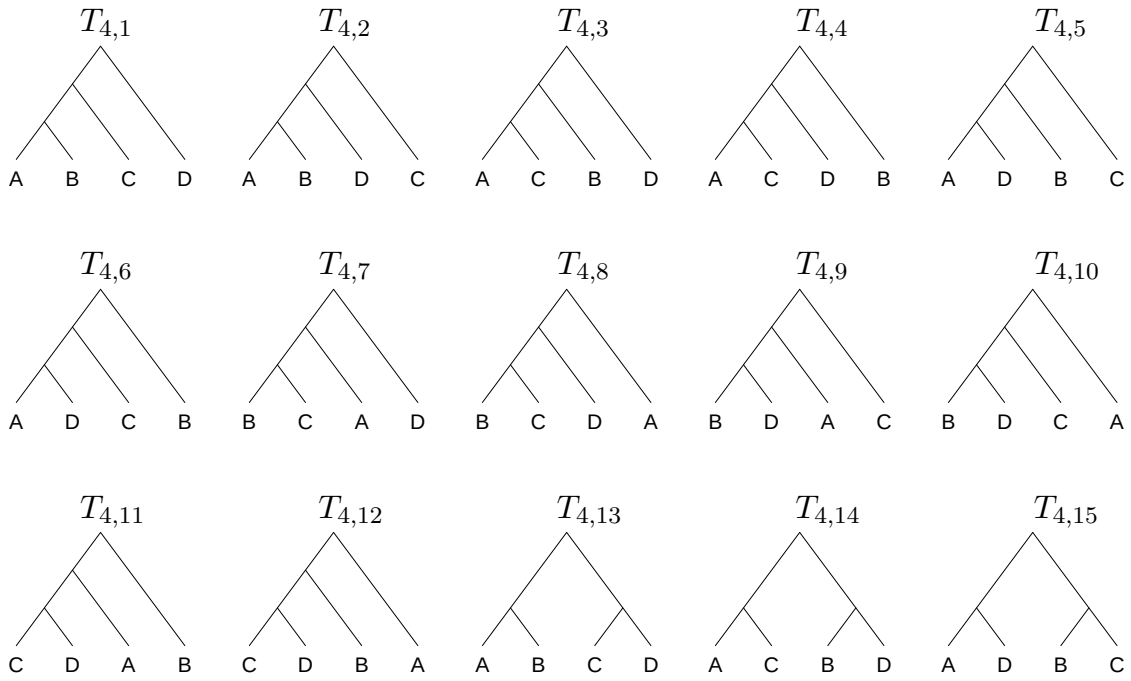

Fig. 1. Four-taxon trees used to determine expected values of the STAR distance matrix in the four-taxon case.

be the distance between taxa i and j at locus ℓ, if there are N loci, then $\overline{D}_{ij} = (1/N)\sum_{i=1}^{N} D_{ij}^{(\ell)}$.

For the 4-taxon case, the standard STAR algorithm uses $(a_0, a_1, a_2) = (4, 3, 2)$. In the standard STAR numbering scheme, all internal branches are equal in length and external branch lengths can be chosen to make the gene tree ultrametric (so that the distance from root to tip is constant). Translating the distances (adding a constant to each distance) or multiplying each by a constant factor should not affect the clustering applied to the distance matrix generated by STAR. Hence for the 4-taxon case, we can consider a generalized numbering scheme $(1, a, 0)$ and try to determine the optimal value of a, where $a = 1/2$ yields the same species tree estimate as the original STAR numbering scheme. More generally, we can consider a numbering scheme $\mathbf{a} = (a_0, \ldots, a_{n-2})$ to be equivalent to the numbering scheme $(\mathbf{a} - a_0)/(a_0 - a_{n-2})$, which fixes the smallest and largest values at 0 and 1, respectively. To determine consequences of different choices of a for $(1, a, 0)$, formulas for moments of STAR distances are shown next.

3. Expected values and variances of STAR distances

Explicit calculations of expected values, variances, and covariances of STAR distances can be used to estimate sample sizes necessary for the STAR tree to have certain relationships over others. For the 4-taxon species tree $\sigma_{4,1} = (((A, B){:}x, C){:}y, D)$, we are particularly interested in the sample size necessary for the STAR tree to have clade $\{ABC\}$ as opposed to clade $\{CD\}$. For notation, we let D_{ij} be the distance between taxa i and j on a single random gene tree occurring on the species tree. We let $\mathbb{E}[D_{ij}]$ be the expected distance between taxa i and j. Thus, as the number of loci goes to infinity STAR tree has clade $\{ABC\}$ as opposed to clade $\{CD\}$ for species tree $\sigma_{4,1}$ if $\mathbb{E}[D_{AB}] < \mathbb{E}[D_{AC}] = \mathbb{E}[D_{BC}] < \mathbb{E}[D_{CD}]$. The greatest difficulty is in

being confident (having enough loci) that the last inequalities, $\mathbb{E}[D_{AC}], \mathbb{E}[D_{BC}] < \mathbb{E}[D_{CD}]$ hold.

We can determine expected values and higher moments for the random distances D_{ij} for a generalized star scheme by

$$\mathbb{E}[D_{ij}^k] = \sum_{y=1}^{(2n-3)!!} (d_{ij}(y))^k \, p_{n,y}(\lambda), \tag{1}$$

where y indexes the gene tree topology, $d_{ij}(y)$ is the observed value of the random variable D_{ij} ($d_{ij}(y)$ depends on the topology y), $p_{n,y}$ is the probability of gene tree topology y in some ordering of tree topologies for n taxa, and λ is the set of internal branch lengths on the species tree. Four-taxon tree topologies are listed and enumerated as $T_{4,y}$, $y = 1, \ldots, 15$, in Figure 1, so that $p_{4,y}$ is the probability that a gene tree has topology $T_{4,y}$. The probabilities $p_{n,y}$ can be computed symbolically using the software COAL.[16]

Additionally, we will need covariances, which can be obtained from

$$\mathbb{E}[D_{ij}D_{k\ell}] = \sum_{y=1}^{(2n-3)!!} d_{ij}(y)\, d_{k\,\ell}(y)\, p_{n,y}(\lambda) \tag{2}$$

where at least two of $\{i, j, k, \ell\}$ are distinct.

From the Central Limit Theorem, the random variables $\overline{D}_{BC}$, $\overline{D}_{CD}$, and $\overline{D}_{CD} - \overline{D}_{BC}$ converge in distribution to normal random variable as the number of loci goes to infinity. We know that $\mathbb{E}[D_{CD} - D_{BC}] > 0$, so that given enough loci, C will be likely to be clustered with B (and therefore also with A) rather than D. We therefore need the variance of $D_{CD} - D_{BC}$ to determine how many loci will be needed with a given probability for the inequality to be positive. Here we have

$$\mathbb{V}(D_{CD} - D_{BC}) = \mathbb{V}(D_{CD}) + \mathbb{V}(D_{BC}) - 2Cov(D_{CD}, D_{BC}), \tag{3}$$

where $\mathbb{V}$ and Cov are the variance and covariance, respectively. These can be evaluated using equations (1) and (2). Knowing the approximate normal distribution for $\overline{D}_{CD} - \overline{D}_{BC}$ as a function of the numbering scheme (a_0, a_1, a_2) also allows us to compare the relative efficiencies of different numbering schemes in terms of the sample size needed to have a high probability of obtaining the correct species tree estimate.

Although the Central Limit Theorem applies asymptotically, in practice, the distances $\overline{D}_{BC}$, $\overline{D}_{CD}$, $\overline{D}_{CD} - \overline{D}_{BC}$ have detectable deviations from normality with 10 loci, and are slightly left-skewed. Simulations were done with STAR to test the applicability of the Central Limit Theorem for finite samples of size 10, 50, 100, and 500 loci on the species tree $\sigma_{4,1}$. The normality of $\overline{D}_{CD} - \overline{D}_{BC}$ was tested using the Shapiro-Wilks test in R,[17] and results are listed in Table 1 for the numbering schemes (4,3,2) and (4,3,0). Statistically significant deviations are detectable with a sample size of 100 or less, but are difficult to detect with samples of size 500 loci. We note that although deviations from normality are detectable, the power to detect deviations is fairly high, since there are 1000 observations, and deviation from normality is difficult to detect by eye using histograms.

Table 1 also lists the c.o.v. (estimated from the simulations), and the proportion of estimated species trees that are correctly inferred using UPGMA implemented in Phybase[18] on the estimated distance matrix, both of which can be used as measures of the efficiency of the

Table 1. Expected values, variances, tests of normality for $\overline{D}_{CD} - \overline{D}_{BC}$ estimated from finite numbers of loci, and proportion of times the correct species tree was estimated under the STAR algorithm. The standard deviation and c.o.v. are based on the sample size, and are $\sqrt{v(a)/n}$ and $\sqrt{v(a)/n}/e(a)$, respectively. P-values are for the normality of $\overline{D}_{CD} - \overline{D}_{BC}$.

Branch lengths		$\overline{D}_{CD} - \overline{D}_{BC}$					proportion
(x, y)	(a_0, a_1, a_2)	loci	mean	sd	c.o.v.	p-value	correct
$(0.05, 0.05)$	$(4, 3, 2)$	10	0.047	0.325	6.919	0.023	0.170
$(0.05, 0.05)$	$(4, 3, 2)$	50	0.056	0.140	2.337	0.076	0.253
$(0.05, 0.05)$	$(4, 3, 2)$	100	0.061	0.098	1.619	0.190	0.363
$(0.05, 0.05)$	$(4, 3, 2)$	500	0.063	0.046	0.718	0.868	0.793
$(0.05, 0.05)$	$(4, 3, 0)$	10	0.107	0.570	5.350	0.000	0.145
$(0.05, 0.05)$	$(4, 3, 0)$	50	0.118	0.246	2.093	0.349	0.275
$(0.05, 0.05)$	$(4, 3, 0)$	100	0.120	0.173	1.438	0.555	0.394
$(0.05, 0.05)$	$(4, 3, 0)$	500	0.122	0.079	0.646	0.225	0.849
$(1.00, 0.05)$	(4,3,2)	10	0.052	0.273	5.234	0.000	0.452
$(1.00, 0.05)$	(4,3,2)	50	0.055	0.122	2.204	0.004	0.535
$(1.00, 0.05)$	(4,3,2)	100	0.053	0.088	1.651	0.069	0.619
$(1.00, 0.05)$	(4,3,2)	500	0.055	0.034	0.707	0.604	0.894
$(1.00, 0.05)$	(4,3,0)	10	0.076	0.380	5.022	0.000	0.452
$(1.00, 0.05)$	(4,3,0)	50	0.075	0.176	2.273	0.070	0.551
$(1.00, 0.05)$	(4,3,0)	100	0.077	0.125	1.617	0.137	0.652
$(1.00, 0.05)$	(4,3,0)	500	0.079	0.056	0.708	0.340	0.905

two numbering schemes. For the species tree with branches $(x, y) = (0.05, 0.05)$, for each given number of loci, the scheme $(4, 3, 2)$ has a higher c.o.v. than $(4, 3, 0)$, although proportions of correctly inferred trees are only statistically significantly better for $(4, 3, 0)$ when sample sizes reach 500 loci. Note, however, that both in simulation (Table 1) and based on theoretical sample size calculations in Table 2, $(4, 3, 2)$ and $(4, 3, 0)$ are approximately equally good for $(x, y) = (1.0, 0.05)$. We note that $(x, y) = (0.05, 1.0)$ leads to more gene tree discordance than $(1.0, 0.05)$

4. Evaluation of variations on STAR

4.1. *The 4-taxon case*

To evaluate generalized STAR in the 4-taxon case, we let the numbering scheme be $(1, a, 0)$. To find an optimal value of a, set $e(a) := \mathbb{E}_a[D_{CD} - D_{BC}]$ and $v(a) = \mathbb{V}_a[D_{CD} - D_{BC}]$, i.e., taking means and variances parameterized by a. Using the normal approximation, the probability that

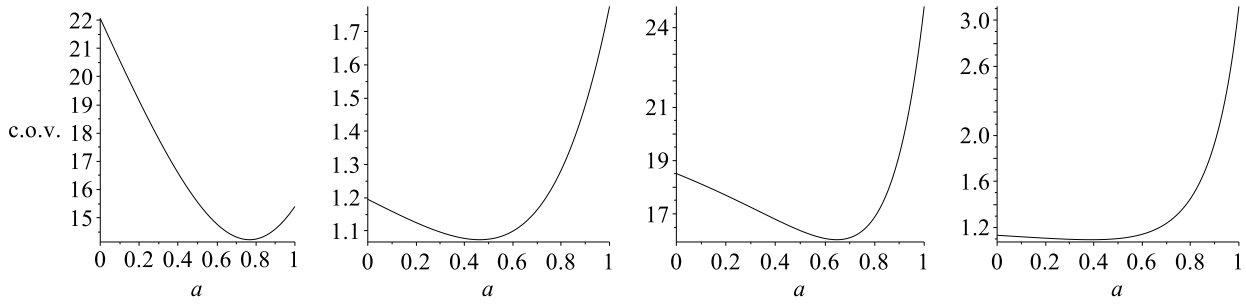

Fig. 2. Coefficient of variation for $D_{CD} - D_{BC}$ as a function of a using the STAR numbering scheme $(1, a, 0)$ for species tree $\sigma_{4,1}$ with $(x, y) = (0.05, 0.05), (0.05, 1.0), (1.0, 0.05), (1.0, 1.0)$.

$\overline{D}_{CD} - \overline{D}_{BC}$ is greater than 0 is approximately $\mathbb{P}_a[Z < (0 - e(a))/\sqrt{v(a)/n}] = \Phi(\sqrt{n}e(a)/\sqrt{v(a)})$, where Z is a standard normal random variable and Φ is the standard normal cumulative distribution function. Thus the sample size, N, needed to have confidence $1-\alpha$ that $\mathbb{E}[D_{CD} - D_{BC}] > 0$ is approximately

$$N = \lceil (\Phi^{-1}(1-\alpha)\text{c.o.v.}(a))^2 \rceil \tag{4}$$

where $\text{c.o.v.}(a) = \sqrt{v(a)}/e(a)$ is the coefficient of variation. We consider the optimal value of a is the value that minimizes N in equation (4), or equivalently, that minimizes the coefficient of variation, $\sqrt{v(a)}/e(a)$. For species tree $(((A, B):x, C):y, D)$, the coefficient of variation under the scheme $(1, a, 0)$ can be written analytically using

$$\begin{aligned}
v(a) = {} & \Big(- \,\mathrm{e}^{-2x} - 9\mathrm{e}^{-2y} - 7\,\mathrm{e}^{-x-3y} + 6\,\mathrm{e}^{-4y-x} + 15\,\mathrm{e}^{-y} + 2\,\mathrm{e}^{-2x-3y} + 3\,\mathrm{e}^{-x} \\
& - 3\,\mathrm{e}^{-x-y} - \,\mathrm{e}^{-2x-6y}\Big) a^2/9 + \Big(- 30\,\mathrm{e}^{-y} + 3\,\mathrm{e}^{-x-2y} - 1\,\mathrm{e}^{-2x-3y} + 18\,\mathrm{e}^{-2y} \\
& + 10\,\mathrm{e}^{-x-3y} - 9\mathrm{e}^{-4y-x} + \,\mathrm{e}^{-2x-6y} + \,\mathrm{e}^{-2x-y} - \,\mathrm{e}^{-2x-4y}\Big) a/9 \\
& - 1/3\,\mathrm{e}^{-x-2y} + 1/3\,\mathrm{e}^{-4y-x} + 1/18\,\mathrm{e}^{-2x-4y} - \mathrm{e}^{-2y} - 1/36\,\mathrm{e}^{-2x-2y} - 1/36\,\mathrm{e}^{-2x-6y} \\
& + 5/3\,\mathrm{e}^{-y} - 5/18\,\mathrm{e}^{-x-3y} + 1/6\,\mathrm{e}^{-x-y} \\
e(a) = {} & \left(1/3\,\mathrm{e}^{-x} - 1 + \mathrm{e}^{-y} - 1/3\,\mathrm{e}^{-x-3y}\right) a + 1 - \mathrm{e}^{-y} - 1/6\,\mathrm{e}^{-x-y} + 1/6\,\mathrm{e}^{-x-3y}
\end{aligned}$$

where these values were computed symbolically using equations (1)-(3), using COAL for the gene tree probabilities $p_{n,i}(\lambda)$, and simplifying in the software MAPLE.

The optimal value of a is difficult to find analytically as a function of x and y; however, for fixed x and y, one can equivalently find the optimal value of $v(a)/e^2(a)$, which is a rational function with both numerator and denominator being quadratic functions in a, and the minimum of this function can be found analytically. For $(x, y) = (0.05, 0.05)$, for example, the optimal value is $a \approx 0.767$. This value is close to $a = 3/4$, which is equivalent to the numbering scheme $(4, 3, 0)$. The coefficient of variation as a function of a is shown in Figure 2 for a few choices of (x, y) and for species trees $\sigma_{4,1}$.

We compute sample sizes required to get a 95% chance that a random sample of N loci results in $D_{CD} - D_{BC} > 0$ for two choices of (x, y) in Table 2. In the table, the root is difficult to resolve, and for $x = 1.0$, the fact that A and B form a clade is less to difficult to infer. We

note that for $(x, y) = (0.05, 0.05)$, the numbering scheme $(4, 3, 1)$ does best among those listed, while for $(x, y) = (0.05, 1.0)$, the numbering scheme $(4, 3, 0)$ does best amongst the same set of numbering schemes.

We note that choosing a to maximize the probability that $D_{CD} - D_{BC} > 0$ does not necessarily maximize the probability that the STAR tree matches the species tree. In particular, for (x, y), if x is small and y is large, then $D_{CD} - D_{BC} > 0$ with high probability, and the more difficult relationships to resolve will be those between taxa A, B, and C. In this case, it might make sense to find a that maximizes the probability that $D_{BC} - D_{AB} > 0$, and sample sizes sufficient for $\overline{D}_{CD} - \overline{D}_{BC} > 0$ are unlikely to be sufficient for $\overline{D}_{BC} - \overline{D}_{AB} > 0$ to obtain.

The sample sizes here are only for being 95% confident that $\overline{D}_{CD} - \overline{D}_{BC} > 0$, which does not guarantee that the correct species tree will be estimated, although in practice, this is often the case. For the scheme $(4, 3, 0)$, a sample size of 548 is needed for 95% confidence that $\overline{D}_{CD} - \overline{D}_{BC} > 0$ when $(x, y) = (0.05, 0.05)$. In simulation, a sample size of 500 recovered the species tree only 84.9% of the time, although by formula (4), a sample size of 500 should have a 94% $(= \Phi(1.571))$ that $\overline{D}_{CD} - \overline{D}_{BC} > 0$. It is not surprising that sample sizes needed to recover the entire tree are somewhat larger than what is needed to estimate the inequality, as for example, $\overline{D}_{CD} - \overline{D}_{BC} > 0$ does not guarantee that $\overline{D}_{AB}$ is the smallest estimated distance, although this is necessary to correctly estimate the species tree.

An alternative approach to guaranteeing that a particularly difficult inequality is estimated correctly with high probability is to guarantee that all pairwise inequalities are estimated correctly. Given the lack of independence between pairwise distances, this is difficult to do exactly. However, using Bonferroni's inequality, k events (not necessarily independent or equiprobable), that each have probability at least $1 - \varepsilon/k$, all occur with probability at least $1 - \varepsilon$.[19] Thus, one could choose, for example, the sample size needed to correctly determine $D_{CD} - D_{BC} > 0$ with probability $1 - \alpha = 0.99$, and conclude that all $\binom{4}{2} = 6$ pairwise relationships (and therefore the correct tree) will be inferred with probability at least $1 - 6\alpha = 0.94$. In general, this approach will be quite conservative (i.e., will overestimate the number of loci needed) if it is based on the most difficult pairwise inequality. Sample sizes needed for 99% confidence can be obtained from 95% values by multiplying by $[\Phi^{-1}(1 - 0.99)/\Phi^{-1}(1 - 0.95)]^2 = (2.326/1.645)^2 \approx 2.00$. Thus, this approach suggests that samples sizes being doubled (for the 4-taxon case) would give approximately at least as much confidence that the entire tree was estimated correctly as well as the inequality $D_{CD} - D_{BC} > 0$.

From the 4-taxon examples, the branch lengths $(x, y) = (0.05, 0.05)$ are in the *anomaly zone*, in which the most likely gene tree topology is $((AB)(CD))$ rather than $(((AB)C)D)$.[20] However, $(x, y) = (1.0, 0.05)$ is not in the anomaly zone (i.e., the most likely gene tree topology matches the species tree topology) but requires similarly large samples (hundreds of loci) to recover the species tree with high probability (Table 1). The results are similar to other studies that have shown that hundreds of loci might be needed to accurately reconstruct the species tree from gene tree topologies when gene tree discordance is this high.[9,21]

Table 2. Samples sizes and c.o.v. needed for approximate 95% confidence that $\overline{D}_{CD} - \overline{D}_{BC} > 0$. The c.o.v. is based on $\sqrt{v(a)}/e(a)$ for a single locus.

(x, y)	(a_0, a_1, a_2)	$(1, a, 0)$	number of loci needed	c.o.v.
$(0.05, 0.05)$	$(4, 3, 2)$	$(1, 0.5, 0)$	655	15.553
$(0.05, 0.05)$	$(4, 3, 1)$	$(1, 0.67, 0)$	564	14.428
$(0.05, 0.05)$	$(4, 3, 0)$	$(1, 0.75, 0)$	548	14.230
$(0.05, 0.05)$	$(4, 3.5, 0)$	$(1, 0.875, 0)$	567	14.474
$(0.05, 0.05)$	$(4, 2, 1)$	$(1, 0.33, 0)$	817	17.375
$(1.00, 0.05)$	$(4, 3, 2)$	$(1, 0.5, 0)$	726	16.371
$(1.00, 0.05)$	$(4, 3, 1)$	$(1, 0.67, 0)$	697	16.038
$(1.00, 0.05)$	$(4, 3, 0)$	$(1, 0.75, 0)$	725	16.358
$(1.00, 0.05)$	$(4, 3.5, 0)$	$(1, 0.875, 0)$	919	18.428
$(1.00, 0.05)$	$(4, 2, 1)$	$(1, 0.33, 0)$	791	17.097

4.2. *A 5-taxon example*

Another example of using different numbering schemes to distinguish difficult-to-resolve relationships is for the two species trees $\sigma_{5,1} = (((A, B){:}x, C){:}y, (D, E){:}z)$ and $\sigma_{5,2} = ((A, B){:}u, (C, (D, E){:}v){:}w)$. For $\sigma_{5,1}$, if x and y are small while z is relatively large, the most likely gene tree could have the same topology as $\sigma_{5,2}$. Similarly, if v and w are small, while u is relatively large, a gene tree with the same topology as $\sigma_{5,1}$ could be the most likely gene tree when $\sigma_{5,2}$ is the species tree. This example with these two candidate species trees is actually the smallest example of a "wicked forest", in which for each of two or more candidate species trees, the most likely gene tree topology matches a different species tree.[20,22] In this example, the clades $\{AB\}$ and $\{DE\}$ might not be very difficult to estimate, and the greatest difficulty is in deciding on which side of the root taxon C belongs. We note that this example was also one of the more difficult cases for estimating rooted species trees from unrooted gene trees.[23]

To get a sense of sample sizes that might be needed to correctly place taxon C, and to find an optimal numbering scheme (a_0, a_1, a_2, a_3) to use with STAR, we consider $D_{CD} - D_{BC}$. Here we map the smallest and largest values of the numbering scheme to 0 and 1, respectively, and consider schemes $(1, a_1, a_2, 0)$ with $1 > a_1 > a_2 > 0$. A plot of the coefficient of variation is given in Figure 3 for the species tree $(((A, B){:}x, C){:}y, (D, E){:}z)$ with $(x, y, z) = (0.05, 0.05, 1.0)$, which shows that larger values of a_1 tend to be more efficient, although some efficiency is lost with value of a_1 too close to 1, and that the choice of a_1 is more important than the choice of a_2.

Sample size calculations can be done as in the 4-taxon case, using $\mathbf{a} = (a_0, a_1, a_2, a_3)$ in place of a in equation (4). Here, a near optimal choice for $\mathbf{a}$ is $(1.0, 0.88, 0.50, 0.0)$. This is equivalent to $(5.00, 4.64, 3.5, 2.00)$ when the smallest and largest values are fixed at 2.0 and 5.0. Similarly, the standard STAR numbering scheme of $(5, 4, 3, 2)$ is equivalent to $(1, 2/3, 1/3, 0)$. Estimated expected values, standard deviations, c.o.v. (both estimated and theoretical), and proportion

Table 3. Expected values standard deviation, c.o.v., and for $\mathbb{E}[D_{CD} - D_{BC}]$ estimated from finite numbers of loci, and proportion of times the correct species tree was estimated under the STAR algorithm using species tree $(((A, B){:}x, C){:}y, (D, E){:}z)$. The theoretical c.o.v. is $\sqrt{\mathbb{V}([D_{CD} - D_{BC}]/n)}/\mathbb{E}[D_{CD} - D_{BC}]$.

Branch lengths (x, y, z)	numbering scheme	loci	mean	sd	c.o.v. (theoretical)	proportion correct
$(0.05, 0.05, 1.0)$	(5,4,3,2)	10	0.0691	0.336	4.861 (4.841)	0.144
$(0.05, 0.05, 1.0)$	(5,4,3,2)	50	0.071	0.150	2.105 (2.165)	0.255
$(0.05, 0.05, 1.0)$	(5,4,3,2)	100	0.073	0.104	1.434 (1.531)	0.375
$(0.05, 0.05, 1.0)$	(5,4,3,2)	500	0.068	0.046	0.670 (0.684)	0.800
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	10	0.062	0.266	4.308 (4.439)	0.152
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	50	0.054	0.106	1.964 (1.985)	0.273
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	100	0.058	0.080	1.376 (1.403)	0.405
$(0.05, 0.05, 1.0)$	(5,4.64,3.5,2)	500	0.055	0.034	0.611 (0.628)	0.865

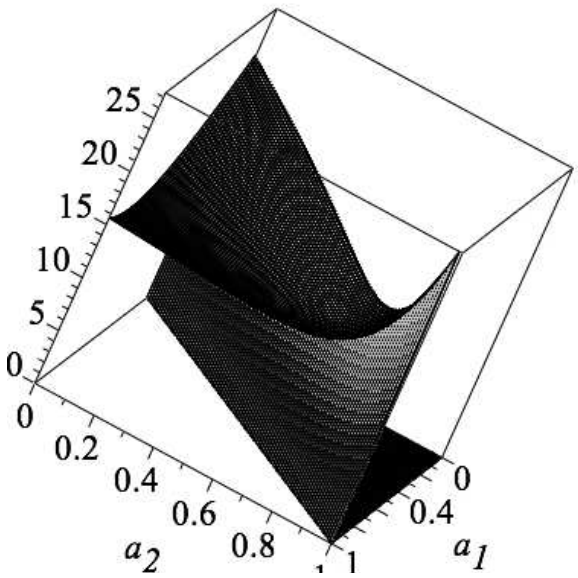

Fig. 3. C.o.v. as a function of a_1 and a_2 for the numbering scheme $(1, a_1, a_2, 0)$ for the species tree $(((A, B){:}x, C){:}y, (D, E){:}z)$ with $(x, y, z) = (0.05, 0.05, 1.0)$. The drop along the plane $a_1 = a_2$ occurs because of the assumption that $a_1 > a_2$.

of STAR trees matching the species tree are shown in Table 3. The sample size needed to determine $D_{CD} - D_{BC} > 0$ with 95% confidence is roughly $N = 534$ with $(a_0, a_1, a_2, a_3) = (5.00, 4.64, 3.50, 2.00)$ and $N = 634$ with $(a_0, a_1, a_2, a_3) = (5, 4, 3, 2)$.

5. Discussion

This paper has shown a framework for investigating variations on the STAR numbering scheme for the purpose of evaluating the relative efficiency of different schemes. The original STAR numbering scheme is well-chosen in that it is simple and works well in a wide variety of situations – i.e., for both long and short branches in the species trees investigated in this paper, the original STAR numbering of equally spaced branches often had a relatively low coefficient of variation, and optimal values for given species tree branch lengths are not necessarily optimal for other branch lengths. Overall, there is no numbering scheme that is uniformly optimal — that performs better than any other scheme for all species tree branch lengths.

If there is some knowledge of the species tree topology, in particular nodes that might be especially difficult to resolve, alternatives to the original STAR numbering scheme can perform better in some situations. In particular, if a node in the species tree is not very resolved, then making genes more star-like in the sense of making internal nodes closer to the root than under the standard STAR algorithm, can lead to improvements in estimating species trees in terms of the number of loci needed. For a fixed number of loci, this could result in improved bootstrap support for the problematic nodes. The sample size calculations used in this paper assume approximately normal distributions for the distances between taxa averaged over many loci. The normality assumption is more reasonable with large numbers of loci; thus, for branch lengths for which equation (4) returns a small number of loci, the normality assumption is less plausible. Instead, equation (4) is intended for use with difficult species trees for which large sample sizes might be required, making the normality assumption more reasonable.

In this paper, only known gene trees have been used, although in practice gene trees are estimated with some error. Because topologies can typically be estimated more reliably than branch lengths, however, STAR and its variations should be less sensitive to misestimation of gene trees than methods that use branch lengths.[9] Although the effects of misestimation on species tree inference can be simulated directly, we note that theoretical expected values, variances, and covariances, and therefore sample size calculations do not assume that gene tree probabilities are obtained directly from the multispecies coalescent. Instead, the probabilities $p_{n,i}$ used in equations (1) and (2) can come from any model for the gene tree topologies, including a model that includes error in the gene trees. In particular, if a distribution on estimated gene trees is obtained, say $\{\widehat{p}_i\}$, then this distribution can be used in equations (1) and (2), and the relative efficiency of different numbering schemes can be compared on different distributions of estimated trees. Similarly, effects of other processes, such as horizontal gene transfer,[24] gene duplication,[25,26] and hybridization[27,28] can be studied as long as distributions of gene tree topologies can be obtained (either theoretically or estimated through simulations).

Some unanswered questions raised by this study is whether the original STAR numbering scheme performs best "on average", perhaps averaged over species trees generated on a Yule model, and whether one STAR numbering scheme can dominate another — that is, could one STAR numbering scheme always perform better than another for all possible topologies and branch lengths in the species tree? The framework used in this paper of using expected pairwise distances as well as their variances and covariances could be used to investigate these questions further.

Acknowledgments

This work was supported by the New Zealand Marsden Fund. The author is grateful for comments from the anonymous reviewers.

References

1. B. Rannala and Z. Yang, *Annu. Rev. Genom. Human Genet.* **9**, 217 (2008).
2. J. H. Degnan and N. A. Rosenberg, *Trends Ecol. Evol.* **24**, 332 (2009).
3. S. V. Edwards, *Evolution* **63**, 1 (2009).
4. L. Liu, L. Yu, L. S. Kubatko, D. K. Pearl and S. V. Edwards, *Mol. Phylogenet. Evol.* **53**, 320 (2009).
5. L. L. Knowles and L. S. Kubatko, *Estimating species trees: practical and theoretical aspects* (Wiley-Blackwell, Hoboken, NJ, 2010).
6. K. A. Cranston, B. Hurwitz, D. Ware, L. Stein and R. A. Wing, *Syst. Biol.* **58**, 489 (2009).
7. G. B. Ewing, I. Ebersberger, H. A. Schmidt and A. von Haeseler, *BMC Evol. Biol.* **8**, p. 118 (2008).
8. J. H. Degnan, M. DeGiorgio, D. Bryant and N. A. Rosenberg, *Syst. Biol.* **58**, 35 (2009).
9. L. Liu, L. Yu, D. K. Pearl and S. V. Edwards, *Syst. Biol.* **58**, 468 (2009).
10. L. Liu and L. Yu, *Syst. Biol.* **60**, 661 (2011).
11. B. R. Larget, S. K. Kotha, C. N. Dewey and C. Ané, *Bioinformatics* **26**, 2910 (2010).
12. Y. Wang and J. H. Degnan, *Stat. Appl. Genet. Mol.* **10**, p. 21 (2011).
13. T. Gernhard, D. Ford, R. Vos and M. Steel, *Evolutionary Bioinformatics Online* **2**, 285 (2006).
14. J. H. Degnan, N. Rosenberg and T. Stadler, *Math. Biosci.* **235**, 45 (2012).
15. E. S. Allman, J. H. Degnan and J. A. Rhodes, *http://www.arxiv/abs/1204.4413* , 23 (2012).
16. J. H. Degnan and L. A. Salter, *Evolution* **59**, 24 (2005).
17. R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, (2012). ISBN 3-900051-07-0.
18. L. Liu and L. Yu, *Bioinformatics* **26**, 962 (2010).
19. S. Ross, *A First Course in Probability*, 5th edn. (Prentice-Hall, Upper Saddle River, NJ, 1998).
20. J. H. Degnan and N. A. Rosenberg, *PLoS Genet.* **2**, 762 (2006).
21. Y. Wu, *Evolution* **66**, 763 (2012).
22. N. A. Rosenberg and R. Tao, *Syst. Biol.* **57**, 131 (2008).
23. E. S. Allman, J. H. Degnan and J. A. Rhodes, *J. Math. Biol.* **62**, 833 (2011).
24. Y. Chung and C. Ané, *Syst. Biol.* **60**, 261 (2011).
25. O. Åkerbord, B. Sennblad, L. Arvestad and J. Lagergren, *Proc. Natl. Acad. Sci. USA* **106**, 5714 (2009).
26. M. Rasmussen and M. Kellis, *Genome Res.* **22**, 755 (2012).
27. C. Meng and L. S. Kubatko, *Theor. Popul. Biol.* **75**, 35 (2009).
28. Y. Yu, J. H. Degnan and L. Nakhleh, *PLoS Genet.* **8**, p. e1002660 (2012).