AN ANALYTICAL COMPARISON OF MULTILOCUS METHODS UNDER THE MULTISPECIES COALESCENT: THE THREE-TAXON CASE

SEBASTIEN ROCH

Department of Mathematics
University of Wisconsin–Madison
Madison, WI
roch@math.wisc.edu

Incomplete lineage sorting (ILS) is a common source of gene tree incongruence in multilocus analyses. Numerous approaches have been developed to infer species trees in the presence of ILS. Here we provide a mathematical analysis of several coalescent-based methods. The analysis is performed on a three-taxon species tree and assumes that the gene trees are correctly reconstructed along with their branch lengths. It suggests that maximum likelihood (and some equivalents) can be significantly more accurate in this setting than other methods, especially as ILS gets more pronounced.

Keywords: Multispecies coalescent, incomplete lineage sorting, gene tree/species tree.

1. Introduction

Incomplete lineage sorting (ILS) is an important confounding factor in phylogenetic analyses based on multiple genes or loci.[1,2] ILS is a population-level phenomenon that is caused by the failure of two lineages to coalesce in a population, leading to the possibility that one of the lineages first coalesces with a lineage from a less closely related population. As a result, it can produce extensive gene tree incongruence that must be accounted for appropriately in multilocus analyses.[3]

A large number of methods have been developed to address this source of incongruence.[4] Several such methods rely on a statistical model of ILS known as the multispecies coalescent. In this model, populations are connected by a phylogeny. Independent coalescent processes are performed in each population and assembled to produce gene trees. Several methods have been shown to be statistically consistent under the multispecies coalescent, that is, they are guaranteed to return the correct species tree given enough loci.[5–8]

The performance and accuracy of coalescent-based multilocus methods have been the subject of numerous simulation studies.[5,9] In this paper, we complement such studies with a detailed analytical comparison in a tractable test case, a three-taxon species tree. We analyze 7 methods: maximum likelihood (ML),[10] GLASS/Maximum Tree (MT),[7,11] R^*,[12] STAR,[5] minimizing deep coalescences (MDC),[1] STEAC,[5] and shallowest coalescences (SC).[1] Under the assumption that gene trees are reconstructed without estimation error, we derive the exponential decay rate of the failure probability as the internal branch length of the species tree varies. The analysis, which relies on large-deviations theory, reveals that ML and GLASS/MT are more accurate in this setting than the other methods—especially in the regime where ILS is more common.

2. Materials and Methods

2.1. *Multispecies coalescent: Three-taxon case*

We first describe the statistical model under which our analysis is performed, the *multispecies coalescent*. We only discuss the three-taxon case. For more details, see Ref. 2 and references therein.

A weighted rooted tree is called ultrametric if each leaf is exactly at the same distance from the root. For a three-leaf ultrametric tree G with leaves a, b, and c, we denote by $ab|c$ the topology where a and b are closer to each other than to c, and similarly for $ac|b$, $bc|a$. The topology of G is denoted by $\mathcal{T}[G]$.

Let S be an ultrametric species phylogeny with three taxa. We assume that all haploid populations in S have population size N. We denote the current populations by A, B and C (which we identify with the leaves of S) and we assume that S has topology $\mathrm{AB}|\mathrm{C}$. The ancestral populations are AB (corresponding to the immediate ancestor to populations A and B) and ABC (corresponding to the ancestor of populations A, B and C). The corresponding divergence times (backwards in time from the present) are denoted by τ_{AB} and τ_{ABC} with the assumption $\tau_{\mathrm{AB}} \leq \tau_{\mathrm{ABC}}$. All times are given in units of N generations. For a population X, we let $\tau_{\mathrm{X}}^{\mathrm{P}}$ be the divergence time of the parent population of X. Let $\mathbb{X} = \{\mathrm{A}, \mathrm{B}, \mathrm{C}, \mathrm{AB}, \mathrm{ABC}\}$ be the set of all populations in S.

We consider L loci $\ell = 1, \ldots, L$ and, for each locus, we sample one lineage from each population at time 0. For locus ℓ, we denote by $I_{\mathrm{X}}^{(\ell)}$ the number of lineages entering population X and by $O_{\mathrm{X}}^{(\ell)}$ the number of lineages exiting population X (backwards in time), where necessarily $I_{\mathrm{X}}^{(\ell)} \geq O_{\mathrm{X}}^{(\ell)}$. Similarly, for $k = O_{\mathrm{X}}^{(\ell)} + 1, \ldots, I_{\mathrm{X}}^{(\ell)}$, the time of the coalescent event bringing the number of lineages from k to $k-1$ in population X and locus ℓ is $T_{\mathrm{X}}^{(\ell,k)}$. We denote by $G_1, \ldots, G_L$ the corresponding ultrametric gene trees (including both topology and branch lengths).

Then, under the multispecies coalescent, assuming the loci are unlinked, the likelihood of the gene trees is given by

$$f(G_1, \ldots, G_L | S) = \prod_{\ell=1}^{L} \exp\Bigg(- \sum_{\mathrm{X} \in \mathbb{X}} \Bigg\{ \binom{O_{\mathrm{X}}^{(\ell)}}{2} \left(\tau_{\mathrm{X}}^{\mathrm{P}} - T_{\mathrm{X}}^{(\ell, O_{\mathrm{X}}^{(\ell)}+1)} \right) - \sum_{k=O_{\mathrm{X}}^{(\ell)}+1}^{I_{\mathrm{X}}^{(\ell)}} \binom{k}{2} \left(T_{\mathrm{X}}^{(\ell,k+1)} - T_{\mathrm{X}}^{(\ell,k)} \right) \Bigg\} \Bigg), \tag{1}$$

where we let $T_{\mathrm{X}}^{(\ell, I_{\mathrm{X}}^{(\ell)}+1)} = \tau_{\mathrm{X}}$ for convenience.[13]

The parameter governing the extent of incomplete lineage sorting is the length of the internal branch of S

$$t = \tau_{\mathrm{ABC}} - \tau_{\mathrm{AB}}.$$

The probability that the lineages from A and B fail to coalesce in branch AB, an event we denote by $\textsc{Fail}_\ell$ for locus ℓ (and its complement by $\textsc{Success}_\ell$), is

$$1 - p = e^{-t}.$$

Note that, in that case, all three gene-tree topologies are equally likely. Of course, $1-p \to 1$ as $t \to 0$.

2.2. *Multilocus methods*

A basic goal of multilocus analyses is to reconstruct a species phylogeny (including possibly estimates of the divergence times) from a collection of gene trees. Here we assume that the data consist of L gene trees $G_1, \ldots, G_L$ corresponding to L unlinked loci generated under the multispecies coalescent. We assume further that the gene trees are ultrametric and that their topologies and branch lengths are estimated without error.

We consider several common multilocus methods. In our setting, several of these methods are in fact equivalent and we therefore group them below. Note further that we only consider statistically consistent methods, that is, methods that are guaranteed to converge on the right species phylogeny (at least its topology) as the number of loci L increases to $+\infty$ in the test case we described above. We briefly describe these methods. For more details, see e.g. Ref. 4 and references therein.

ML/GLASS/MT Under the multispecies coalescent, maximum likelihood (ML) selects the topology and divergence times that maximizes the likelihood (1). ML is implemented in the software package STEM.[10]

In the GLASS method,[7] the species phylogeny is reconstructed from a distance matrix in which the entries are the minimum gene coalescence times across loci. The equivalent Maximum Tree (MT) method was introduced and studied in Refs. 8,11,14.

A key result in Ref. 8 is that, in the constant-population case, the term inside the exponential in the likelihood (1) is monotonically decreasing in the divergence times. As a result, because GLASS and MT select the phylogeny with the largest possible divergence times, maximum likelihood is equivalent to GLASS and MT in this context. See Ref. 8 for details.

R^*/STAR/MDC In the R^* consensus method,[6,12] for each three-taxon set (here, we only have one such set), we include the topology that appears in highest frequency among the loci (breaking ties uniformly at random) and we reconstruct the most resolved phylogeny that is compatible with these three-taxon topologies.

In the STAR method,[5] the species phylogeny is reconstructed from a distance matrix in which the entries are the average ranks of gene coalescence times across loci. Here the root has the highest rank and the rank decreases by one as one goes from the root to the leaves.

The minimizing deep coalescences (MDC) method[1,15] selects the species phylogeny that requires the smallest number of "extra lineages," that is, lineages that fail to coalesce in a branch of the species phylogeny (breaking ties uniformly at random).

On a three-taxon phylogeny, there are only three distinct rooted topologies. In each case, the most recent divergence is assigned rank 1 in STAR and the other divergence is assigned rank 2. Hence selecting the topology corresponding to the lowest average rank is equivalent to selecting the most common topology among all loci—which is what R^* does. A similar argument shows that MDC also selects the R^* consensus tree in our test case.

Other common topology-based methods fall in this class, for instance, Rooted Triple Consensus.[16]

STEAC/SC In the STEAC method,[5] the species phylogeny is reconstructed from a distance matrix in which the entries are the average coalescence times across loci. The shallowest coalescences (SC)

method is similar to STEAC in that it uses average coalescence times. The difference between the two methods is in how they deal with multiple alleles per population. Since we only consider the single-allele case, the two methods are equivalent here.

2.3. *Large-deviations approach*

As mentioned above, we consider estimation methods that are statistically consistent in the sense that they are guaranteed to converge on the correct species phylogeny as the number of loci L increases to $+\infty$. To compare different methods, we derive the rate of exponential decay of the probability of failure. Let S be a species phylogeny with internal branch length t and assume that $G_1, \ldots, G_L$ are unlinked gene trees generated under the multispecies coalescent. As we explain next, large-deviations theory (see e.g. Ref. 17) allows us to compute the (exponential) decay rate

$$\alpha_{\mathbb{M}}(t) = -\lim_{L\to+\infty} \frac{1}{L} \ln \mathbb{P}[\text{Method } \mathbb{M} \text{ fails given } L \text{ loci from } S],$$

that is, roughly

$$\mathbb{P}[\text{Method } \mathbb{M} \text{ fails given } L \text{ loci from } S] \approx e^{-L\alpha_{\mathbb{M}}(t)},$$

for large L. As the notation indicates, the key parameter that influences the decay rate is the length of the internal branch t of the species phylogeny. In particular, we expect that $\alpha_{\mathbb{M}}(t)$ is increasing in t as a larger t makes the reconstruction problem easier.

To derive $\alpha_{\mathbb{M}}(t)$, we first need to express the probability of failure as a large deviation event of the form

$$\mathbb{P}[\text{Method } \mathbb{M} \text{ fails given } L \text{ loci from } S] = \mathbb{P}\left[\sum_{\ell=1}^{L} Y_\ell > yL\right],$$

where y is a constant and $\{Y_\ell\}_{\ell=1}^{L}$ are independent identically distributed random variables. The particular choice of random variables depends on the method, as we describe below. Let

$$\phi(s) = \mathbb{E}[e^{sY_\ell}],$$

be the moment-generating function of Y_ℓ (which does not depend on ℓ by assumption). Then large-deviations theory stipulates (see e.g. Theorem 2.6.3 in Ref. 17) that the decay rate is given by

$$\alpha_{\mathbb{M}}(t) = ys_* - \ln \phi(s_*), \tag{2}$$

where $s_* > 0$ is the solution (if it exists) to

$$\frac{\phi'(s_*)}{\phi(s_*)} = y,$$

provided there is an $s > 0$ such that $\phi(s) < +\infty$, $y > \mathbb{E}[Y_\ell]$ and Y_ℓ is not a point mass at $\mathbb{E}[Y_\ell]$.

3. Results: Derivations of decay rates

3.1. *A domination result*

We first argue that, given perfectly reconstructed unlinked gene trees under the multispecies coalescent, ML/GLASS/MT always has a greater probability of success than R^*/STAR/MDC and STEAC/SC—or, in fact, any other method. Indeed note that the probability of success can be divided into two cases:

(1) The case where $\textsc{Success}_\ell$ occurs for at least one locus ℓ, an event of probability $(1-(1-p)^L)$. In that case, ML/GLASS/MT necessarily succeeds whereas the other two methods succeed with probability < 1.
(2) The case where $\textsc{Fail}_\ell$ occurs for all loci ℓ, an event of probability $(1-p)^L$. In that case, all methods succeed with probability $1/3$ by symmetry. For instance, for ML/GLASS/MT, any pair of populations is equally likely to lead to the smallest inter-species distance. A similar argument applies to the other two methods.

Hence, overall ML/GLASS/MT succeeds with greater probability.

3.2. *Decay rates*

We derive the decay rates for the methods above. The results are plotted in Figure 1. The asymptotic regimes are highlighted in Figures 2 and 3. For lack of space, all proofs can be found in Ref. 18.

ML/GLASS/MT In this case, the decay rate can be derived directly without using (2). Following the derivation in Ref. 7 (see also Ref. 8 for a similar argument), ML/GLASS/MT fails with probability

$$1-\left[(1-(1-p)^L)+\frac{1}{3}(1-p)^L\right]=\frac{2}{3}(1-p)^L=\frac{2}{3}e^{-tL}.$$

Then we get the following:

Claim 3.1 (ML/GLASS/MT). *The decay rate of ML/GLASS/MT on S is*

$$\alpha_{\mathrm{ML}}(t)=t.$$

***R**/STAR/MDC** For a locus ℓ, we let $Z_{\mathrm{AB}}^{(\ell)}$ be 1 if $\textsc{Fail}_\ell$ occurs and $\mathcal{T}[G_\ell]=\mathrm{AB}|\mathrm{C}$, and 0 otherwise (where recall that $\mathcal{T}[G_\ell]$ is the topology of G_ℓ). We let

$$\mathcal{Z}_{\mathrm{AB}}=\sum_{\ell=1}^{L} Z_{\mathrm{AB}}^{(\ell)}.$$

Similarly, we define $Z_{\mathrm{AC}}^{(\ell)}$, $Z_{\mathrm{BC}}^{(\ell)}$, $\mathcal{Z}_{\mathrm{AC}}$ and $\mathcal{Z}_{\mathrm{BC}}$. Then R^*/STAR/MDC fails if

$$\mathcal{Z}_{\mathrm{AB}}+(L-\mathcal{Z}_{\mathrm{AC}}-\mathcal{Z}_{\mathrm{BC}}-\mathcal{Z}_{\mathrm{AB}})<\max\{\mathcal{Z}_{\mathrm{AC}},\mathcal{Z}_{\mathrm{BC}}\}.$$

It can be shown that

$$\alpha_{\mathrm{R}^*}(t)=-\lim_{L\to+\infty}\frac{1}{L}\ln\mathbb{P}[2\mathcal{Z}_{\mathrm{AC}}+\mathcal{Z}_{\mathrm{BC}}>L].$$

Then we get the following:

Claim 3.2 (R^*/STAR/MDC). *The decay rate of R^*/STAR/MDC on S is*

$$\alpha_{\mathrm{R}^*}(t)=-\ln\left(2\sqrt{\frac{1}{3}e^{-t}\left(1-\frac{2}{3}e^{-t}\right)}+\frac{1}{3}e^{-t}\right).$$

As $t\to 0$,

$$\alpha_{\mathrm{R}^*}(t)=\frac{3}{4}t^2+O(t^3),$$

and, as $t \to +\infty$*,*

$$\alpha_{\mathrm{R}^*}(t) \approx \frac{t}{2} - \frac{1}{2}\ln\frac{4}{3}.$$

STEAC/SC For a locus ℓ, we let $D_{\mathrm{AB}}^{(\ell)}$ be the time to the most recent common ancestor of A and B in G_ℓ (in unit of N generations). We let

$$\mathcal{D}_{\mathrm{AB}} = \sum_{\ell=1}^{L} D_{\mathrm{AB}}^{(\ell)}.$$

Similarly, we define $D_{\mathrm{AC}}^{(\ell)}$, $D_{\mathrm{BC}}^{(\ell)}$, $\mathcal{D}_{\mathrm{AC}}$ and $\mathcal{D}_{\mathrm{BC}}$. Then STEAC/SC fails if

$$\mathcal{D}_{\mathrm{AB}} > \min\{\mathcal{D}_{\mathrm{AC}}, \mathcal{D}_{\mathrm{BC}}\}.$$

It can be shown that

$$\alpha_{\mathrm{STEAC}}(t) = \lim_{L\to+\infty} -\frac{1}{L}\ln\mathbb{P}[\mathcal{D}_{\mathrm{AB}} - \mathcal{D}_{\mathrm{AC}} > 0].$$

Then we get the following:

Claim 3.3 (STEAC/SC). *The decay rate of STEAC/SC on* S *is*

$$\alpha_{\mathrm{STEAC}}(t) = -\ln\left(\frac{3e^{-s_* t} - s_*^2 e^{-t}}{3(1-s_*^2)}\right),$$

where $0 < s_* < 1$ *is the unique solution to the fixed-point equation*

$$s_* = \frac{1}{2}[6s_* - 3t(1-s_*^2)]e^{(1-s_*)t}.$$

Further, as $t \to 0$*,*

$$\alpha_{\mathrm{STEAC}}(t) = \frac{3}{8}t^2 + O(t^3),$$

and, as $t \to +\infty$*,*

$$\alpha_{\mathrm{STEAC}}(t) \approx t - \ln t - 0.1656.$$

4. Discussion

As can be seen from Figures 1 and 3 as well as from the asymptotics, ML/GLASS/MT does indeed give a larger decay rate for all t. In fact, the decay rate of ML/GLASS/MT is significantly higher, especially as $t \to 0$ that is, under high levels of incomplete lineage sorting. For instance, to be concrete, if $L = 500$ loci and $t = 0.1$ (in units of N generations), the probability of failure is approximately: 1.9×10^{-22} for ML/GLASS/MT; 0.038 for R^*/STAR/MDC; 0.16 for STEAC/SC. Intuitively, this difference in behavior arises from the fact that ML/GLASS/MT requires only *one* successful locus, whereas R^*/STAR/MDC and STEAC/SC rely on an *average* over all loci.

Comparing R^*/STAR/MDC and STEAC/SCin Figures 1, 2 and 3, note that $\alpha_{\mathrm{R}^*}(t)$ is higher than $\alpha_{\mathrm{STEAC}}(t)$ for small t but that the situation is reversed for large t. In fact, in the limit $t \to +\infty$, $\alpha_{\mathrm{STEAC}}(t)$ grows at roughly the same rate as $\alpha_{\mathrm{ML}}(t)$ (which is optimal by the domination result). At large t, STEAC/SC has somewhat of an advantage in that the expectation gap in the failure event increases linearly with t, whereas it saturates under R^*/STAR/MDC.

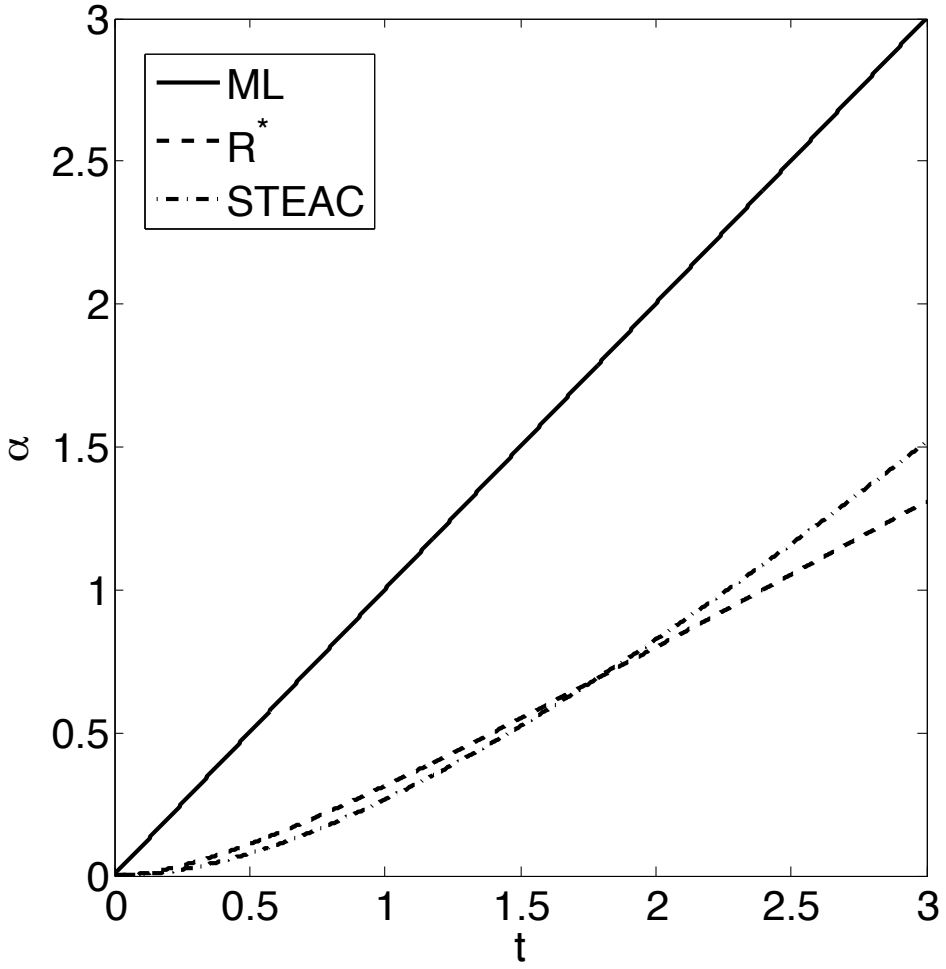

Fig. 1. Decay rates.

The analysis described here ignores several features that influence the accuracy of species tree reconstruction. Notably we have assumed that gene trees, including their branch lengths, are reconstructed without error. On real sequence datasets, the uncertainty arising from gene-tree estimation plays an important role. For instance, although GLASS/MT achieves the optimal decay rate in our setting, these methods are in fact sensitive to sequence noise because they rely on the computation of a minimum over loci—the very feature that leads to their superior performance here. See Ref. 5 for simulation results. Extending our analysis to incorporate gene tree estimation error is an important open problem which should help in the design of multilocus methods. It is important to note that, under appropriate modeling of sequence data, ML is *not* in general equivalent to GLASS/MT and comparing the sensitivity of these methods to estimation error is an interesting problem.

Other extensions deserve further study. Often many alleles are sampled from each population. Note that the benefit of multiple alleles is known to saturate as the number of alleles increases.[19] This is because the probability of observing any number of alleles at the top of a branch is uniformly bounded in the number alleles existing at the bottom.

Further, the molecular clock assumption, although it may be a reasonable first approximation in

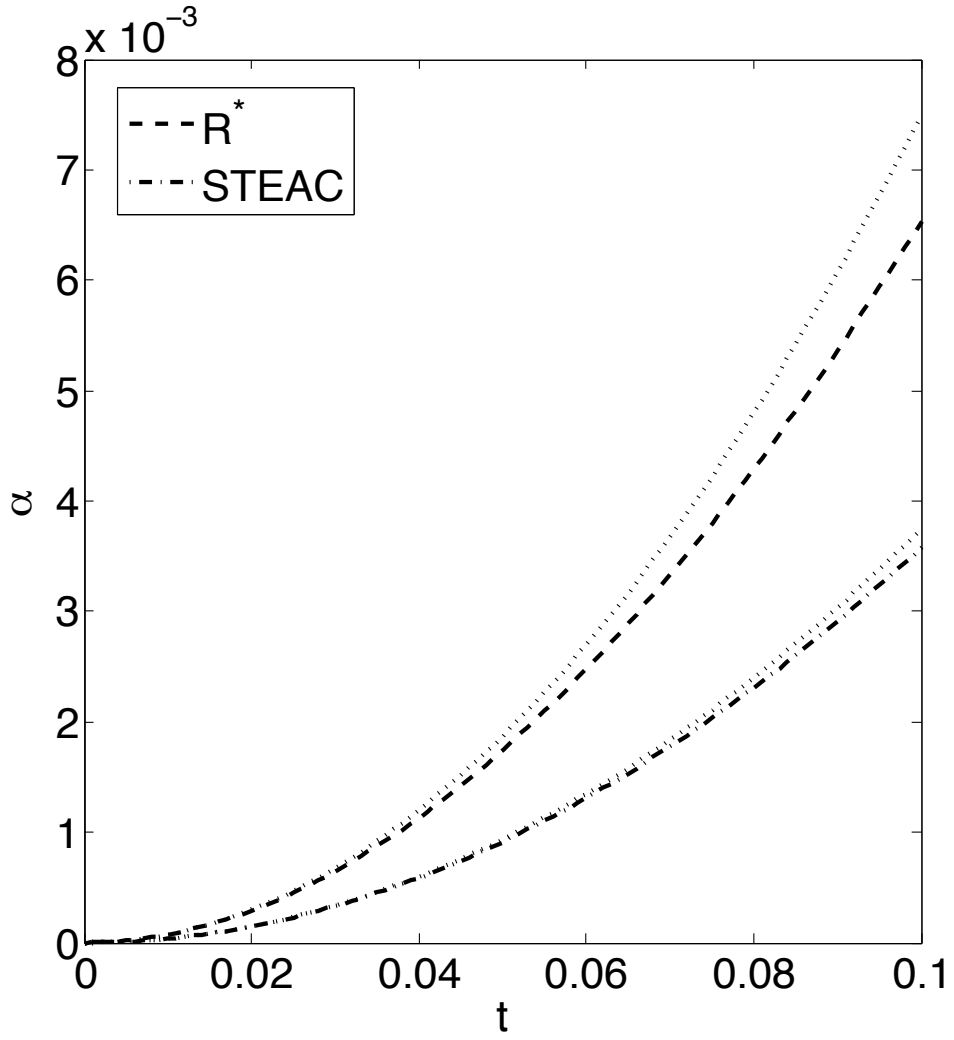

Fig. 2. Decay rates as $t \to 0$. The dotted lines indicate the respective predicted asymptotics. The decay rate for ML is not shown as it would be almost vertical.

the context of recently diverged populations, should not be necessary for our analysis. One should also consider larger numbers of taxa, varying population sizes, etc.

Simulation studies may provide further insight into these issues. However an analytical approach, such as the one we have used here, is valuable in that it allows the study of an entire class of models in one analysis. It can also provide useful, explicit predictions to guide the design of reconstruction procedures.

5. Supplementary Material

For lack of space, all proofs can be found in Ref. 18.

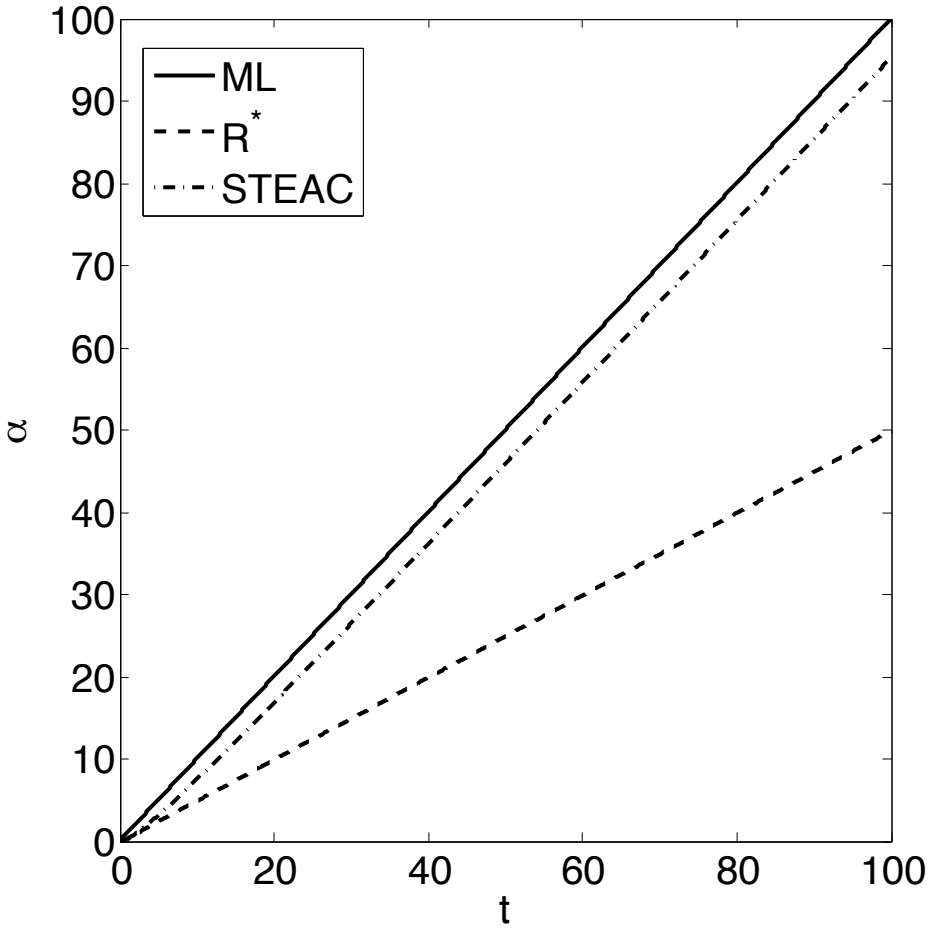

Fig. 3. Decay rates as $t \to +\infty$.

6. Acknowledgments

This work was supported by NSF grant DMS-1007144, NSF grant DMS-1149312 (CAREER), and an Alfred P. Sloan Research Fellowship. Part of this work was performed while the author was visiting the Institute for Pure and Applied Mathematics (IPAM) at UCLA.

References

1. W. P. Maddison, *Systematic Biology* **46**, 523 (1997).
2. J. H. Degnan and N. A. Rosenberg, *Trends in ecology and evolution* **24**, 332 (2009).
3. J. H. Degnan and N. A. Rosenberg, *PLoS Genetics* **2** (May 2006).
4. L. Liu, L. Yu, L. Kubatko, D. K. Pearl and S. V. Edwards, *Molecular Phylogenetics and Evolution* **53**, 320 (2009).
5. L. Liu, L. Yu, D. K. Pearl and S. V. Edwards, *Systematic Biology* **58**, 468 (2009).
6. J. H. Degnan, M. DeGiorgio, D. Bryant and N. A. Rosenberg, *Systematic Biology* **58**, 35 (2009).
7. E. Mossel and S. Roch, *IEEE/ACM Trans. Comput. Biology Bioinform.* **7**, 166 (2010).
8. L. Liu, L. Yu and D. Pearl, *Journal of Mathematical Biology* **60**, 95 (2010), 10.1007/s00285-009-0260-0.
9. A. D. Leaché and B. Rannala, *Systematic Biology* **60**, 126 (2011).
10. L. S. Kubatko, B. C. Carstens and L. L. Knowles, *Bioinformatics* **25**, 971 (2009).

11. L. Liu and D. K. Pearl, *Systematic Biology* **56**, 504 (2007).
12. D. Bryant, A classification of consensus methods for phylogenetics, in *Bioconsensus (Piscataway, NJ, 2000/2001)*, , DIMACS Ser. Discrete Math. Theoret. Comput. Sci. Vol. 61 (Amer. Math. Soc., Providence, RI, 2003) pp. 163–183.
13. B. Rannala and Z. Yang, *Genetics* **164**, 1645 (2003).
14. S. V. Edwards, L. Liu and D. K. Pearl, *Proceedings of the National Academy of Sciences* **104**, 5936 (2007).
15. C. Than and L. Nakhleh, *PLoS Comput Biol* **5**, p. e1000501 (09 2009).
16. G. Ewing, I. Ebersberger, H. Schmidt and A. von Haeseler, *BMC Evolutionary Biology* **8**, p. 118 (2008).
17. R. Durrett, *Probability: theory and examples*Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge Series in Statistical and Probabilistic Mathematics, fourth edn. (Cambridge University Press, Cambridge, 2010).
18. S. Roch, An analytical comparison of coalescent-based multilocus methods: The three-taxon case, Supplementary material available at http://arxiv.org/abs/1207.4074.
19. N. A. Rosenberg, *Theor. Popul. Biol.* **61**, 225 (March 2002).