Systematic Biology Advance Access published July 9, 2009

Point of View

DOI:10.1093/sysbio/syp034

Phylogenetic Analysis in the Anomaly Zone

LIANG LIU* AND SCOTT V. EDWARDS

Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA;
E-mail: sedwards@fas.harvard.edu;
**Correspondence to be sent to: Department of Organismic and Evolutionary Biology, Harvard University, 26 Oxford Street, Cambridge, MA 02138, USA; E-mail: lliu@oeb.harvard.edu.*

The concatenation method has been widely used as a means of combining data to estimate phylogenetic trees (Huelsenbeck et al. 1996a, 1996b; Glazko and Nei 2003). However, simulation studies have shown that the maximum likelihood (ML) estimate of the species tree for concatenated sequences may be statistically inconsistent if the gene trees are highly heterogeneous (Kolaczkowski and Thornton 2004; Kubatko and Degnan 2007). Recently, Degnan and Rosenberg (2006) defined an “anomaly zone”—a set of short internal branches in species trees that will generate gene trees that are discordant with the species tree more often than gene trees that are concordant. Kubatko and Degnan (2007) went on to show that when DNA sequences are generated from gene trees simulated from species trees in the anomaly zone, as well as from species trees slightly outside this zone but still with short internal branches, the ML estimate of the species tree for the concatenated sequences can be inconsistent, resulting in increasing certainty in the wrong species tree. These studies were all performed with a molecular clock on rooted gene and species trees within the variation realized in stochastic simulations of DNA sequences under the Jukes and Cantor (1969) model of nucleotide substitution. They applied the ML method with a clock to recover phylogenetic trees from their simulated concatenated data sets.

In this paper, we show that phylogenetic methods that solely utilize the relative order of divergences among a set of DNA sequences as a criterion for inferring phylogenies, such as the unweighted pair group method with arithmetic mean (UPGMA), are statistically consistent even when DNA sequences are generated from gene trees simulated from species trees in the anomaly zone. In addition, we use simulation to assess the performance of a variety of tree construction methods when analyzing concatenated sequences generated from 4 and 5-taxon species tree located in the anomaly zone and show that a variety of methods do in fact recover the correct species tree topology, whereas ML, with or without a molecular clock, remains inconsistent. However, the branch lengths of the tree inferred from concatenated data are inevitably overestimated, as predicted by theory. Finally, simulations also suggest that a newly proposed Bayesian approach for estimating species trees from multiple unlinked loci, BEST (Liu and Pearl 2007; Liu et al. 2008), is consistent in both topology and branch lengths on data sets generated from species trees in the anomaly zone.

Theoretical Results

Expected Coalescence Time of Alleles Sampled from 2 Species in a Multispecies Tree

The mathematical results in this section are based on the following assumptions: 1) the coalescence time of any 2 sequences follows the coalescent process (Kingman 1982, 2000), 2) the gene trees and species trees are rooted and ultrametric, 3) the DNA sequences are generated from a Jukes–Cantor model (Jukes and Cantor 1969), 4) the species tree is strictly bifurcating (dichotomous), and 5) all lineages in the gene trees have the same mutation rate (molecular clock). Throughout, population sizes θ are equal to $\theta = 4N_e u$ and divergence times τ are equal to $\tau = ut$, where N_e is the effective population size, t is the species divergence time in generations, and u is the mutation rate per generation per site.

Coalescent theory has played a central role in investigating the relationship between gene coalescence times and species divergence times (Nielsen 1998; Rannala and Yang 2003; Liu and Pearl 2007; Mossel and Roch 2007; Efromovich and Kubatko 2008; Liu et al. 2008; Seo 2008). A species tree represents the history of species and lineages evolving through time. All species arise from a common ancestral population, and the genetic material is transmitted from ancestral populations to descendant populations along the branches of the species tree (Fig. 1). A gene tree represents the coalescence process of the sequences sampled from the contemporary populations (the tips of the species tree), looking backward in time along the ancestral branches in the species tree. Consider 2 sequences, s_1 and s_2, randomly selected from species S1 and S2 (Fig. 1). Let $\mathrm{AP}^0(s_1, s_2)$ be the most recent ancestral population of s_1 and s_2. To simplify the notation, we will drop (s_1, s_2) throughout this paper. The population $\mathrm{AP}^i(\mathrm{i} \geqslant 1)$ denotes the first ancestral population of the sequences in AP^{i-1} if AP^{i-1} is

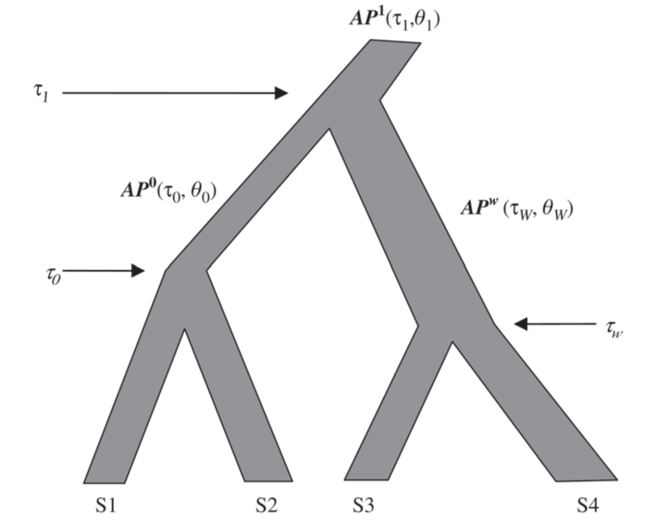

FIGURE 1. The species phylogeny of 4 species S1, S2, S3, and S4. AP, ancestral populations; τ, species divergence time; and θ, population sizes. The most recent common ancestral population of S1 and S2 is AP^0. The root population AP^1 is the parent population of AP^0. The population AP^w is the most recent common ancestral population of S3 and S4.

not the root population (Fig. 1). The set $\{\mathrm{AP}^0, \ldots, \mathrm{AP}^k\}$ in which AP^k is the root population includes all the ancestral populations of s_1 and s_2. The species divergence time and population size of the ancestral population AP^i are τ_i and θ_i. The sequences s_1 and s_2 may coalesce in any of the ancestral populations $\{\mathrm{AP}^0, \ldots, \mathrm{AP}^k\}$. For the species tree defined in Figure 1, they may coalesce in either the population AP^0 or the AP^1 (Fig. 1). Under the coalescence with speciation model, or the multispecies coalescent (Nielsen 1998; Rannala and Yang 2003), if the species tree S is given, the coalescence time T of the sequences s_1 and s_2 within a particular population has an exponential distribution. Thus, the conditional probability density function of coalescence time T within the ancestral population AP^0 given the species tree S and population size θ_0 is $f(T|\theta_0, S) = \frac{2}{\theta_0}e^{-2(T-\tau_0)/\theta_0}$ for $\tau_0 \leqslant T < \tau_1$, which indicates that the probability that s_1 and s_2 do not coalesce in the population AP^0 is equal to $e^{-2(\tau_1-\tau_0)/\theta_0}$. Therefore, the conditional probability density function of time T within population AP^1 is $f(T|\theta_0, \theta_1, S) = \frac{2}{\theta_1}e^{-2(T-\tau_1)/\theta_1} * e^{-2(\tau_1-\tau_0)/\theta_0}$ for $\tau_1 \leqslant T < \tau_2$. Similarly, the conditional probability density function of time T within population AP^i given the species tree S and population sizes $\{\theta_0, \ldots, \theta_i\}$ is

$$f(T|\theta_0, \ldots, \theta_i, S) = \frac{2}{\theta_i}e^{-2(T-\tau_i)/\theta_i} * \prod_{j=0}^{i-1} e^{-2(\tau_{j+1}-\tau_j)/\theta_j} \quad \text{for } \tau_i \leqslant T < \tau_{i+1} \tag{1}$$

in which $\prod_{j=0}^{i-1} e^{-2(\tau_{j+1}-\tau_j)/\theta_j}$ is the probability that the sequences do not coalesce in the previous populations $\{\mathrm{AP}^0, \ldots, \mathrm{AP}^{i-1}\}$. Note that $\tau_{i+1} = \infty$ if $\mathrm{AP}^i = \mathrm{AP}^k$, where AP^k is the root population. It follows that if the species tree S and population sizes $\{\theta_0, \ldots, \theta_k\}$ are given, the conditional expectation of the coalescence time for s_1 and s_2 is equal to

$$\begin{aligned} E(T|\theta_0, \ldots, \theta_k, S) &= \sum_{i=0}^{k-1}\left\{\int_{\tau_i}^{\tau_{i+1}} T^* f(T|\theta_0, \ldots, \theta_i)\mathrm{dT}\right\} \\ &\quad + \left\{\int_{\tau_k}^{\infty} T^* f(T|\theta_0, \ldots, \theta_k)\mathrm{dT}\right\} \\ &= \tau_0 + \theta_0/2 + \sum_{i=0}^{k-1}\left\{(\theta_{i+1}/2 - \theta_i/2) * \right. \\ &\quad \left. \times \prod_{j=0}^{i} e^{2(\tau_j-\tau_{j+1})/\theta_j}\right\}. \end{aligned} \tag{2}$$

It follows from Equation (2) that the expected gene coalescence time overestimates the species divergence time, as noted by many authors (Gillespie and Langley 1979; Nei 1987; Nei and Kumar 2000). The level of the overestimation depends on the population sizes and species divergence times as indicated by Equation (2). If $\theta_{i+1} < \theta_i$, the sum in Equation (2) is negative and the overestimation will be reduced. The overestimation can be further reduced when the lengths of internodes in the species tree approach 0 or when $\tau_{i+1} = \tau_i$ for $i = 0, \ldots, k-1$. This condition amounts to a star phylogeny in the species tree. If all populations have the same population size θ, the expected coalescence time is reduced to $E(T) = \tau_0 + \theta/2$, which is identical to previous results (Edwards and Beerli 2000).

Ancestral Order of Populations and Species Tree Estimation

We use the symbol $\prec$ to denote the ancestral order of populations along a lineage. For 2 ancestral populations AP and AP^*, $\mathrm{AP} \prec \mathrm{AP}^*$ indicates that AP is the descendant population of AP^* and the time before present associated with AP is less than that of AP^*. For example, populations AP^0 and AP^1 in Figure 1 have ancestral order $\mathrm{AP}^0 \prec \mathrm{AP}^1$, but the population AP^0 and AP^w do not have an ancestral order because AP^0 is not the ancestral population of AP^w, even though AP^0 diverged before AP^w. The ancestral order is transitive, that is, if $\mathrm{AP} \prec \mathrm{AP}^*$ and $\mathrm{AP}^* \prec \mathrm{AP}^{**}$ then $\mathrm{AP} \prec \mathrm{AP}^{**}$. The species tree topology is determined by the combination of all ancestral orders of populations. For example, the topology of the species tree in Figure 1 is determined by the ancestral order

$$S1 \prec \mathrm{AP}^0 \prec \mathrm{AP}^1, \quad S2 \prec \mathrm{AP}^0 \prec \mathrm{AP}^1,$$

$$S3 \prec \mathrm{AP}^W \prec \mathrm{AP}^1, \quad S4 \prec \mathrm{AP}^W \prec \mathrm{AP}^1.$$

Therefore, estimating the species tree topology can be equivalent to estimating the ancestral order of populations.

Consider the ancestral populations AP^0 and AP^1 in Figure 1 with ancestral order $\mathrm{AP}^0 \prec \mathrm{AP}^1$. The divergence times of the 2 populations are τ_0 and τ_1. Note that $\tau_0 < \tau_1$. Let s_a and s_b be the sequences from the 2 species (S1 and S2) whose most recent common ancestor is AP^0. The coalescence time of s_a and s_b is T^0. Let s_c and s_d be the sequences from the 2 species whose most recent common ancestor is AP^1. For example, the sequence s_c is sampled from species S1 and s_d is sampled from S3. The coalescence time of s_c and s_d is T^1. When s_a and s_b coalesce within the population AP^1, that is, $T^0 > \tau_1$, the coalescence times T^0 and T^1 have the same expectation because their distributions are identical to each other according to coalescent theory. When $T^0 \leqslant \tau_1$, the expected value of T^0 is smaller than that of T^1 because T^1 is always greater than τ_1. Overall, the expectation of T^0 is smaller than the expectation of T^1, that is, $E(T^0) < E(T^1)$. This result can be generalized to any 2 ancestral populations AP and AP^* with $\mathrm{AP} \prec \mathrm{AP}^*$(Appendix 1). Consider a set of ancestral populations along a lineage in an arbitrary species tree. We denote the ancestral populations as $\mathrm{AP}^0, \mathrm{AP}^1, \ldots, \mathrm{AP}^K$ by looking backward in time. Note these populations have the ancestral order $\mathrm{AP}^0 \prec \mathrm{AP}^1, \ldots, \prec \mathrm{AP}^K$. Let $E(T^i)$ denotes the expected coalescence time for population AP^i as defined above. It follows from Appendix 1 that $E(T^i) < E(T^j)$ for any $i < j$ and thus $E(T^0) < E(T^1) \cdots < E(T^K)$, which indicates that the order of expected coalescence times is consistent with the ancestral order of populations $\mathrm{AP}^0, \mathrm{AP}^1, \ldots, \mathrm{AP}^K$. Because the ancestral order of populations in the species tree is equivalent to the species tree topology, the tree of expected coalescence times has the same topology as that of the species tree, even though the expect coalescence times are biased estimators of the species divergence times. This statement applies equally to species trees inside and outside of the anomaly zone.

By inferring trees based on values of DNA sequence divergence from a distance matrix, phylogenetic methods such as UPGMA make use of estimates of species coalescence times whose distribution we have derived above. The combination of Equation (2) and the assumption of a molecular clock (Assumption 5) imply that any pair of species that have the same most recent common ancestor will have identical expectations of sequence divergences. Thus, the expected distances are additive, which suggests that the tree of expected coalescence times is ultrametric, because the expected distances are twice the expected coalescence times. Consequently, the UPGMA tree constructed from expected distances is identical to the tree of expected coalescence times and the species tree. In fact, other distance methods such as neighbor joining (NJ; Saitou and Nei 1987) also utilize the expected coalescence times and are predicted by our theory to produce correct unrooted species trees even in the anomaly zone. If the root of the species tree is known, the tree built by these distance methods is identical to the true rooted species tree. Hence, the UPGMA tree constructed from consistent estimates of the expected distances is a consistent estimate of the species tree topology, as is the NJ tree (or trees constructed by other distance methods) when the tree root is known. This result contrasts with the result found by Kubatko and Degnan (2007) who showed that ML was not a consistent estimator of the species tree in some situations.

To estimate the species tree topology using DNA sequences, we assume that sequences of multiple genes are generated independently from each ultrametric gene tree in the species tree under the Jukes–Cantor substitution model. Consider the 2 ancestral populations AP^0 and AP^1 in Figure 1. We use the same notation, s_a,s_b, s_c, s_d, T^0, and T^1, for the sequences and their coalescence times as defined in the previous paragraph. Given the coalescence time T^0, the probability that 2 sequences s_a and s_b have different nucleotides at a site under the Jukes–Cantor model is

$$p(T^0) = \frac{3}{4}(1 - e^{-8T^0/3}). \tag{3}$$

Integrating with respect to T^0, the probability p that 2 sequences are observed to be different at a site is equal to

$$p = \int_T p(T^0) f(T^0) \mathrm{d}\mathrm{T}^0 \tag{4}$$

in which $f(T^0)$ is a truncated exponential distribution defined in Equation (1). The probability p is just the expectation of $p(T^0)$ with respect to the distribution $f(T^0)$ or $p = E(p(T^0))$. In addition, the probability p^* that s_c and s_d are different at a site is $p^* = E(p(T^1))$.

We can show that $p < p^*$ (Appendix 2), which implies that the order of the expected proportions of sites having different nucleotides among sequences is identical to the order of ancestral populations $\mathrm{AP}^0 \prec \mathrm{AP}^1$. Similar to the arguments for the expected coalescence times, we therefore expect trees constructed by distance methods such as UPGMA and NJ, based as they are on the expected proportions of sites having different nucleotides between 2 sequences, to be consistent estimates of the species tree topology. In this case, estimation of the species tree becomes the problem of consistently estimating the order of expected proportions.

This suggests that under the assumptions used by Kubatko and Degnan (2007), namely a molecular clock and ultrametric species tree, distance methods should be able to produce consistent estimates of the species tree when DNA sequences are concatenated whether produced from trees within or outside of the anomaly zone. If DNA sequences across genes have equal length, the observed proportion of different sites between 2 concatenated sequences is equal to $\hat{p} = \frac{1}{m}\sum_{i=1}^{m} \hat{p}_i$ in which m

is the number of genes and $\hat{p}_i$ is the proportion of different sites between 2 sequences for gene i. By the law of large numbers, the proportion $\hat{p}$ is a consistent estimate of the expected proportion p of sites having different nucleotides between 2 sequences across m genes. We can also show (Appendix 3) that the observed proportion $\hat{p}$ consistently estimates p when lengths of genes vary. Hence, the proportion of sites observed to have different nucleotides between 2 concatenated sequences is a consistent estimate of p, the expected proportion of sites having different nucleotides between 2 sequences across genes. Under the Jukes–Cantor model, the distance between 2 sequences is equal to $-\frac{3}{4}\ln(1-\frac{4\hat{p}}{3})$, which is a monotonically increasing function of the proportion $\hat{p}$ indicating that the distance and the proportion $\hat{p}$ have the same order. Hence, tree construction methods that capitalize on the order of the distances among sequences should be able to produce a consistent estimate of the species tree topology when sequences evolve on genes according to a coalescent process even in the anomaly zone.

Simulations and Phylogenetic Analysis of Concatenated Sequences under Coalescence

The simulation conducted by Kubatko and Degnan (2007) showed that the ML method consistently estimated the wrong tree when the species tree was in the anomaly zone (Kubatko and Degnan 2007). In our simulations of the 4-taxon case, we confirmed the results of Kubatko and Degnan (2007) but also confirm that UPGMA can consistently estimate the anomalous species tree as expected from the theory we derived above (not shown). We note, however, that the "wrong" tree in the paper of Kubatko and Degnan (tree S1) is wrong only in the placement of the root; ignoring the root placement yields an unrooted tree that is not different from the true tree (their Matching Tree). Throughout their study, the estimated trees are rooted at the midpoint. Incorrect rooting at the midpoint may result in the wrong estimate of the rooted species tree (tree S1 in the paper of Kubatko and Degnan), a situation that many systematists would not consider "inconsistent."

Therefore, we specified a 5-taxon tree that was still in the anomaly zone (due to having short internal branches among the 4 in-group species) but that also had a long branch leading to an out-group species E (Fig. 2a). This tree was ((((A:0.05,B:0.05):0.005,C:0.055):0.005, D:0.06): 1.0,E:1.06). Simulating from a species tree of this shape ensured not only that the root would be correctly placed but also that the tree of the 4 in-group taxa is still in the anomaly zone because the most probable gene tree is (((A,B),(C,D)),E), which is different from the species tree (Fig. 2b and c). Gene trees matching the species tree occurred only 10.4% of the time or 2.4% less frequently than the most common gene tree (Fig. 2b and c). DNA sequences of 500 bp were generated from 100, 500, 1000, 2000, 4000, and 6000 gene trees simulated from this 5-taxon species tree using the program MCMCcoal (Rannala and Yang 2003). The concatenated sequences were analyzed in a phylogenetic program PAUP (Swofford 2003) to estimate phylogenetic trees using UPGMA, NJ, maximum likelihood with (MLK) and without (ML) a molecular clock, and maximum parsimony (MP) (these methods are reviewed in Nei 1987; Nei and Kumar 2000). The ML and MLK analyses were performed with an exhaustive search of all possible tree topologies. We also applied the traditional Bayesian method with and without a clock as implemented in MrBayes (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) using the Jukes–Cantor model. In this case, the chain was run for 100 000 generations, with samples taken every 100 generations. The first 100 trees were discarded as burn-in. We used the consensus tree built from the estimated posterior distribution as the Bayesian estimate of the species tree. The simulations with each phylogenetic method were repeated 100 times for each number of genes simulated.

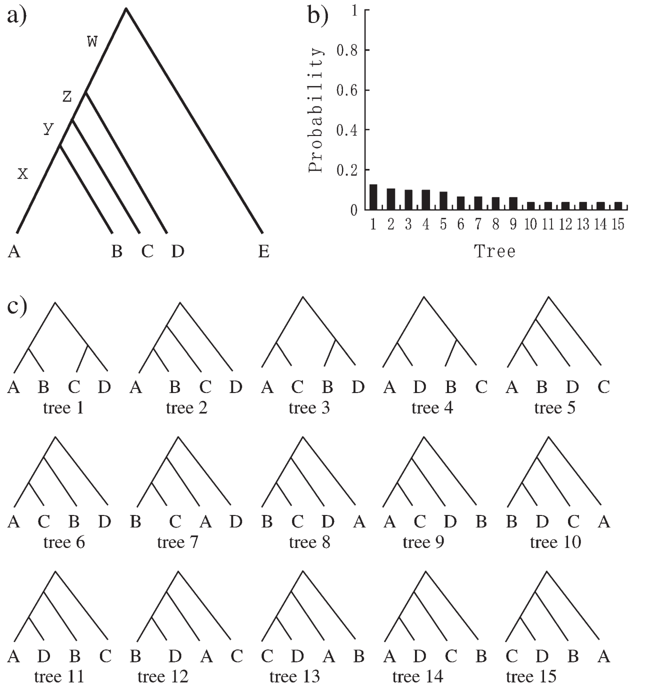

FIGURE 2. The 5-taxon species tree and the probability distribution of the gene trees generated from the species tree. a) The 5-taxon anomalous species tree used in simulations. The 5-taxon trees is ultrametric with population size of 0.1 for all populations. The lengths of branches are $x = 0.05$, $y = 0.005$, $z = 0.005$, and $w = 1.0$ in mutation units. The out-group is species E; b) the probability distribution of the 15 gene tree topologies generated from the species tree in (a); c) the list of the 15 gene tree topologies on the x-axis in (b). To simplify the notation, the out-group (species E) is not included in the trees. Tree 1 is the most probable gene tree, whereas Tree 2 is the gene tree that matches the species tree.

The trees constructed by the ML, NJ, MP, and Bayesian methods were rooted by the out-group E and compared with the true species tree in Figure 2. The result shows that the proportion of trials in which UPGMA, NJ, and MP can correctly estimate the species tree topology approaches 1.0, whereas the proportion of trials in which

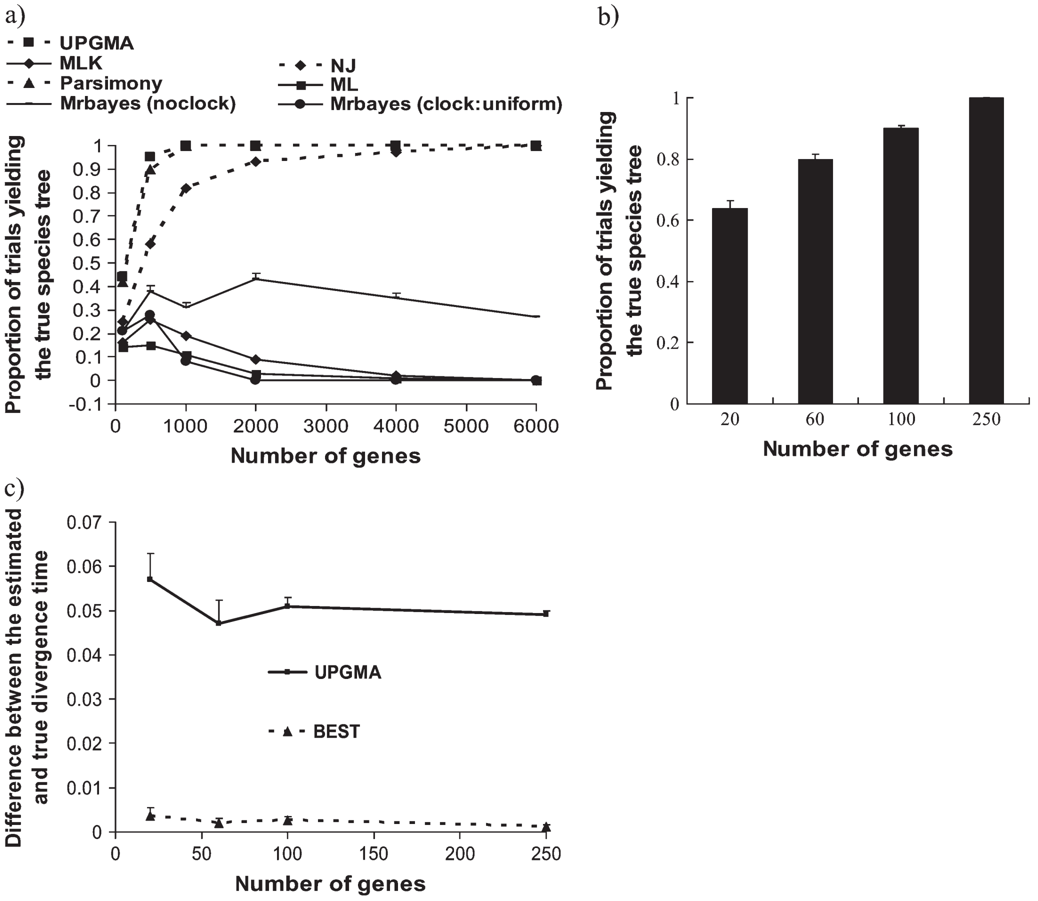

FIGURE 3. The performance of the tree construction methodologies in estimating the species tree topology. a) DNA sequences of 500 bp were generated from 100, 500, 1000, 2000, 4000, and 6000 gene trees simulated from the 5-taxon tree in Figure 2a and used to estimate the species tree using UPGMA, NJ, ML, MLK, Mrbayes(clock:uniform), and Mrbayes(noclock) methods; b) the proportion of trials that BEST estimates the 5-taxon species tree; c) the differences between the true divergence time of species A and C in the 5-taxon species tree and the estimates given by BEST and UPGMA. The heights of the vertical bars represent standard errors.

ML, MLK, and the Bayesian method with a clock estimate the true species tree topology approaches 0 as the number of genes increases (Fig. 3a). The simulation result for the distance approaches (UPGMA and NJ) confirms the theoretical result we have derived that distance approaches are statistically consistent in estimating species trees under the 5 assumptions described above. In addition, we found that MP was statistically consistent in this 5-taxon simulation. However, we also discovered that the MP method consistently estimated the wrong tree for sequences generated from the species tree (((A:0.05,B:0.05):0.005,(C:0.05,D:0.05):0.005):0.1,E: 0.155). This tree has 4 long terminal branches leading to taxa A–E but is not in the anomaly zone because the most probable gene tree matches the species tree. Although it is well known that parsimony is inconsistent for trees with long terminal branches in the Felsenstein zone and elsewhere (Felsenstein 1978; Huelsenbeck 1998), the behavior of parsimony has not been studied for species trees corresponding to these zones and under coalescence. Thus, our result extends the results from traditional gene tree studies to the case where gene trees vary on the species tree according to the coalescent.

To understand why MP performs well in the first 5-taxon simulation in the anomaly zone, the gene trees estimated by MP for 1000 genes were compared with the true gene trees generated from the 5-taxon species tree. The most common tree among the simulated gene trees is (((A,B),(C,D)),E) occurring 13.25% of the time, whereas 11.75% of the simulated gene trees have topology ((((A,B),C),D),E). By contrast, the most common MP tree reconstructed from the DNA sequences is ((((A,B),C),D),E), which occurs 14.8% of the time, whereas only 10.4% of the MP trees correspond to (((A,B),(C,D)),E). The bias of parsimony against the misleading gene tree (((A,B),(C,D)),E) appears to be due to long branch attraction (LBA) (Felsenstein 1978). This result is supported by the distribution of site patterns in 4-taxon subtrees for concatenated sequences under coalescence in the anomaly zone, which result in significantly more informative sites supporting ((B,C)(D,E)) than ((C,D)(B,E)) (Table 1). The true gene trees generated

TABLE 1. Informative sites for the 4 taxa: B, C, D, and E

Four-taxon subtree	Number of informative sites	Standard error	% of total informative sites
5 loci (−)			
(BC)(DE)[1]	38	4	24.20
(CD)(BE)	44	8	28.03
(BD)(CE)	75	9	47.77
5 loci (+)			
(BC)(DE)[1]	63	3	40.38
(CD)(BE)	43	2	27.56
(BD)(CE)	50	3	32.06
4000 loci			
(BC)(DE)[1]	33 441	446	38.11
(CD)(BE)	27 255	234	31.04
(BD)(CE)	27 082	342	30.85

Note: We generated 20 sets of DNA sequences of 500 bp from 5 gene trees and 10 sets of DNA sequences from 4000 gene trees simulated from the 5-taxon species tree in Figure 3. Among the 20 sets of DNA sequences, 10 data sets (followed by (+)) produced MP trees matching the species tree, whereas the other data sets (followed by (−)) produced the wrong tree (((A,B),(C,D)),E). There are 3 types of informative sites for the 4 taxa B, C, D, and E. The average number and standard error of such informative sites supporting a particular unrooted tree were calculated across the 10 samples.

[1] The 4-taxon subtree that is consistent with the true 5-taxon tree.

from the 5-taxon species tree are characterized by a long branch leading to the root and by terminal branch lengths that are all much longer than the internal branches in the in-group. The LBA we observe is not solely due to the long branch leading to the root because when we repeated the simulation with a tree that is still in the anomaly zone but with a shorter branch leading to the root ((((A:0.05,B:0.05):0.005,C:0.055):0.005,D:0.06):0.01,E:0.07), MP still consistently estimated the true species tree.

In the paper of Felsenstein (1978), the trees that suffer LBA have 2 short terminal branches and 2 long terminal branches. By contrast, all terminal branches in gene trees generated in these 5-taxon simulation are long branches, that is, there are no short terminal branches. However, LBA generally refers to the situation in which MP wrongly groups 2 long branches that actually belong to 2 different groups (Huelsenbeck 1998). In our first 5-taxon simulation, we found that MP sometimes wrongly supported the group (D, E) for the gene trees with topology (((A,B),(C,D)),E). Thus, LBA indeed occurred in our simulation when estimating such gene trees using MP, and there was a good chance that the out-group attached to the branch leading to species D, which results in a gene tree that is congruent with the true species tree. It appears that LBA sometimes cancels out signals in the data introduced by deep coalescence and result in the correct estimate of the species tree. Thus, it is plausible that MP performs well in our first 5-taxon simulation because parsimony has a bias against tree (((A,B),(C,D)),E) due to LBA.

The result for the Bayesian method without clock shows that proportion of the trials yielding the true species tree decreases as the number of genes increases (Fig. 3a). However, the proportions for the Bayesian method without clock are always greater than those for the ML method indicating that the Bayesian method performs better than ML in this 5-taxon simulation. The Bayesian estimate of the species tree is based on the estimated posterior distribution of tree topologies, marginalizing over branch lengths. By contrast, the ML tree is a single tree (topology and branch lengths) that can maximize the likelihood. The topology of the ML tree may not have the maximum posterior probability. We have found that in some cases, the ML tree found in the Markov Chain Monte Carlo (MCMC) chain in the Bayesian analysis matches the ML tree, but both are different from the maximum posterior probability tree. The most frequent result in this 5-taxon simulation corresponded to the situation in which the Bayesian estimate had topology ((((A,B),C),D),E), but the ML tree had topology (((A,B),(C,D)),E). This suggests that marginalization over branch lengths caused the Bayesian approach to outperform the ML method in this particular simulation (Fig. 3a). In other simulations on different species trees, we found that the nonclock Bayesian and ML methods performed equally (not shown). Thus, the effect of marginalization over branch lengths in Bayesian analysis causes unanticipated effects that may be reminiscent of the star tree paradox (Lewis et al. 2005; Yang and Rannala 2005).

We also conducted simulations to evaluate the performance of a new Bayesian method that utilizes a coalescence model to estimate the species tree for multilocus sequences data (Liu and Pearl 2007; Liu et al. 2008). It has been shown to perform well in situations with high gene tree heterogeneity and in which a low proportion of gene trees matches the species tree (Edwards et al. 2007), but it has not yet been tested for gene trees generated from species trees in the anomaly zone. We generated DNA sequences of 500 bp from 20, 60, 100, and 250 gene trees simulated from the 5-taxon species tree (Fig. 2a) with the Jukes–Cantor model. The DNA sequences were analyzed by BEST to infer species trees. The prior of θ was the inverse gamma distribution with $(\alpha = 3, \beta = 0.2)$. The simulation was repeated 100 times. The result shows that the proportion of trials yielding the true species tree approaches 1.0 as the number of genes increases, which suggests that the BEST method is consistent even when the species tree is in the anomaly zone (Fig. 3b). To compare the performance of the BEST and distance methods in the estimation of divergence times, the divergence times of species A and C were estimated by BEST and UPGMA, respectively. The divergence time estimates were compared with the true divergence time 0.055. The differences between the true divergence time and its estimates given by the BEST method are almost 0, whereas the differences are around 0.05 for the UPGMA method (Fig. 3c). This suggests that the BEST method is consistent in estimating divergence times in the anomaly zone, whereas the UPGMA method overestimates the divergence time by $\theta/2 = 0.05$ as the theory predicted.

Discussion

Under general conditions, the ML approach has been shown to be consistent for estimating gene trees, but recent simulation studies suggest that when applied to concatenated sequences from highly heterogeneous gene trees, ML may be inconsistent for estimating species trees (Kubatko and Degnan 2007). For concatenated sequences, the ML approach assumes homogeneous gene trees across loci. When the species tree is in the anomaly zone (Degnan and Salter 2005; Degnan and Rosenberg 2006), gene trees generated from the species tree are highly heterogeneous, a condition that seriously violates the homogeneous gene tree assumption and results in inconsistency of the ML approach. This violation has also been shown to mislead standard Bayesian methods (Mossel and Vigoda 2005). The ML method attempts to use a single tree to explain the DNA sequences generated from multiple gene trees. Because the majority of nucleotides evolve from the most probable gene tree, it appears that the ML method often chooses the most probable gene tree as the estimate of the phylogenetic tree (our simulations showed that ML indeed chose the most probable gene tree [the wrong tree] as the estimate of the species tree).

Our simulations also showed that distance and parsimony methods outperformed the model-based approaches such as the ML and Bayesian methods. In addition, a new Bayesian method using a coalescence model, which nonetheless relies on likelihood calculations (Liu and Pearl 2007; Liu et al. 2008), appears to perform well in the anomaly zone, as judged by our simulations in the 5-taxon case. Although we have proved theoretically that the UPGMA and NJ methods are consistent in estimating the species tree for concatenated sequences in and outside of the anomaly zone, the theoretical proof for the consistency of parsimony methods is challenging but appears to result from long branch attraction and parsimony being “right for the wrong reason” (see Results and Table 1).

The anomaly zone and associated “wicked forests” (Degnan and Rosenberg 2006; Rosenberg and Tao 2008) have been studied primarily with regard to gene tree topologies (Degnan et al. 2009) and less so with regard to branch lengths of gene trees. The 5-taxon tree (Fig. 2a) used in the simulation can produce anomalous gene trees because its short internal branches (y and z) increase the probability that the lineages A, B, C, and D coalesce in the population with branch length w and the coalescence process for the 4 lineages favors the symmetric tree over the asymmetric tree even though the lineages coalesce randomly. By contrast, the order of the coalescence times in the same population is totally random and the chance that 1 coalescence time is smaller than another is 0.5. For example, if the 3 sequences from species B, C, and D enter into the population with branch length w, the chance that the sequence from B coalesces with the sequence from C before it coalesces with the sequence from D is 0.5 (Fig. 2a). When the sequences from B and C coalesce in the population with branch length z, its coalescence time is definitely smaller than that of the sequences from B and D. This additional probability, even though it is very small, results in the order of the coalescence time of the sequences from B and C, and the sequences from B and D, being consistent with the ancestral order of the population with branch length w and the population with branch length z. Distance approaches work in the anomaly zone because they utilize the order information of the coalescence times when estimating species trees.

We have conducted our simulations only in the presence of a molecular clock, and so our findings, like those of Kubatko and Degnan (2007), are restricted to these conditions. In the absence of a clock, the expected distances may not be twice the expected coalescence times. Thus, the UPGMA tree constructed from the matrix of expected distances may not be identical to the tree of expected coalescence times, which implies that the distance methods may consistently estimate the wrong topology of the species tree when the assumption of a molecular clock does not hold. Furthermore, distance methods only use partial information of DNA sequences and they therefore may require more genes in order to achieve the same level of confidence for the estimate of the species tree as that found by model-based approaches. Nevertheless, distance methods are simple and consistent in the situations we have studied and therefore have the potential to estimate the species tree for large data sets where model-based methods may not be practical due to their requirement for intensive computation.

Funding

This research was supported by National Science Foundation grant DEB 0743616 to S.V.E. and Dennis Pearl.

Acknowledgments

We thank John Wakeley for comments on the distribution of coalescence times under the coalescence model. We thank Laura Kubatko for helpful discussion about 4-taxon trees in the anomaly zone. We thank Mike Steel and James Degnan for constructive comments on the first draft of the manuscript. We are grateful for the insightful comments of Jack Sullivan, Mark Holder, Paul Lewis, and 3 anonymous reviewers on the manuscript.

References

Degnan J.H., DeGiorgio M., Bryant D., Rosenberg N.A. 2009. Properties of consensus methods for inferring species trees from gene trees. Syst. Biol. 58:35–54.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:762–768.

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

Edwards S.V., Beerli P. 2000. Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. Evolution. 54:1839–1854.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104:5936–5941.

Efromovich S., Kubatko L.S., 2008. Coalescent time distributions in trees of arbitrary size. Stat. Appl. Genet. Mol. Biol. 7:Article2 1–28.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Gillespie J.H., Langley C.H. 1979. Are evolutionary rates really variable? J. Mol. Evol. 13:27–34.

Glazko G.V., Nei M. 2003. Estimation of divergence times for major lineages of primate species. Mol. Biol. Evol. 20:424–434.

Huelsenbeck J.P. 1998. Systematic bias in phylogenetic analysis: is the strepsiptera problem solved? Syst. Biol. 47:519–537.

Huelsenbeck J.P., Bull J.J., Cunningham C.W. 1996a. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

Huelsenbeck J.P., Bull J.J., Cunningham C.W. 1996b. Combining data in phylogenetic analysis—reply. Trends Ecol. Evol. 11:335–335.

Huelsenbeck J.P., Ronquist F. 2001. Mrbayes: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Jukes T.H., Cantor C.R. 1969. Evolution of protein molecules. In: Munro H.N. editor. Mammalian protein metabolism. New York: Academic Press. p. 21–123.

Kingman, J.F.C. 1982. On the genealogy of large populations. Stoch. Proc. Appl. 13:235–248.

Kingman, J.F.C. 2000. Origins of the coalescent: 1974–1982. Genetics. 156:1461–1463.

Kolaczkowski B., Thornton J.W. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.

Kubatko L.S., Degnan J.H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Syst. Biol. 56: 17–24.

Lewis P.O., Holder M.T., Holsinger K.E. 2005. Polytomies and Bayesian phylogenetic inference. Syst. Biol. 54:241–53.

Liu L., Pearl D.K. 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., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution. 62:2080–2091.

Liu L., Yu L., Pearl D.K., Edwards S.V. 2009. Estimating species phylogenies using coalescence times among sequences. Syst. Biol. doi: 10.1093/sysbio/syp031.

Mossel E., Roch S. 2007. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. IEEE/ACM Transactions on Computational biology and Bioinformatics. doi: 10.1109/TCBB.2008.66.

Mossel E., Vigoda E. 2005. Phylogenetic mcmc algorithms are misleading on mixtures of trees. Science. 309:2207–2209.

Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.

Nei M., Kumar S., 2000. Molecular evolution and phylogenetics. New York: Oxford University Press.

Nielsen R. 1998. Maximum likelihood estimation of population divergence times and population phylogenies under the infinite sites model. Theor. Popul. Biol. 53:143–51.

Rannala B., Yang Z.H. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.

Ronquist F., Huelsenbeck J.P. 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19:1572–1574.

Rosenberg N.A., Tao R. 2008. Discordance of species trees with their most likely gene trees: the case of five taxa. Syst. Biol. 57:131–140.

Saitou N., Nei M. 1987. The neighbor-joining method—a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4:406–425.

Seo, T.-K. 2008. Calculating bootstrap probabilities of phylogeny using multilocus sequence data. Mol. Biol. Evol. 25:960–971.

Swofford D.L. 2003. Phylogenetic analysis using parsimony. Sunderland (MA): Sinauer Associates.

Yang Z., Rannala B. 2005. Branch-length prior influences bayesian posterior probability of phylogeny. Syst. Biol. 54:455–470.

Received 7 October 2008; reviews returned 17 December 2008; accepted 6 April 2009
Associate Editor: Mark Holder

Appendix 1

Suppose AP and AP* are 2 ancestral populations in the species tree. Let T^{ab} be the coalescence time of the 2 sequences whose most recent ancestral population is AP and T^{cd} be the coalescence time of the 2 sequences whose most recent common ancestral population is AP* . If AP and AP* have an ancestral order then AP $\prec$ AP* or AP* $\prec$ AP if and only if $E(T^{\mathrm{ab}}) < E(T^{\mathrm{cd}})$ or $E(T^{\mathrm{cd}}) < E(T^{\mathrm{ab}})$.

Proof. we first prove that if AP $\prec$ AP* then $E(T^{\mathrm{ab}}) < E(T^{\mathrm{cd}})$. Since the coalescence time T^{ab} is either less than τ^* or greater than or equal to τ^*, the expectation of T^{ab} is

$$\begin{aligned} E(T^{\mathrm{ab}}) &= E(T^{\mathrm{ab}}|T^{\mathrm{ab}} \geqslant \tau^*) \times Prob(T^{\mathrm{ab}} \geqslant \tau^*) \\ &\quad + E(T^{\mathrm{ab}}|T^{\mathrm{ab}} < \tau^*) \times Prob(T^{\mathrm{ab}} < \tau^*). \end{aligned} \tag{5}$$

When $T^{\mathrm{ab}} > \tau^*$, the coalescence time T^{ab} and T^{cd} have the same distribution. Thus,

$$E(T^{\mathrm{ab}}|T^{\mathrm{ab}} > \tau^*) = E(T^{\mathrm{cd}}). \tag{6}$$

It follows that

$$E(T^{\mathrm{ab}}|T^{\mathrm{ab}} < \tau^*) < \tau^* < E(T^{\mathrm{cd}}). \tag{7}$$

Combine Equations (5), (6), and (7), we have

$$\begin{aligned} E(T^{\mathrm{ab}}) &= E(T^{\mathrm{cd}}) \times Prob(T^{\mathrm{ab}} \geqslant \tau^*) \\ &\quad + E(T^{\mathrm{ab}}|T^{\mathrm{ab}} < \tau^*) \times Prob(T^{\mathrm{ab}} < \tau^*) \\ &< E(T^{\mathrm{cd}}) \times Prob(T^{\mathrm{ab}} \geqslant \tau^*) \\ &\quad + E(T^{\mathrm{cd}}) \times Prob(T^{\mathrm{ab}} < \tau^*) = E(T^{\mathrm{cd}}). \end{aligned}$$

Next, we will prove that if $E(T^{\mathrm{ab}}) < E(T^{\mathrm{cd}})$ then AP $\prec$ AP* by contradiction. Suppose AP $\prec$ AP* is not true by which we have AP* $\prec$ AP because AP* is either the ancestral or the descendant population of AP under the assumption that 2 populations AP and AP* have ancestral order. Apparently, if AP* $\prec$ AP then $E(T^{\mathrm{ab}}) > E(T^{\mathrm{cd}})$, which contradicts the statement $E(T^{\mathrm{ab}}) < E(T^{\mathrm{cd}})$. The proof is complete. □

Appendix 2

Suppose AP and AP* are 2 ancestral populations in the species tree. Let T^{ab} and T^{cd} be the coalescence times as defined in Appendix 1. Let $p = E(p(T^{\mathrm{ab}}))$ and $p* = E(p(T^{\mathrm{cd}}))$, where $E(p(.))$ denotes the expected proportion of sites having different nucleotides among sequences. If gene trees are ultrametric trees and 2 populations AP and AP* have ancestral order, then AP $\prec$ AP* if and only if $p < p*$.

Proof: The expectation of $p(T^{\mathrm{ab}})$ is equal to $E(p(T^{\mathrm{ab}})=E(p(T^{\mathrm{ab}})|T^{\mathrm{ab}} > \tau^*) \times \mathrm{Prob}(T^{\mathrm{ab}} > \tau^*) + E(p(T^{\mathrm{ab}})|T^{\mathrm{ab}} \leqslant \tau^*) \times \mathrm{Prob}(T^{\mathrm{ab}} \leqslant \tau^*)$. Under the coalescence model, $p(T^{\mathrm{ab}})$ and $p(T^{\mathrm{cd}})$ have the same distribution if $T^{\mathrm{ab}} \geqslant \tau^*$. Thus,

$$E(p(T^{\mathrm{ab}})|T^{\mathrm{ab}} > \tau^*) = E(p(T^{\mathrm{cd}})).$$

Because $p(.)$ is a monotonically increasing function,

$$E(p(T^{\mathrm{ab}})|T^{\mathrm{ab}} \leqslant \tau^*) < E(p(\tau^*)) < E(p(T^{\mathrm{cd}})).$$

It follows that

$$\begin{aligned} p &= E(p(T^{\mathrm{ab}})) \\ &= E(p(T^{\mathrm{ab}})|T^{\mathrm{ab}} > \tau^*) \times Prob(T^{\mathrm{ab}} > \tau^*) \end{aligned}$$

$$+E(p(T^{\mathrm{ab}})|T^{\mathrm{ab}} \leqslant \tau^*) \times Prob(T^{\mathrm{ab}} \leqslant \tau^*)$$

$$< E(p(T^{\mathrm{cd}})) \times Prob(T^{\mathrm{ab}} > \tau^*)$$

$$+E(p(T^{\mathrm{cd}})) \times Prob(T \leqslant \tau^*) = E(p(T^{\mathrm{cd}})) = p*$$

The reverse statement can be easily proved by contradiction. The proof is complete.

Appendix 3

If sequence length L_is are independently identically distributed from a discrete distribution $\mathrm{Prob}(L = l) = P_l$, where $P_l > 0$ for $0 < l \leqslant C$ and otherwise $P_l = 0$, the observed proportion $\hat{p}$ is a consistent estimate of p.

Proof. To calculate the proportion of sites having different nucleotides for the concatenated sequences, we categorize the genes by their sequence length. There are C categories because the sequence length is less than or equal to C. Let m_j be the number of genes in category j in which the sequence length is equal to l_j. The proportion of sites having different nucleotides between 2 concatenated sequences is equal to

$$\hat{p} = \sum_{j=1}^{k} m_j l_j \hat{p}_j / \sum_{j=1}^{k} m_j l_j$$

in which $\hat{p}_j$ is the average proportion of sites having different nucleotides for genes in category j. Because the sequence lengths are equal across genes in each category, the law of large numbers applies and $\hat{p}_j$ converges to p in probability as $m_j \to \infty$. The number of genes in each category m_j will go to infinity when the total number of genes m goes to infinity because the genes are assigned to each category with a positive probability under the assumption $P_l > 0$ for $0 < l \leqslant C$. Thus, we have $\hat{p}_j \to p$ in probability as $m \to \infty$ for all j, which shows that as the number of genes $m \to \infty$, $\hat{p} = \sum_{j=1}^{k} m_j l_j \hat{p}_j / \sum_{j=1}^{k} m_j l_j \to p$ in probability. The proof is complete. □