J Mol Evol (1996) 42:308–312

JOURNAL OF MOLECULAR EVOLUTION

The Effect of Topology on Estimates of Among-Site Rate Variation

Jack Sullivan,* Kent E. Holsinger, Chris Simon

Department of Ecology and Evolutionary Biology, U-43, University of Connecticut, Storrs, CT 06269-3043, USA

Received: 27 April 1995 / Accepted: 1 September 1995

Abstract. Among-site rate variation, as quantified by the gamma-distribution shape parameter, *a* or α, and the ratio of transition rate to transversion rate (*Ts*/*Tv*) influence phylogenetic inference. We examine the effect of topology on estimates of these two parameters in 12S rRNA sequences from nine species of mice belonging to the genera *Onychomys* and *Peromyscus* by generating 100 random topologies and estimating these parameters using parsimony and maximum-likelihood methods for each of the random topologies. The parsimony-based estimate of *Ts*/*Tv* from the well-corroborated topology falls within the distribution of estimates based on random topologies, whereas the maximum-likelihood estimate of *Ts*/*Tv* based on the well-corroborated topology lies well outside the distribution of estimates derived from random topologies. The *Ts*/*Tv* ratio derived via maximum-likelihood estimation is three times the parsimony-based estimate, suggesting that parsimony-based estimates are severe underestimates even when the correct topology is used. Both parsimony- and likelihood-based estimates of the gamma-distribution shape parameter (α) are sensitive to topology because the best estimates based on the well-corroborated topology are well outside the distributions of estimates derived from random topologies for both methods. We show that the reason for topology dependence is the presence of long internal branches in the underlying topology.

Key words: Gamma shape parameter — Among-site rate variation — Phylogeny

* *Present address:* Laboratory of Molecular Systematics, MSC, MRC-534, Smithsonian Institution, Washington, DC 20560, USA
Correspondence to: J. Sullivan

Introduction

It has become increasingly obvious that among-site rate variation influences phylogenetic estimation from nucleotide sequence data and therefore must be incorporated in phylogenetic models (Kuhner and Felsenstein 1994; Tateno et al. 1994; Yang 1994; Yang et al. 1994; Sullivan et al. 1995). Yang et al. (1994) tested a series of models typically used in estimation of molecular phylogenies and made important contributions in showing that the use of complex models which incorporate nucleotide and substitution bias as well as rate variation among sites led to significant improvement in likelihood scores relative to use of more simple models. The model of among-site rate variation most widely used is the Γ-distributed rates model (e.g., Tateno et al. 1994; Gaut and Lewis 1995; Sullivan et al. 1995). The important parameter of this model is the shape parameter (*a* or α), which is inversely related to the coefficient of variation. Low values of α, therefore, imply substantial rate heterogeneity across sites. A shape parameter of 0.5 has been used by Tateno et al. (1994) and Gaut and Lewis (1995) to model extreme among-site rate variation, but estimates derived from real data sets are often significantly lower (e.g., Kocher and Wilson 1991).

Yang et al. (1994) found little variation in estimates of α derived from different topologies for several real four- and five-taxon data sets. If it is true generally that estimates are not topology dependent, this provides an important escape from the impossible situation of trying to incorporate tree-based estimates of these parameters into models used for estimating tree topology. However, Yang (1994) subsequently showed that basing estimates of α on a star topology can lead to significant overesti-

mates of among-site rate variation (underestimates of α) in a larger data set. It is therefore important to assess the variation in tree-based estimates of α over a wide range of alternative topologies and to ascertain under what conditions estimates will vary across topologies.

In this paper we examine variation in the estimates of the Γ-distribution shape parameter (α) and the ratio of transition rate to transversion rate (*Ts*/*Tv*) over a wide range of tree space by generating a distribution of these tree-based estimates over 100 random trees for some of the sequences examined in Sullivan et al. (1995). We then compare the best estimates of these parameters derived from the well-corroborated relationships among these taxa to the distribution of estimates from random topologies.

Methods

The data set used here consists of nine 12S rRNA sequences (775 bp) from grasshopper mice (*Onychomys*) and deer mice (*Peromyscus*), a subset of those presented in Sullivan et al. (1995). Divergence levels range up to ca. 13% among the sequences, and the relationships among these mice are well understood based on congruence among morphological, allozyme, chromosomal, and DNA hybridization and sequencing studies. Further, it has been shown that the mitochondrial DNA gene tree is equivalent to the species tree for these samples (Sullivan et al. 1995).

MacClade (Maddison and Maddison 1992) was used to generate 100 random trees for these nine taxa using the equiprobable trees option. Transition/transversion ratio (*Ts*/*Tv*) was estimated for each of these trees by direct count of changes inferred for each topology under the assumption of parsimony, and by the maximum-likelihood method of Yang (1994) using Baseml and the F84 model for each of the topologies. Wakeley (1994) pointed out a relationship between among-site rate variation and *Ts*/*Tv*. To examine this relationship in our data we estimated *Ts*/*Tv* with Baseml using both a single rate model and a Γ-distributed rates model.

Shape parameters were calculated for each of the random topologies using the discrete gamma option of Baseml (Yang 1994) with 40 categories and the F84 model. In addition, MacClade was used to generate parsimony-based distributions of the number of changes inferred at each site for each of the random topologies and Γ-distribution shape parameters (α) were estimated by finding the maximum-likelihood fit to the negative binomial using the program GAMMA (Sullivan et al. 1995). Tree-based estimates of α derived from the well-corroborated relationships of these taxa using both of the above methods were compared to the distributions of estimates generated by using the random topologies.

To test for the effect of long internal branches, we took three subsets of four taxa such that two subsets contained a long internal branch (between *Onychomys* and *Peromyscus*) and one contained a short internal branch. Subset A included *O. leucogaster, O. torridus, P. leucopus,* and *P. eremicus;* in subset B, *P. keeni* replaced *P. leucopus;* subset C contained *O. leucogaster, P. eremicus, P. keeni,* and *P. leucopus.* We then estimated α for all three possible topologies plus the star topology for all three data subsets using Baseml as above (Yang 1994). Estimates of α derived from the alternative topologies were fixed in Baseml and likelihood scores were calculated using the well-corroborated topology. The likelihood scores were then compared using likelihood-ratio tests.

Results and Discussion

Ts/Tv *Estimates*

The estimate of *Ts*/*Tv* derived using parsimony and direct counts from the well-corroborated relationships among these taxa is 1.75 and lies well within the distribution of parsimony-based estimates from random topologies (Fig. 1A). However, as pointed out by Wakeley (1994), tree-based parsimony estimates of *Ts*/*Tv* are typically conflated with among-site rate variation and thus are underestimates when there is a large amount of rate heterogeneity. The underestimation of *Ts*/*Tv* due to parsimony is uniform across the random topologies in this data set.

The maximum-likelihood estimate of *Ts*/*Tv* based on the well-corroborated relationships when among-site rate variation is accommodated (using F84 and discrete gamma options of Baseml) is 6.04, much higher than any of the parsimony-based estimates (Fig. 1). This estimate lies well outside the distribution of estimates derived from random topologies (Fig. 1B). However, when among-site rate variation is ignored (Fig. 1C) the estimate of *Ts*/*Tv* from the well-corroborated topology is well within the distribution of estimates from the random topologies. This suggests that accurate maximum-likelihood estimation of *Ts*/*Tv* is topology dependent when among-site rate variation is accommodated.

The difference between the results for parsimony-based estimates (Fig. 1A) and maximum-likelihood estimates (Fig. 1B) is due to the bias in estimation of *Ts*/*Tv* from direct counts of changes inferred on a tree using parsimony. The magnitude of this bias can be seen by the displacement of the distribution in Fig. 1A relative to Fig. 1B and C. Any estimate of the underlying ratio of rate of transitional substitutions to rate of transversional substitutions derived by counting observable substitutions on a tree is a severe underestimate. When the more accurate of the maximum-likelihood methods is used to estimate *Ts*/*Tv* (incorporating among-site rate variation) the effect of using an incorrect topology becomes apparent (Fig. 1B).

Estimates of Shape Parameters

The parsimony-based estimate of α derived (using GAMMA) from the well-corroborated relationships clearly falls outside the distribution of estimates based on random trees (Fig. 2A). Similarly, the Baseml estimate derived from the well-corroborated topology falls outside the distribution of estimates from random topologies (Fig. 2B). The estimate derived from the star phylogeny falls in the middle of the distribution, indicating that estimates based on the star topology are inadequate. We

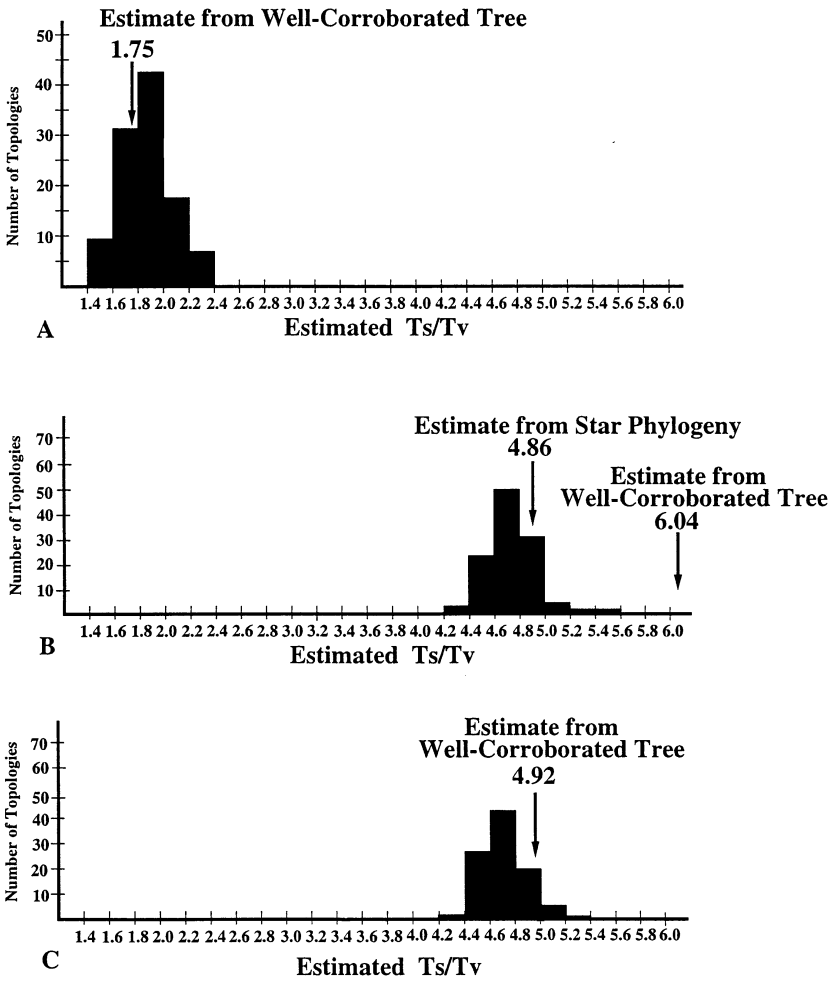


Fig. 1. **A** The distribution of parsimony-based estimates of *Ts*/*Tv* for 100 random topologies. Variation in 12S rRNA sequences among nine taxa of *Peromyscus* and *Onychomys* was optimized on each topology using MacClade and estimates were based of direct counts of changes on each tree. **B** The distribution of maximum-likelihood estimates of *Ts*/*Tv* for 100 random topologies when among-site rate variation is accommodated. **C** The distribution of maximum-likelihood estimates across topologies when a single rate is assumed across all sites.

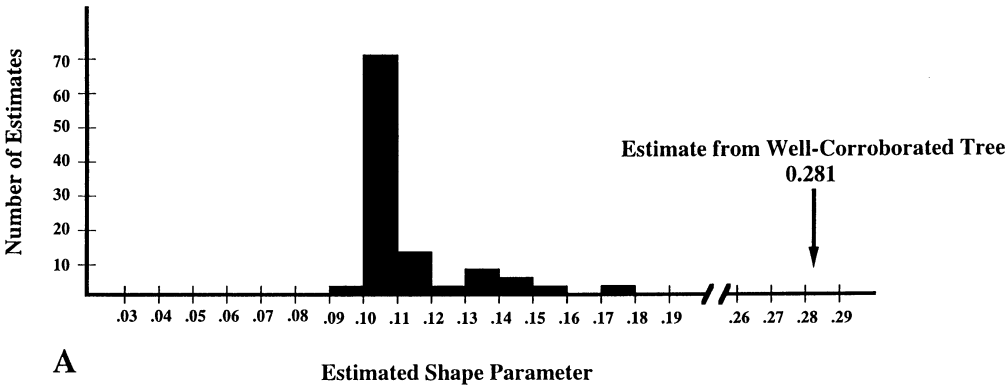
interpret these results to indicate that topology has a dramatic effect on estimates of α regardless of whether Yang's (1994) method or the method of Sullivan et al. (1995) is used; estimates of α are clearly topology dependent for this data set.

The Effect of Long Internal Branches

In each of the five small data sets (four or five taxa) examined by Yang et al. (1994) there is a very short internal branch. In our data set, the genera *Peromyscus* and *Onychomys* are separated by a relatively long internal branch. In our three four-taxon subsets, two (Fig. 3A,B) contained a long internal branch and one did not (Fig. 3C). When we used Baseml (Yang 1994) to estimate α for all three alternatively resolved topologies plus the star topology for each subset, the estimates of α from the true topology (α_1) for subsets A and B (long internal branch maintained) were significantly higher (as judged by likelihood ratio tests) than the estimates derived using any of the other topologies (α_2, α_3) including the star topology (α_0). In data subset C, where there was a very short internal branch, the estimates of the shape parameter were not significantly different across the four topologies (Fig. 3).

Interestingly, in data subset C, the maximum-likelihood tree unites *P. eremicus* with *P. leucopus* (Table 1) and the likelihood score was significantly higher than that for the star topology. Although the likelihood-ratio test is difficult to interpret in this case, this resolution conflicts with relationships based on all data examined to date for these taxa, including data from linked cytochrome *b* (Sullivan et al. 1995). Thus, maximum-likelihood estimation fails to find the well-corroborated relationships among *Peromyscus* species with the 12S data, even when among-site rate variation and unequal substitution probabilities and base composition are included in the model of evolution.

The conclusion of Yang et al. (1994) that there is little variation in estimates of α across topologies was based primarily on four-taxon data sets. Upon inclusion of additional taxa, Yang (1994) found that there is indeed variation in estimates of α across topologies. Specifically, adding a gibbon (*Hylobates*) sequence to the primate mtDNA data set generated a long internal branch



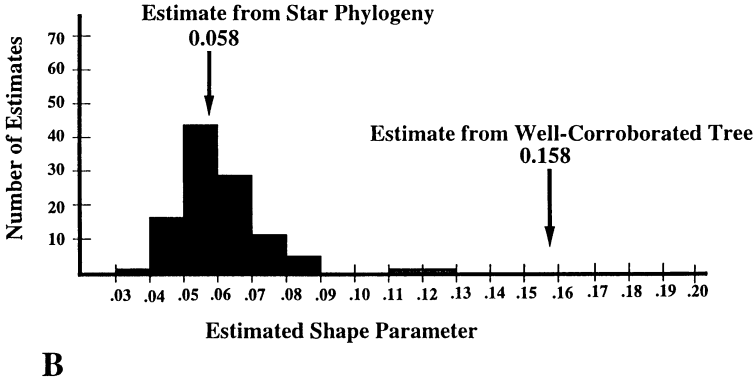


Fig. 2. **A** The distribution of shape parameter estimates (α) calculated by the GAMMA program for 100 random topologies. Note that the x-axis is not continuous. **B** The distribution of shape parameter estimates (α) calculated by the maximum-likelihood method of Baseml for 100 random topologies.

Subset	Topology	a_1	a_0	a_2	a_3
A	*O. leucogaster*, *P. leucopus*, 0.052, *O. torridus*, *P. eremicus*	0.284	0.032***	0.032***	0.032***
B	*O. leucogaster*, *P. keeni*, 0.057, *O. torridus*, *P. eremicus*	0.270	0.046***	0.054***	0.048***
C	*O. leucogaster*, 0.0001, *P. keeni*, *P. leucopus*, *P. eremicus*	0.143	0.143	0.186	0.152

*** Significantly different from a_1 at P < 0.001

Fig. 3. The effects of long internal branch on estimates of α. For all three subsets, topology 1 is the well-corroborated topology, topologies 2 and 3 represent the alternative resolutions, and topology 0 represents the star phylogeny. a_x represents estimates of the shape parameter for each of the above topologies. Significance tests were conducted by comparing the likelihood score for topology 1 and a_1 with the likelihood score calculated using topology 1 and each a_x, $\chi^2 = 2(l_1 - l_x)$.

between *Homo-Pan-Gorilla* and *Pongo-Hylobates.* Topologies that did not maintain this bipartition of taxa produced underestimates of α. This is consistent with the results seen here (Fig. 3).

To further examine the importance of long internal branches, we estimated α from the nine-taxon data set using a topology with only one bipartition of taxa, that separating *Onychomys* from *Peromyscus.* All intrage-

Table 1. Likelihood scores of three trees for data subset C of Fig. 3[a]

Topology	*l*	χ^2 vs tree 0
0) (*Ole, Ple, Per, Pke*)	−1,435.005	
1) ((*Ole, Per*), *Ple, Pke*)	−1,435.007	0.004
2) ((*Ole, Pke*), *Ple, Per*)	−1,433.064	3.887*

* $P < 0.05$

[a] Topology 0 is the star topology, topology 1 represents the well-corroborated relationships among these taxa, and topology 2 is the maximum-likelihood and maximum parsimony tree (Sullivan et al. 1995). Topology 2 has a higher likelihood score than the star topology, whereas topology 1 does not. Standard likelihood ratio tests are applicable when testing a resolved topology vs star topologies because the star topology is the same as any resolved topology, but with zero-length internal branches

neric relationships remained as unresolved star topologies. The estimate of α for this topology was 0.125 (compared to a value of 0.158 when the well-corroborated relationships were used, Fig. 2B) and a likelihood-ratio test showed that these estimates are statistically indistinguishable. Thus, in this data set, using any topology that maintains the bipartition of taxa into their appropriate genera will generate a reasonable estimate of α.

Accommodating Among-Site Rate Variation

One potential limitation of using maximum-likelihood estimation of parameters in complex (realistic) models is the intensive computation required. This is especially problematic for large data sets. Yang (1994) suggested using subsets of taxa to obtain maximum-likelihood estimates directly from the sequence data. The results of the taxon subsampling here (Fig. 3) suggest that there is a taxon sampling effect. The estimates of the shape parameter produced by Baseml for two of the four-taxon subsets (i.e., subset A, $\alpha_1 = 0.284$) are no closer to the Baseml estimates from the whole data set (Fig. 2B; $\alpha = 0.158$) than is the estimate of α derived by the method of Sullivan et al. (1995), which is based on the entire data set (Fig. 2A; $\alpha = 0.281$). However, a very good estimate of α is obtained from the four taxa in subset C (Fig. 3; $\alpha_1 = 0.143$). Thus, in this data set, some subsets of taxa provide accurate estimates of α, whereas others do not. We are currently examining the effect of data set size and taxon subsampling on estimates of *a* using large (ca. 40 taxa) data sets.

Topology clearly has a dramatic effect on estimates of the Γ-distribution shape parameter, regardless of the method used in estimation, if the history of the included taxa involves long internal branches. Fortunately, long internal branches are easy to resolve using virtually any phylogenetic inference method. It is therefore possible to conduct an initial phylogenetic analysis based on a simple model of evolution to identify long internal branches, incorporate them into estimates of among-site rate variation, and then refine phylogenetic analyses using more complex models of nucleotide substitution. This approach is conceptually very similar to successive approximations (Farris 1969) and dynamic parsimony (Williams and Fitch 1989) in that an initial tree is used to evaluate character variability. Phylogenetic analysis then becomes a recursive procedure.

Acknowledgments. We would like to thank Dr. Z. Yang for providing his Baseml programs and comments on the manuscript. Drs. J. Wakeley and D. Pollock also provided comments that improved the manuscript. This study was conducted while J.S. was supported by an NSF Graduate Research Traineeship in the evolution, ecology, and conservation of biodiversity (BIR-9256616). This study was supported by NSF grant BSR-8822710 to C.S.

References

Farris JS (1969) A successive approximations approach to character weighting. Syst Zool 18:374–385

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162

Kocher TD, Wilson AC (1991) Sequence evolution of mitochondrial DNA in human and chimpanzees: control region and protein coding region. In: Osawa S, Honjo T, (eds) Evolution of life: fossils, molecules, and culture. Springer, Tokyo, pp 391–413

Kuhner MK, Felsenstein J (1994) A simulation of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Maddison WP, Maddison DR (1992) MacClade: analysis of phylogeny and character evolution. Version 3.0. Sinauer Associates, Sunderland, MA

Sullivan J, Holsinger KE, Simon C (1995) Among-site rate variation and phylogenetic analysis of 12S rRNA in Sigmodontine rodents. Mol Biol Evol 12:988–1001

Tateno Y, Takezaki N, Nei M (1994) Relative efficiencies of the maximum likelihood, neighbor joining, and maximum parsimony methods when substitution rate varies with site. Mol Biol Evol 11:261–277

Wakeley J (1994) Substitution-rate variation among sites and the estimation of transition bias. Mol Biol Evol 11:426–442

Williams PL, Fitch WM (1990) Phylogeny determination using the dynamically weighted parsimony method. Methods Enzymol 183: 615–627

Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39:306–414

Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11:316–324