Syst. Biol. 53(3):485–495, 2004

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150490445779

Modeling Compositional Heterogeneity

PETER G. FOSTER

Department of Zoology, The Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom;
E-mail: p.foster@nhm.ac.uk

Abstract.—Compositional heterogeneity among lineages can compromise phylogenetic analyses, because models in common use assume compositionally homogeneous data. Models that can accommodate compositional heterogeneity with few extra parameters are described here, and used in two examples where the true tree is known with confidence. It is shown using likelihood ratio tests that adequate modeling of compositional heterogeneity can be achieved with few composition parameters, that the data may not need to be modelled with separate composition parameters for each branch in the tree. Tree searching and placement of composition vectors on the tree are done in a Bayesian framework using Markov chain Monte Carlo (MCMC) methods. Assessment of fit of the model to the data is made in both maximum likelihood (ML) and Bayesian frameworks. In an ML framework, overall model fit is assessed using the Goldman-Cox test, and the fit of the composition implied by a (possibly heterogeneous) model to the composition of the data is assessed using a novel tree- and model-based composition fit test. In a Bayesian framework, overall model fit and composition fit are assessed using posterior predictive simulation. It is shown that when composition is not accommodated, then the model does not fit, and incorrect trees are found; but when composition is accommodated, the model then fits, and the known correct phylogenies are obtained. [Compositional heterogeneity; Markov chain Monte Carlo; maximum likelihood; model assessment; model selection; phylogenetics.]

Markov process models used for phylogenetic analysis of DNA sequences have become more realistic. The simple Jukes-Cantor model has been extended to take into account unequal nucleotide composition, different rates of change from one nucleotide to another, and among-site rate variation in the form of a proportion of invariant sites, and discrete gamma-distributed rates of variable sites (Swofford et al., 1996; Whelan et al., 2001). We want to reflect the important features of evolution in our models, without adding unimportant parameters that would increase the variance and decrease the power of our conclusions. Although a badly fitting model does not guarantee obtaining the wrong tree, not accommodating important features is a primary reason for failure to get the correct topology (e.g., Sullivan and Swofford, 1997).

The process of evolution can differ over the tree. Trivially, overall rates of evolution can differ over the tree, and we often see fast and slow lineages. The rate of evolution of individual sites can change over time as well, presumably reflecting changes in functional constraints (Penny et al., 2001; Lopez et al., 2002). That we see different compositions in the terminal taxa tells us that the compositional part of the process of evolution also differs over the tree; the present study focuses on this. When compositional heterogeneity is pronounced, then accuracy of phylogenetic methods is compromised (Mooers and Holmes, 2000; Lockhart et al., 1992; Lake, 1994; Foster and Hickey, 1999; Tarrío et al., 2001).

The process of evolution can also differ over the data. Phylogenetic models that differ over the data were developed by Yang (1996). In these models, data were divided into site classes, each of which could have their own overall rates, transition-transversion ratios, and among-site rate heterogeneity. This class of models is appropriate for combined analysis of genes with different evolutionary dynamics, and also for data such as the three different positions of codons, where typically the third position evolves much faster than the other two. A version of this model has been implemented in PAUP* as the site-specific rates model, and in MrBayes as the site-specific model and the site-specific gamma model (Swofford, 2002; Huelsenbeck and Ronquist, 2001). For the present study I have reimplemented this class of models, allowing overall rates, among-site rate variation, composition, and rate matrix to differ in different data partitions.

In a maximum likelihood (ML) analysis we can compare models with a likelihood ratio test or with the Akaike information criterion (AIC) (Huelsenbeck and Rannala, 1997; Akaike, 1974; Posada and Crandall, 1998). There are, however, tests that address the question of whether the model fits the data in an absolute sense, asking whether the chosen model is appropriate for the data at hand (Yang et al., 1994). We can test for overall model fit in an ML context using the Goldman-Cox test (Goldman, 1993; Whelan et al., 2001). The fit of the composition of the model to the composition of the data can be tested separately. For this, since models in common use are homogeneous, the χ^2-test for compositional homogeneity is often used, even though it is widely recognized that this test does not take into account tree-based correlation of compositions among taxa. In the present study a novel tree- and model-based composition fit test is proposed, which asks whether the composition of the (possibly heterogeneous) model fits the composition of the (possibly heterogeneous) data. Bayesian methods such as posterior predictive simulation offer new possibilities for tests for adequacy of the model (Gelman et al., 1995). In a Bayesian framework we do not need to test the fit on a particular tree as is the case with the Goldman-Cox test, as the tree topology is considered another nuisance parameter over which the analysis is integrated (Huelsenbeck et al., 2001).

We seldom know the true tree in phylogenetic analyses; however, in this study I have chosen two examples where we can have confidence in the true tree, and thereby test the methods. The first example uses

bacterial 16S genes, where convergent compositional attraction causes unrelated lineages to group together (Embley et al., 1993; Mooers and Holmes, 2000). In the second example outgroup rooting in *Xdh* genes from *Drosophila* is incorrect or unstable because the compositions of the outgroup and the ingroup differ (Tarrío et al., 2000). It is telling that distance-based analyses using LogDet/Paralinear distances (Steel, 1994; Lockhart et al., 1994; Lake, 1994) alleviate both of these problems. However, using these distances is not a panacea for compositional heterogeneity problems (Foster and Hickey, 1999; Tarrío et al., 2001). Additionally, using LogDet/Paralinear distances does not allow us to calculate the likelihood, and so we cannot use likelihoods to compare results using this approach with results using Markov models.

Models in current use are for the most part tree-homogeneous; however, some models have been proposed that allow the composition to differ over the tree (Yang and Roberts, 1995; Galtier and Gouy, 1998; Galtier et al., 1999). In these the composition of the model differs on each branch of the tree, or on each terminal branch. In the Yang and Roberts N2 model the composition of each nucleotide on each branch is estimated, and so would have three free parameters per branch. The Yang and Roberts N1 model is less parameter-rich in that all the internal branches are given a single composition. In the Galtier and Gouy model the composition is described by the GC content, and so has only one composition parameter per branch. These approaches do not scale well, and in large trees may result in over-parameterization. Here I approach the analysis of bacterial 16S genes and *Xdh* genes from *Drosophila* using novel heterogeneous models that accommodate compositional differences over the tree. These models do not require that each branch get its own composition, and so can accommodate compositional heterogeneity with few additional parameters. This would accommodate those cases when compositional differences are localized in the tree, rather than differing continuously over the tree. Phylogenetic analysis, including placement of composition vectors on the tree, is done using Markov chain Monte Carlo (MCMC) methods. When there is more than one composition vector, both the composition parameters and the placement of the composition vectors on the tree are free parameters. The fit of the model to the data is examined using both ML and Bayesian methods. It is shown that when composition is not accommodated, then the model does not fit, and erroneous trees are found; but when composition is accommodated, the model then fits, and the known correct phylogenies are obtained.

METHODS

Phylogenetic Model

I model character change as a continuous-time Markov process. Using notation similar to that given in Swofford et al. (1996), a model can be fully described by its instantaneous rate matrix $\boldsymbol{Q}$, where $\boldsymbol{Q}$ can be decomposed to $\boldsymbol{Q} = \boldsymbol{R\Pi} - \mathrm{diag}(\boldsymbol{R\Pi}[1])$, where [1] is a column vector of 1's, where the diagonal elements of $\boldsymbol{R}$ are set to zero, and where the diagonal matrix $\boldsymbol{\Pi}$ is diag($\boldsymbol{\pi}$), where $\boldsymbol{\pi}$ is the vector of the proportions of the character states in the model. Therefore we have two different instantaneous rate matrices—$\boldsymbol{R}$ without the composition, and $\boldsymbol{Q}$ with the composition. In this study the rate matrix $\boldsymbol{R}$ is symmetrical, implying time-reversibility. I model compositional heterogeneity over the tree by allowing more than one composition vector while having a tree-wide $\boldsymbol{R}$ matrix. In such cases, due to composition differences, there will be more than one $\boldsymbol{Q}$ on the tree, but each $\boldsymbol{Q}$ taken by itself is reversible. However, because the overall model is tree-heterogeneous, the analysis as a whole is not reversible, and the likelihood depends the position of the root (Yang and Roberts, 1995). The present study places the root of the tree on internal trifurcating nodes as an approximation to the location of the biological root, expected to be on internal bifurcating nodes.

I allow different model parameters over the data. There may be more than one data partition, and each of the data partitions can have different models, with different rate matrices, among-site rate variation, and (perhaps heterogeneous) composition. The rate of evolution of the data partitions may differ, and so the rate of each data partition is a free parameter, constrained such that the overall rate of the data is 1. Although there may be fast and slow data partitions, branch lengths in the different data partitions are constrained to be proportional to each other.

The probability matrix $\boldsymbol{P} = e^{\boldsymbol{Q}\nu}$ for a given branch in the tree depends on the branch length ν. The likelihood of a site in the alignment of DNA sequences can be calculated from the $\boldsymbol{P}$ matrices for a tree using Felsenstein's pruning algorithm (Felsenstein, 1981). The likelihood of a data partition is the product of the site likelihoods, and the likelihood of the data as a whole is the product of the likelihoods of the data partitions. I allow for among-site rate heterogeneity using either a free proportion of invariant sites, or allowing the sites to have different rates in a discrete gamma distribution with a free α shape parameter, or a combination of these.

Bayesian Phylogenetic Analysis

The posterior probability of the trees and model parameters given the data was approximated using MCMC methods. The Metropolis-coupled MCMC variant was used, with four chains, as it gives good mixing of parameters and tree topologies (Huelsenbeck and Ronquist, 2001). Flat priors were used throughout. The MCMC started with a random tree, and was allowed to run until well after the likelihood values of the chain reached a plateau. Convergence was checked by comparing the consensus of different ranges of sampled trees and parameters after burn-in.

Move types included "Local" nearest neighbor interchange (NNI) as described in (Larget and Simon, 1999). This move type adjusts both the topology and branch lengths. The proposal ratio is the square of the branch length multiplier (Larget and Simon, 1999). Proposals to

change the rate matrix $\boldsymbol{R}$, the gamma shape parameter α, the proportion of invariant sites, and the relative rate of the data partitions were drawn uniformly from a window centered on the current state. If the proposed state was less than a minimum or more than a maximum then the difference was reflected off the limit. The proposal ratios are 1 for each of these move types. Changes to composition parameters were made by drawing from a Dirichlet distribution centered on the current state, for which the proposal ratio is the ratio of the two Dirichlet densities as described in Larget and Simon (1999).

Two new move types are proposed here. The composition can be allowed to differ over the tree, by allowing the tree to have more than one composition vector. In this implementation the number of composition vectors is fixed for each analysis (an obvious improvement would be to allow the number of composition vectors to be free, but that was not done here). Each branch is associated with a node distal to the root, and at any point in the MCMC each node is associated with one of the available composition vectors. One new move type is to allow randomly chosen nodes to choose at random from one of the other available composition vectors. The position of the root can affect the likelihood of the tree, and therefore the posterior probability calculations. As mentioned above, in this study the root is placed on an internal, trifurcating node, and so the second new move type allows the tree to root on a different, randomly chosen internal trifurcating node. The proposal ratio is 1 for both of these move types.

Assessment of Model Fit

In a maximum likelihood framework, I use Goldman's version of the Cox test to assess overall adequacy (Goldman, 1993; Whelan et al., 2001). Also in ML, I use a tree- and model-based composition fit test, described here, to assess fit of the composition implied by the model to the data.

Tree- and model-based composition fit test.—I begin by examining the widely used χ^2 test for compositional homogeneity. This test uses a contingency or $R \times C$ table of the compositions of the taxa against the mean composition. This test uses a statistic X^2 (in the sense used by Sokal and Rohlf [1981], *cf* χ^2, which is a distribution), and uses a null distribution $\chi^2_{df=(R-1)\times(C-1)}$ to assess significance. It is commonly appreciated that this test does not take into account correlation due to relatedness of the taxa on a tree (this caveat is printed out when this test is done using PAUP*; see also Rzhetsky and Nei [1995]). The extent of this problem is shown in Figure 1, where a true null distribution was generated by simulation on a four-taxon tree. When the sequences are unrelated (approximated here by setting branch lengths to 10 mutations per site) and all sites are free to vary, the χ^2 null distribution is valid (Fig. 1a). However, if there are invariant sites or the sequences are related (branch lengths 0.1 mutations per site), then the χ^2 distribution is no longer valid, and can have a large probability of type II error. This test can be made valid by assessing significance using a null distribution from X^2 statistics from simulations based on the tree and model being tested (for example the distribution in Fig. 1d), rather than using χ^2 to assess significance (D. L. Swofford and J. Sullivan, personal communication).

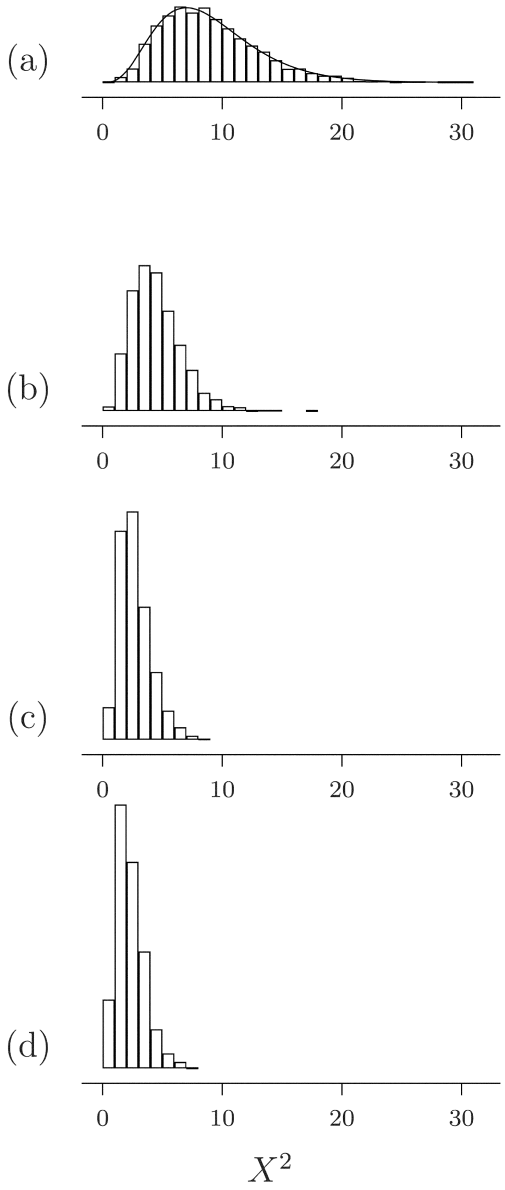

FIGURE 1. Type II error in the χ^2 compositional homogeneity test. Bars show distributions of the statistic X^2 for the composition χ^2 test generated by simulation of DNA sequences on a four-taxon tree with equal branch lengths. Simulations used the Jukes-Cantor model of evolution (Jukes and Cantor, 1969). (a) Branch lengths were set to expected 10.0 mutations per site, effectively making the sequences unrelated. The curve is $\chi^2_{df=9}$. (b) Same as (a), but half of the sites were held invariant in the simulation. (c) Branch lengths were set to expected 0.1 mutations per site, with all sites free to vary. (d) Same as (c), but half of the sites were held invariant. Note that an observed X^2 statistic of 10 would not reject compositional homogeneity based on the χ^2 curve, but that same statistic would be significant using the simulation histograms.

Because the expected composition comes from the mean observed composition, the test for compositional homogeneity described above (however its significance is assessed) can be considered a test for model fit only in the case of homogeneous models with empirical composition; it is not an appropriate test for models that allow compositional heterogeneity over the tree. The compositional homogeneity test can be considered a special case of a more general test, a tree- and model-based composition-fit test, introduced here. Rather than asking whether the data are compositionally homogeneous, we can instead ask whether the possibly heterogeneous data fit the possibly heterogeneous model. The compositional homogeneity test uses the test statistic $X^2 = \sum[(obs - exp)^2/exp]$, where the expected values are from the mean composition of the data. The composition fit test described here uses a similar statistic X_m^2, where the m subscript indicates that the expected values come from the model, not from the data. Expected values can differ in different taxa.

The composition implied by the model is calculated on the tree being tested starting from the root. The root model composition is given. The composition for nodes above the root can be obtained directly from the composition of the parent of the node and the probability matrix $\boldsymbol{P}$ of the model for the branch leading to the node, taking into account among-site rate variation. By iterating from the root to the terminal nodes of the tree in this way the model composition can be calculated and used as the expected value in calculating the X_m^2 statistic. A null distribution of X_m^2 values is made with simulations, to which the realized X_m^2 value from the original data can be compared. Generally the model parameters and branch lengths of the simulations need to be optimized by ML. The realized value is considered significant if it is larger than 95% of the null distribution.

Posterior predictive simulation.—Model fit was assessed in Bayesian analyses using a posterior predictive distribution of a test quantity $T(\cdot)$, which was approximated by simulations during the MCMC (Gelman et al., 1995; Huelsenbeck et al., 2001; Bollback, 2002). Model fit was measured by tail-area probability p_t

$$p_t = \frac{1}{N}\sum_{i}^{N} I\big(T(\boldsymbol{X}_i) \geq T(\boldsymbol{X})\big)$$

where N is the number of samples i taken during the MCMC, at which the model parameters are θ_i, which are used to simulate data sets $\boldsymbol{X}_i$. The test quantity from the simulated data set is compared to the test quantity of the original data set $\boldsymbol{X}$. I is an indicator function, which is 1 when the relation is true, and 0 otherwise. The two test quantities suggested in Huelsenbeck et al. (2001) were measured: The multinomial, or unconstrained, likelihood is used to assess overall model fit (Bollback, 2002; Goldman, 1993), and the composition X^2 statistic described above is used to assess fit of the model composition.

Phylogenetic Software

Phylogenetic analyses used PAUP* version 4.0b10, MrBayes version 2.01, and ModelTest version 3.06 (Swofford, 2002; Huelsenbeck and Ronquist, 2001; Posada and Crandall, 1998). Simulations, ML and Bayesian analyses with heterogeneous models, posterior predictive simulations, SH tests, and model fit tests used p4 (http://www.nhm.ac.uk/zoology/external/p4.htm). The Shimodaira-Hasegawa (SH) RELL test was done as described in Goldman et al. (2000).

RESULTS

Compositional Heterogeneity in Bacterial 16S Genes

I begin with a reanalysis of a problematic data set of bacterial 16S genes (Embley et al., 1993; Mooers and Holmes, 2000). Our best hypothesis for the true tree is that *Deinococcus* and *Thermus* group together to the exclusion of *Bacillus*, *Thermotoga*, and *Aquifex* (Fig. 2a). *Deinococcus* and *Thermus* share the same peptidoglycan and menaquinone type (Murray, 1991). Also, signature sequences, and phylogenetic analysis based on Ef-Tu, Hsp70, and RecA group *Deinococcus* with *Thermus* (Gupta, 1998; Eisen, 1995).

This alignment is remarkable because it gives incorrect results with most phylogenetic methods. *Deinococcus* and *Thermus* should group together, but instead we find a grouping of the two mesophiles *Deinococcus* and *Bacillus* (Fig. 2b). The erroneous grouping of these two taxa is consistent with the compositions shown in Table 1, where the three thermophiles are GC-rich, but the two

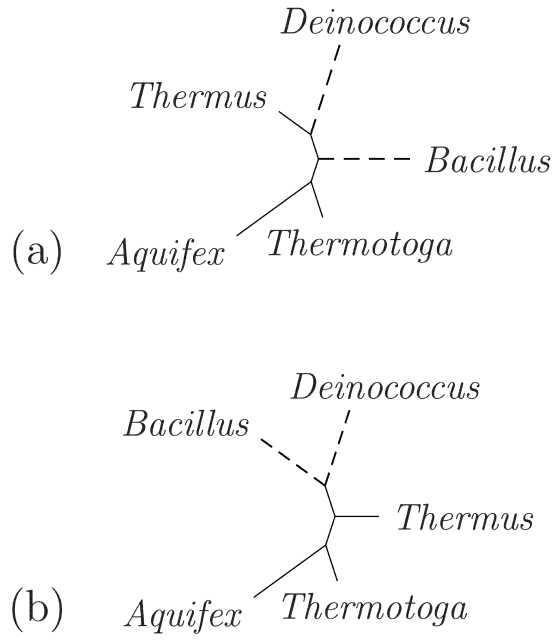

FIGURE 2. Compositional attraction in bacterial 16S sequences. The branches leading to the two mesophiles are indicated by dashed lines. (a) Based on other evidence (see text), this can be considered the correct tree. (b) The tree obtained with most phylogenetic methods, the "attract" tree.

TABLE 1. Composition in bacterial 16S genes.

	A	C	G	T	
Aquifex	0.21	0.29	0.36	0.14	thermophiles
Thermotoga	0.21	0.28	0.36	0.15	
Thermus	0.22	0.28	0.36	0.15	
Deinococcus	0.25	0.23	0.32	0.19	mesophiles
Bacillus	0.25	0.24	0.32	0.19	

mesophiles are less so. The compositional heterogeneity among lineages is even more pronounced if we only look at variable sites. *Deinococcus* and *Bacillus* appear to "attract" each other due to their shared compositional bias. Indeed, a tree made solely from Euclidean distances between compositions of the taxa is as in Figure 2b. Additionally, if *Thermus ruber* (now *Meiothermus ruber*), which has a compositional bias intermediate between the thermophiles and the mesophiles in Table 1, is added to the analysis in addition to *Thermus aquaticus* used in this example, then *Deinococcus* groups with *Thermus* (Embley et al., 1993). Also, *Deinococcus* and *Thermus* group together using a larger number of 16S sequences (Eisen, 1995).

Maximum Likelihood Model Choice, Phylogenetic Analysis, and Assessment of Model Fit

The five sites with alignment gaps were excluded from analysis, leaving 1287 sites. The GTR + Γ (general time-reversible rate matrix, with discrete gamma-distributed among-site rate variation) model was chosen using ModelTest, and the maximum likelihood (ML) tree for this model (Fig. 2b) was found using an exhaustive search. The ln L difference between the true tree and the attract tree is 1.8, which is not significant by the Shimodaira-Hasegawa (SH) RELL test ($P = 0.44$). This tree fails the tree and model-based composition-fit test (see Methods; $P < 0.005$), and also the data fail the χ^2 compositional homogeneity test. Additionally, this tree and model fail the Goldman-Cox test (Fig. 3; $P < 0.005$), a test for overall adequacy of the model (Goldman, 1993).

Because the data are compositionally heterogeneous, it is reasonable to model them using a heterogeneous model, and I begin with an N1-like model (Yang and Roberts, 1995). In its original implementation the N1 model is a derivative of the HKY85 model (Hasegawa et al., 1985) with gamma-distributed rates, and a tree-wide κ transition-transversion parameter. In the N1 model each terminal branch is given its own optimizable composition vector. The internal branches all together are given another optimizable composition vector, and the root given another. In this nonstationary model, the likelihood is affected by the position of the root. The present implementation is similar except that a tree-wide GTR + Γ model is used, with a trifurcating root (i.e., the likelihood is evaluated with the tree rooted at an internal, trifurcating, node), without a separate root composition. When there is no separate root composition in this model, the position of the root among the three possible internal roots does not affect the likelihood. Evaluating only the

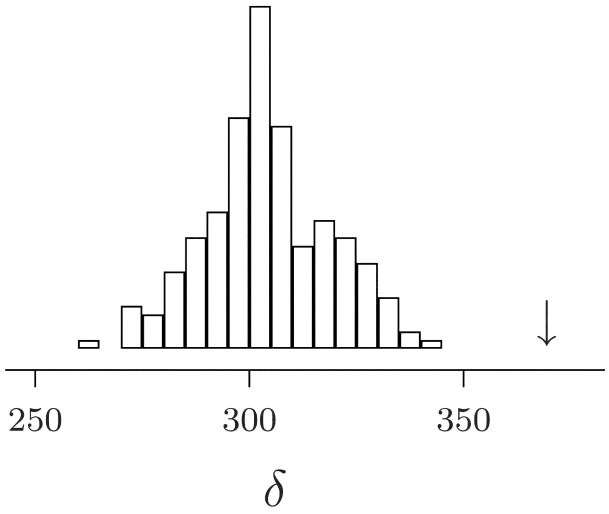

FIGURE 3. Goldman-Cox test for the homogeneous model. The ML tree under a homogeneous GTR + Γ model (Fig. 2b) was tested. The statistic δ is the difference between the optimized log likelihood and the unconstrained or multinomial log likelihood of the data. The arrow shows the position of δ for the original data. The bars are a null distribution generated by 200 simulations on the ML tree, each followed by optimization of branch lengths and model parameters. Because the statistic for the original data is well outside of the null distribution, the ML tree fails this test ($P < 0.005$), showing that this homogeneous model does not adequately describe the data.

two trees in Figure 2, the ML tree is now the true tree, and the log likelihood of the attract tree is now worse by 19.1, a difference which is significant ($P = 0.02$) by the SH test. Under this model, the ML tree passes both the tree and model-based composition-fit test ($P = 0.96$) and the Goldman-Cox test for overall adequacy of the model ($P = 0.25$), both of which the data failed with a homogeneous model.

Compared to the homogeneous model, this model has 15 additional parameters, and gives an increase in log likelihood of 90.7 using the true tree. Because these are nested models, the significance of this increase can be evaluated using the χ^2 approximation to the null distribution of twice the log likelihood ratio, with degrees of freedom equal to the difference in the number of free parameters. By this test, the increase is highly significant ($P \approx 0$).

However significant, this increase in model fit is at a cost of many extra parameters, and may be more than are needed to adequately model these data. Because the compositions fall into two groups (Table 1), we can ask whether a two-part model can adequately model these data. We place the composition vectors on the two trees using the solid/dashed line pattern in Figure 2 such that the branches leading to the mesophiles are given one optimizable composition and the remaining branches (and root) are given another. Using this model, the ML tree is again the true tree, the attract tree is significantly worse by the SH test ($P = 0.03$), and the model passes both the tree and model-based composition-fit test ($P = 0.86$) and the Goldman-Cox test ($P = 0.09$). Compared to the

homogeneous model, the two-part model has only three additional parameters, but gives an increase in log likelihood of 84.2, which is again highly significant ($P \approx 0$). If we compare the two-part with the N1-like model, the increase in log likelihood is only 6.5 at a cost of 12 additional parameters, a difference which is not significant ($P = 0.37$). It can be concluded that although both of these heterogeneous models recover the true tree, and both adequately model the data, the additional parameters of the N1-like model are in this case not worth the cost.

Placement of Composition Parameters on the Tree

In the example above, deciding which branches are associated with the two composition vectors in the tree-heterogeneous model was done by inspection. The arrangement of the compositions on the tree was as shown in Figure 2, where the branches shown by dashed lines get one composition, and the remaining branches get another. There are other plausible arrangements for the compositions on the two trees, examples of which are shown in Figure 4. There are many other possible, although less plausible, arrangements of two composition vectors on the two trees, and if other tree topologies are considered then the number of possibilities is compounded. A complete examination of all the possibilities would be too computationally expensive for all but the smallest data sets.

A solution to this problem is to analyze the phylogeny in a Bayesian framework using an MCMC to approximate the posterior probability distribution, and to let the MCMC place the composition vectors on the trees. The proposals in the MCMC include changes to the composition, rate matrices, among-site rate parameters, topology, and branch lengths, as would be the case for a homogeneous model. The additional proposal of allowing each branch to choose among available composition vectors solves the daunting problem of how best to arrange them on the tree. Composition parameters are free even when there is more than one composition vector. Such an analysis simultaneously explores the posterior probability densities of tree topologies, model parameters, and placement of composition vectors on the trees.

An additional complication is that if the model parameters differ over internal branches then the likelihood depends on the root, and for that reason I used the additional MCMC proposal of a change in the trifurcating internal node root position.

A Bayesian analysis of the bacterial 16S data using a homogeneous model results in the attract tree (Fig. 2b) being most probable, with the split to the two mesophiles having a posterior probability of 0.86, and the other split, to *Thermotoga* and *Aquifex*, having a posterior probability of 1.0. An analysis with a two-part composition heterogeneous model results in the true tree (Fig. 2a) being most probable, with both splits having a posterior probability of 1.0. Repeated runs starting with different random tree topologies and initial placements of the two composition vectors converge to the tree topology shown in Figure 2a, with the two composition vectors arranged as shown by the dashed-solid line pattern in that figure. After burn-in, the trees maintained this composition arrangement essentially all of the time. This suggests that the lineage to *Deinococcus* and *Thermus* was more thermophile-like than mesophile-like (Fig. 4a *cf* Fig. 4b).

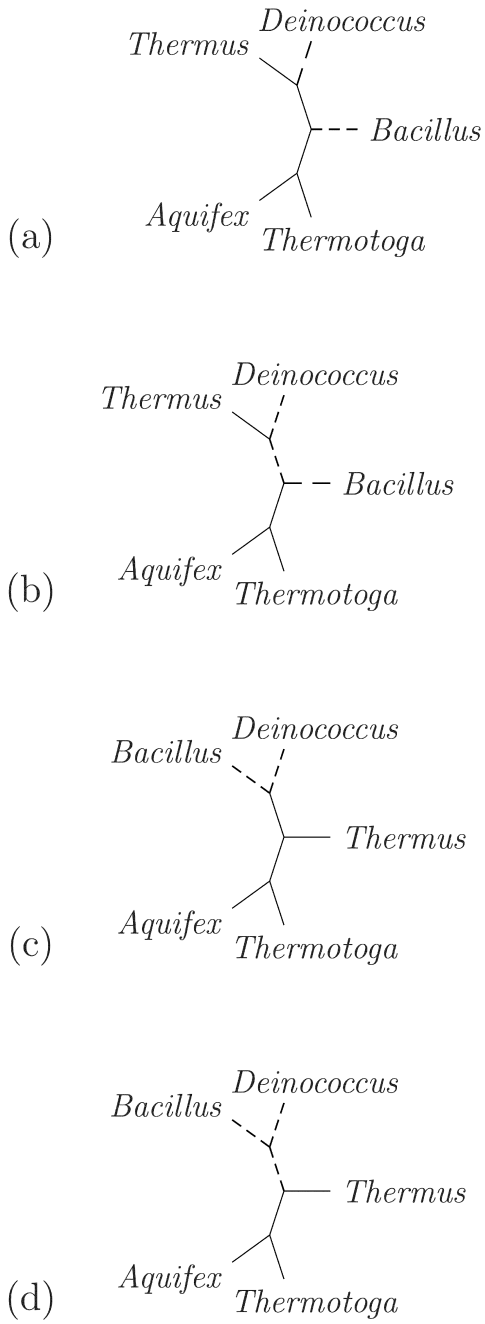

FIGURE 4. Arrangement of composition parameters on the trees. Dashed lines show the arrangement of the composition for the mesophiles, solid lines show the arrangement of the composition for the thermophiles. (a and b) This is the true topology, with two plausible arrangements of the two composition vectors. (a) The arrangement shown in Figure 2, and of the arrangements shown here is the ML arrangement. (c and d) Same as (a) and (b), for the attract tree.

Bayesian Tests for Fit of the Model to the Data

The fit of the model can be assessed in a Bayesian framework by using a posterior predictive distribution of some test quantity from data simulated using the model and comparing that to the test quantity from the original data (Gelman et al., 1995; Huelsenbeck et al., 2001; Bollback, 2002). This approach is based on the idea that data simulated under a model that fits should be similar to the original data. Two test statistics have been suggested to measure this similarity—the unconstrained or multinomial likelihood of the data, and the composition X^2 statistic (Huelsenbeck et al., 2001; Bollback, 2002).

Posterior predictive simulations for the bacterial 16S analysis are shown in Figure 5. The multinomial likelihood (Fig. 5a and b) does not distinguish between the homogeneous and the two-part models. For the homogeneous model the tail area probability is 0.26, and for the two-part model the tail area probability is 0.37. This test statistic appears to indicate that both models fit the data; however, we know from the Goldman-Cox test above that the homogeneous model does *not* fit. It appears that the multinomial likelihood is of limited utility as a test statistic in this context.

The fit of the composition is shown in Figure 5c and d, using the test statistic X^2, the same as is used in the χ^2 test for compositional homogeneity. Here X^2 for the original data is 47.8. Data simulated under a homogeneous model have small X^2 values, and these clearly differ from the original data (Fig. 5c). However, using the two-part model the X^2 distribution from simulated data is much greater, and the X^2 from the original data is within the posterior predictive distribution (Fig. 5d; tail area probability = 0.35), showing by this test that the composition of the model fits the composition of the data. In contrast to the multinomial likelihood, X^2 appears to be a useful test quantity to determine model composition fit.

Xanthine Dehydrogenase from Drosophila

For a second example I turn to the analysis by Tarrío et al. (2000), who showed that outgroup rooting with the *Xdh* gene failed to find the preferred root of the *Drosophila saltans* and *Drosophila willistoni* groups. Root position 1 in Figure 6 is the preferred root based on morphology and on deletion of an intron in the *Adh* gene that is specific to the *willistoni* group. Using the ingroup only, a satisfactory phylogeny was obtained, corroborating known relationships derived from morphological characters. However, when outgroup taxa were added to the analysis, the position of the root of the ingroup became unstable, and changed depending on the model or method of analysis (Fig. 6). This instability was blamed on compositional differences among the taxa, especially between the outgroup and the ingroup, and indeed the data fail the χ^2 composition test. A distance-based analysis using LogDet/paralinear distances, which can overcome compositional heterogeneity, finds root 1.

I begin the reanalysis of these data by choosing a model, arbitrarily using the tree rooted at position 1 in Figure 6, with the expectation that other roots will not affect our choice of model. Confirming Tarrío et al. (2000), of the models available in PAUP* the GTR model was found to be the most suitable overall rate matrix, and the

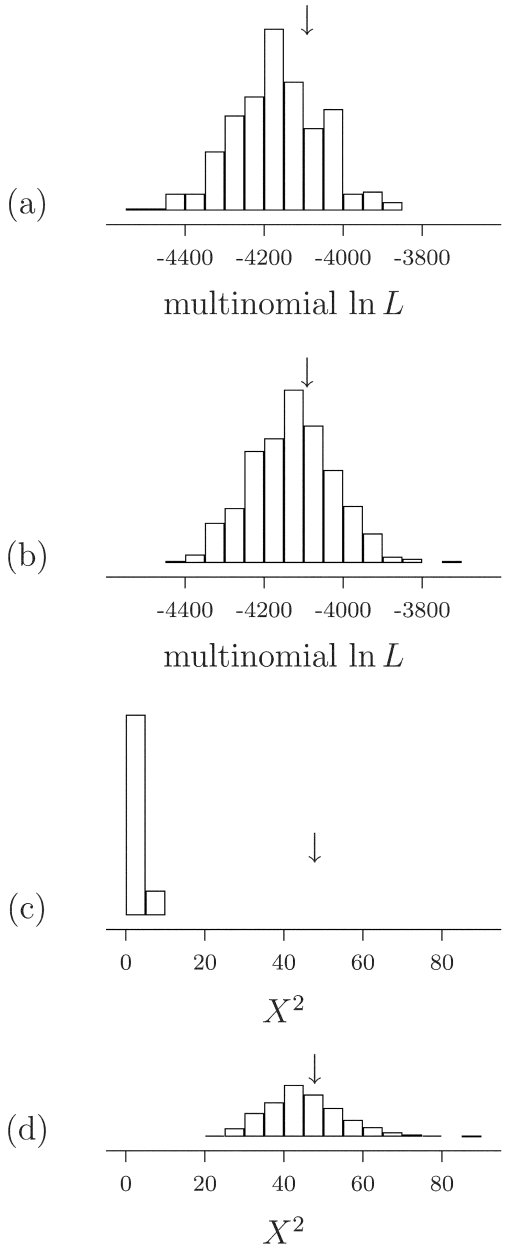

FIGURE 5. Posterior predictive distributions from bacterial 16S. Arrows show the test quantity from the original data. Panels (a) and (b) show a model fit test using the unconstrained or multinomial likelihood. Panel (a) shows the test quantity generated by simulations using the homogeneous model. Panel (b) shows the test quantity for the two-part heterogeneous model. Using this test quantity, because the test quantity from the original data falls within both of the distributions, it appears that both models fit the data. Panels (c) and (d) show a model fit test using X^2 as the test quantity. Panel (c) shows the test quantity for simulations using the homogeneous model and shows marked lack of model fit. Panel (d) shows the test quantity for simulations under the two-part heterogeneous model, which fits the data.

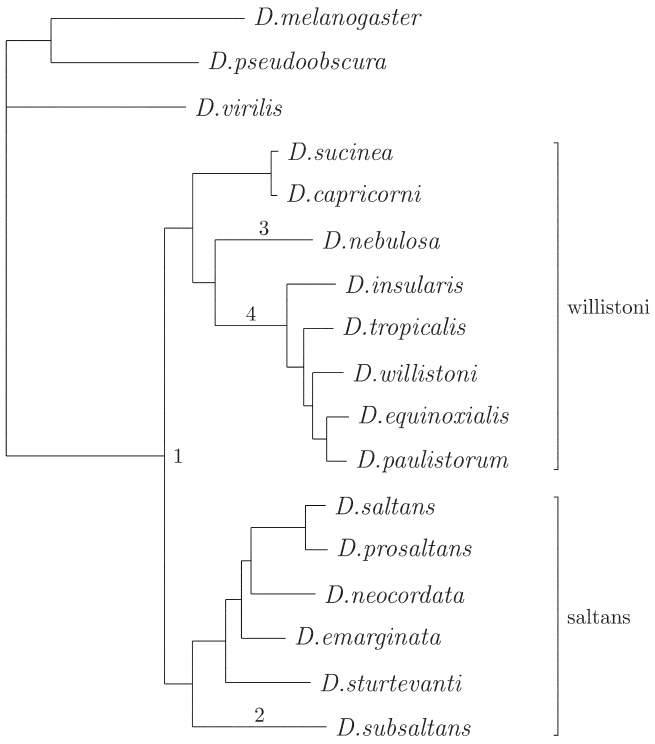

FIGURE 6. Rooting the *Drosophila saltans* and *willistoni* groups. Modified from Tarrío et al. (2000), Figure 1. The outgroup is *D. melanogaster*, *D. pseudoobscura*, and *D. virilis*. Four different roots, attachment points of the outgroup to the ingroup, indicated by 1 to 4, were found by various methods. Roots 1 to 3 were noted in Tarrío et al. (2000), and root 4 is from the present study. The preferred root is as shown, root 1.

codon-based site-specific model (SS) accounted for the among-site rate heterogeneity best (Table 2). This model does not pass the Goldman-Cox test, and none of the three codon positions pass the tree- and model-based composition-fit test. Using PAUP*, a search for the ML tree using the GTR + SS model finds a tree with root 2, supporting Tarrío et al. (2000). A Bayesian analysis using MrBayes also finds root 2, either with the GTR + SS or with the GTR + SSΓ model; in the latter each codon partition is given its own gamma-distributed among-site rate variation.

TABLE 2. *Xdh* model choice. The tree with the outgroup rooted at position 1 in Figure 6 was evaluated with various among-site rate accommodation, all using the GTR rate matrix. Δ is the difference in ln *L* between the indicated model and the model in which it nests as a special case. Among-site rate parameters are I, a proportion of invariant sites, Γ, discrete gamma-distributed rates of variable sites, and SS, a site-specific model where the three codon position classes are allowed their own rates. The overall composition was estimated by maximum likelihood in all cases. The SS model is by far the best of the among-site rate accommodations shown.

Among-site rate variation	ln *L*	Δ	Parameters
—	−15606.8		8
I	−14743.5	863.2	9
Γ	−14696.2	910.6	9
I + Γ	−14673.4	22.8	10
SS	−14264.3	1342.5	10

From Table 2 it is evident that allowing the three codon positions to have their own site-specific rates is important. Tarrío et al. (2000) pointed out that the three codon positions also differ markedly in their base composition. Therefore it is reasonable to ask whether allowing different compositions or even rate matrices in each codon position is worth the cost of the extra parameters. If each codon position is allowed its own composition, at a cost of six additional parameters, the ln *L* increases 49.7, which is highly significant ($P \approx 0$). If in addition we allow each codon position its own GTR rate matrix, at a cost of an additional 10 parameters, the ln *L* increases 70.7, which is again highly significant ($P \approx 0$). Therefore each codon position was given independent GTR rate matrices and composition, as well as partition rates.

Among-site rate variation was examined in the three codon positions. I tested all combinations of using a proportion of invariant sites (I), gamma-distributed variable sites (Γ), both I and Γ, and no among-site rate variation. The combination Γ + ΓI + Γ was chosen because it had the lowest AIC score (Akaike, 1974).

The best-fitting model examined so far is one where the rate matrix, the composition, and among-site rate heterogeneity parameters are free in each of the three codon positions. This model now passes the Goldman-Cox test. The tree- and model-based composition-fit test borderline fails for the first codon position ($P = 0.03$), passes for the second codon position ($P = 0.91$), and fails markedly for the third position ($P < 0.01$).

Bayesian analysis with this model shows that the root 4 tree has the highest posterior probability. Posterior predictive simulation can be used to look at the fit of this model (Fig. 7). The X^2 statistic, used to show fit of the composition of the model, shows borderline lack of fit in the first position, good fit in the second, and markedly poor fit in the third, in agreement with the tree- and model-based composition fit test above.

Using the multinomial log likelihood in posterior predictive simulation (not shown) gives a tail area probability of $p_t = 0.311$ for the first codon position, 0.18 for the second position, and 0.052 for the third position, and so does not indicate lack of fit of this model, although the p_t value is borderline for the third position. As in the bacterial 16S analysis above, we can, however, be skeptical of this result, as there is evidence above that the model composition does not fit.

Clearly compositional heterogeneity in the third position needs to be accommodated in the model. This is done, as in the bacterial 16S analysis above, by adding additional vectors of composition parameters, which are allowed to move among the branches of the tree during the MCMC. Because the position of the overall root can affect the likelihood if there is model heterogeneity in the internal branches, the position of the overall root was

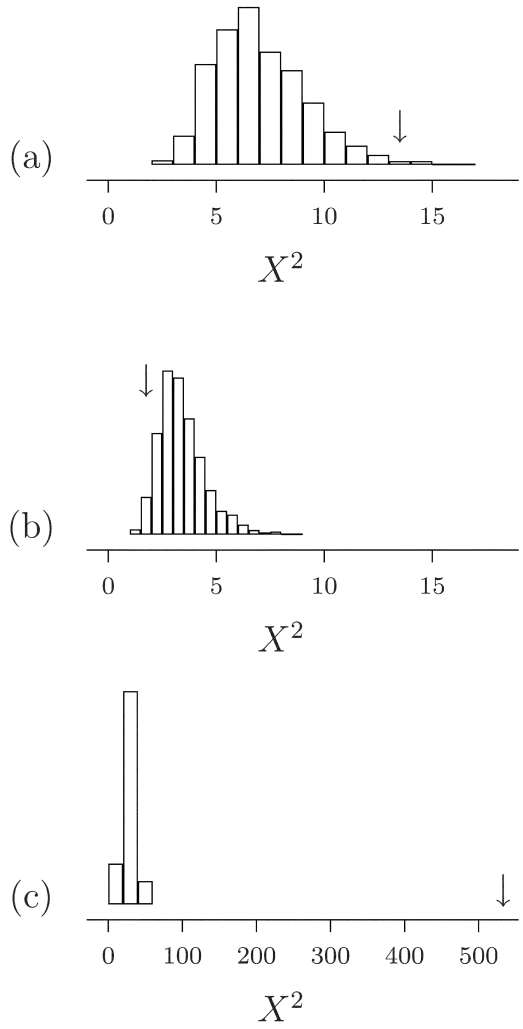

FIGURE 7. Bayesian model composition fit assessed using posterior predictive simulation. The model used separate GTR rate matrix, composition, and among-site rate heterogeneity in each of the 3 codon positions, homogeneous over the tree. Bars show the distribution of X^2 in the posterior predictive simulation (a) for the first codon position, (b) the second codon position, and (c) for the third codon position. Arrows show the realized X^2 for the original data for each codon position.

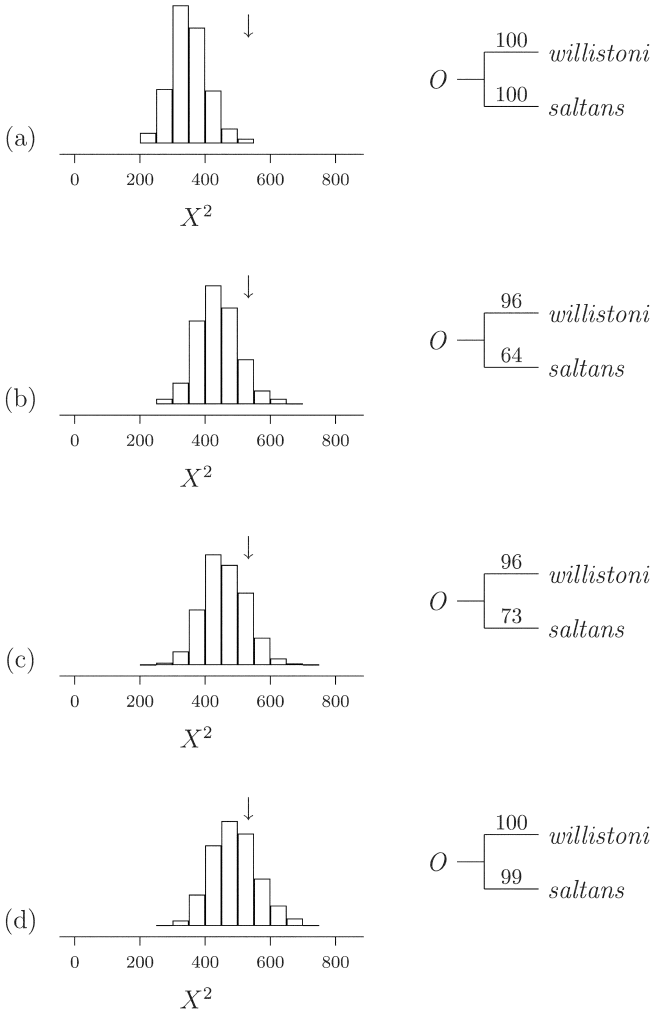

FIGURE 8. Bayesian composition fit tests where the composition differs over the tree. The model used separate GTR rate matrix, composition, and among-site rate heterogeneity in each of the three codon positions. For the third codon position only, composition was accommodated with (a) two, (b) three, (c) four, and (d) five vectors of composition parameters. Left panels show posterior predictive simulation results using the composition X^2 statistic, where the statistic for the original data is shown with an arrow. Tail area probabilities p_t for these analyses are 0.002, 0.077, 0.137, and 0.256, for panels (a), (b), (c), and (d), respectively. Panels on the right show percent posterior probability support for outgroup rooting position 1. The posterior probability of the split to the outgroup was 100% in all cases.

allowed to move during the MCMC as well. (Here the overall root of the entire analysis is distinguished from the outgroup root position, numbered as in Fig. 6.) When heterogeneous composition was accommodated in the model, outgroup root position 1, the preferred root, was recovered (Fig. 8), together with a greatly improved fit of the model to the composition in the third position (Fig. 8 *cf* Fig. 7).

When two composition vectors are used, outgroup root 1 is recovered with high confidence even though the model shows lack of fit (Fig. 8a). When three composition vectors are used, it appears that the model composition now fits the data; however, there is an unexpected decrease in support for root 1 (Fig. 8b). When additional composition vectors are added, the composition fit increases, and support for root 1 increases again (Fig. 8c and d). A steady increase in support for the correct root would have been expected as the model fit improves, and so this unusual trend deserves comment. As mentioned above, the root of the entire analysis was allowed to change during the MCMC. When two composition vectors were used, the overall root was usually located within the *saltans* group, but when three or more composition vectors were used the overall root was usually located within the outgroup. Therefore it appears that the unexpected high support for outgroup root 1 when two composition vectors are used is influenced by the overall root.

DISCUSSION

Previous models that allowed the composition to differ over the tree had free composition parameters on each branch, or on every terminal branch, in the tree (Yang and Roberts, 1995; Galtier and Gouy, 1998). This does not scale well, because large trees will then have many parameters. Based on the notion that if compositional heterogeneity exists it may be localized in part of the tree, I have developed models that do not require that each branch of the tree get a different composition. A simple model may be made with this approach that has only two vectors of composition parameters on the tree, as was done with the bacterial 16S data. If two composition vectors are not enough to adequately model the data, as was the case with the *Xdh* example, then more composition vectors can be added until the model fits. The process of adding or removing parameters can itself be made part of the MCMC, although this was not done in this study (Huelsenbeck et al., 2000; Suchard et al., 2001; Green, 1995).

This study also focused on assessing the overall fit of the model and the fit of the composition of the model to the data, using both ML and Bayesian methods. In an ML context, the Goldman-Cox test was used to assess overall fit. This test was sensitive enough to show lack of fit of the homogeneous model with bacterial 16S. However, it was not sensitive enough to show lack of fit of a tree-homogeneous model using *Xdh*, even though these data failed the composition χ^2 test. In a Bayesian context posterior predictive simulation was used to assess the fit of the model. The multinomial likelihood and the composition X^2 statistic were suggested as test quantities for this purpose (Huelsenbeck et al., 2001; Bollback, 2002). The composition X^2 statistic was useful in the present study. However, the multinomial likelihood was not found to be useful in this study, and was not sensitive to lack of model fit in either of the examples used. The multinomial likelihood may show a badly fitting model in more extreme cases, and this was likely the case in the examples cited (e.g., Fig. 3B in Huelsenbeck et al., 2001).

In many cases where there is compositional heterogeneity among lineages the phylogenetic signal is strong and is not overwhelmed by problems with composition (Conant and Lewis, 2001; Rosenberg and Kumar, 2003). I have used two examples here where compositional problems do indeed affect the analysis, do indeed overwhelm the phylogenetic signal, and required accommodation of the compositional heterogeneity in the model to allow the underlying phylogenetic signal to be seen.

ACKNOWLEDGMENTS

Thanks to Martin Embley for the bacterial 16S alignment, and to Francisco Rodriguez-Trelles for the Drosophila *Xdh* alignment. Thanks go to Chris Simon, Ted Schultz, Mike Steel, and Bret Larget for many helpful suggestions. Thanks especially to Bret Larget for clarifying the notation of the model rate matrices.

REFERENCES

Akaike, H. 1974. A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19:716–723.

Bollback, J. P. 2002. Bayesian model adequacy and choice in phylogenetics. Mol. Biol. Evol. 19:1171–1180.

Conant, G. C., and P. O. Lewis. 2001. Effects of nucleotide composition bias on the success of the parsimony criterion in phylogenetic inference. Mol. Biol. Evol. 18:1024–1033.

Eisen, J. A. 1995. The RecA protein as a model molecule for molecular systematic studies of bacteria: Comparison of trees of RecAs and 16S rRNAs from the same species. J. Mol. Evol. 41:1105–1123.

Embley, T. M., R. H. Thomas, and R. A. D. Williams. 1993. Reduced thermophilic bias in the 16S rDNA sequence from *Thermus ruber* provides further support for a relationship between *Thermus* and *Deinococcus*. Syst. Appl. Microbiol. 16:25–29.

Felsenstein, J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. J. Mol. Evol. 17:368–376.

Foster, P. G., and D. A. Hickey. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J. Mol. Evol. 48:284–290.

Galtier, N., and M. Gouy. 1998. Inferring pattern and process: Maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. 15:871–879.

Galtier, N., N. Tourasse, and M. Gouy. 1999. A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221.

Gelman, A. B., J. S. Carlin, H. S. Stern, and D. B. Rubin. 1995. Bayesian data analysis. Chapman & Hall/CRC, London.

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

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. 49:652–670.

Green, P. J. 1995. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika 82:711–732.

Gupta, R. S. 1998. Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. Microbiol. Mol. Biol. Rev. 62:1435–1491.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22:160–174.

Huelsenbeck, J. P., B. Larget, and D. Swofford. 2000. A compound poisson process for relaxing the molecular clock. Genetics 154:1879–1892.

Huelsenbeck, J. P., and B. Rannala. 1997. Phylogenetic methods come of age: Testing hypotheses in an evolutionary context. Science 276:227–232.

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

Huelsenbeck, J. P., F. Ronquist, R. Nielsen, and J. P. Bollback. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294:2310–2314.

Jukes, T., and C. Cantor. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. Munro, ed.). Academic Press, New York.

Lake, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. U.S.A. 91:1455–1459.

Larget, B., and D. L. Simon. 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. Mol. Biol. Evol. 16:750–759.

Lockhart, P. J., C. J. Howe, D. A. Bryant, T. J. Beanland, and A. W. Larkum. 1992. Substitutional bias confounds inference of cyanelle origins from sequence data. J. Mol. Evol. 34:153–162.

Lockhart, P. J., M. A. Steel, M. D. Hendy, and D. Penny. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

Lopez, P., D. Casane, and H. Philippe. 2002. Heterotachy, an important process of protein evolution. Mol. Biol. Evol. 19:1–7.

Mooers, A. Ø., and E. C. Holmes. 2000. The evolution of base composition and phylogenetic inference. Trends Ecol. Evol. 15:365–369.

Murray, R. G. E. 1991. The family Deinococcaceae. Pages 3733–3744 *in* The prokaryotes, vol. 4 (A. Balows, H. G. Trüper, M. Dworkin, W. Harder, and K.-H. Schleifer, eds.). Springer, London.

Penny, D., B. J. McComish, M. A. Charleston, and M. D. Hendy. 2001. Mathematical elegance with biochemical realism: The covarion model of molecular evolution. J. Mol. Evol. 53:711–723.
Posada, D., and K. A. Crandall. 1998. ModelTest: Testing the model of DNA substitution. Bioinformatics 14:817–818.
Rosenberg, M. S., and S. Kumar. 2003. Heterogeneity of nucleotide frequencies among evolutionary lineages and phylogenetic inference. Mol. Biol. Evol. 20:610–621.
Rzhetsky, A., and M. Nei. 1995. Tests of applicability of several substitution models for DNA sequence data. Mol. Biol. Evol. 12:131–151.
Sokal, R. R., and F. J. Rohlf. 1981. Biometry. W. H. Freeman, 2nd ed.
Steel, M. A. 1994. Recovering a tree from the leaf colorations it generates under a Markov model. Applied Mathematics Letters 7:19–24.
Suchard, M. A., R. E. Weiss, and J. S. Sinsheimer. 2001. Bayesian selection of continuous-time Markov chain evolutionary models. Mol. Biol. Evol. 18:1001–1013.
Sullivan, J., and D. L. Swofford. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. J. Mammal. Evol. 4:77–86.
Swofford, D. L. 2002. PAUP*. Phylogenetic analysis using parsimony (*and other methods), Version 4. Sinauer Associates, Sunderland, MA.
Swofford, D. L., G. J. Olson, P. J. Waddell, and D. M. Hillis. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd ed. (D. M. Hillis, G. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, MA.
Tarrío, R., F. Rodriguez-Trelles, and F. J. Ayala. 2000. Tree rooting with outgroups when they differ in their nucleotide composition from the ingroup: The *Drosophila saltans* and *willistoni* groups, a case study. Mol. Phylogenet. Evol. 16:344–349.
Tarrío, R., F. Rodriguez-Trelles, and F. J. Ayala. 2001. Shared nucleotide composition biases among species and their impact on phylogenetic reconstructions of the Drosophilidae. Mol. Biol. Evol. 18:1464–1473.
Whelan, S., P. Liò, and N. Goldman. 2001. Molecular phylogenetics: State-of-the-art methods for looking into the past. Trends Genet. 17:262–272.
Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.
Yang, Z., N. Goldman, and A. Friday. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. 11:316–324.
Yang, Z., and D. Roberts. 1995. On the use of nucleic acid sequences to infer early branchings in the tree of life. Mol. Biol. Evol. 12:451–458.

First submitted 21 July 2003; reviews returned 31 October 2003;
final acceptance 14 December 2003
Associate Editor: Ted Schultz