Syst. Biol. 56(2):194–205, 2007

ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150701291804

Accounting for Gene Rate Heterogeneity in Phylogenetic Inference

RACHEL B. BEVAN,[1] DAVID BRYANT,[3] AND B. FRANZ LANG[2]

[1]*McGill Centre for Bioinformatics, McGill University, Duff Medical Building, 3775 University Street, Montréal, Quebec, H3A 2B4, Canada; E-mail: rbbevan@gmail.com*
[2]*Program in Evolutionary Biology, Canadian Institute for Advanced Research Centre Robert Cedergren, Départrement de Biochimie, Université de Montréal, 2900 Boulevard Edouard–Montpetit, Montréal, Québec, H3T 1J4, Canada*
[3]*Department of Mathematics, The University of Auckland, Private Bag 92019, Auckland New Zealand*

Abstract.— Traditionally, phylogenetic analyses over many genes combine data into a contiguous block. Under this concatenated model, all genes are assumed to evolve at the same rate. However, it is clear that genes evolve at very different rates and that accounting for this rate heterogeneity is important if we are to accurately infer phylogenies from heterogeneous multigene data sets. There remain open questions regarding how best to incorporate gene rate parameters into phylogenetic models and which properties of real data correlate with improved fit over the concatenated model. In this study, two methods of accounting for gene rate heterogeneity are compared: the *n*-parameter method, which allows for each of the *n* gene partitions to have a gene rate parameter, and the α-parameter method, which fits a distribution to the gene rates. Results demonstrate that the *n*-parameter method is both computationally faster and in general provides a better fit over the concatenated model than the α-parameter method. Furthermore, improved model fit over the concatenated model is highly correlated with the presence of a gene with a slow relative rate of evolution. [AIC; gene rates; phylogenetic integration; phylogenomics; rate heterogeneity.]

The use of multigene data sets in phylogenetic analysis is imperative in order to resolve evolutionary relationships over large taxon sets and deep phylogenetic divergences. Multigene data sets have the advantage of providing greater resolution—with more information it is possible to find trees that more accurately reflect evolutionary history (see, for instance, Gontcharov et al., 2004). However, the heterogeneous nature of the data does present problems. When there are many genes present in the analysis, it is necessary to account for the fact that different genes undergo different selective pressures and that the degree of site rate heterogeneity within a gene may vary from gene to gene. The incorporation of data under different evolutionary pressures (as found in different codon positions, or different genes) should be taken into account when calculating likelihoods (Bull et al., 1993; Yang, 1996; Huelsenbeck et al., 1996; Pupko et al., 2002; Bapteste et al., 2002; Nylander et al., 2004; Cranston and Rannala, 2005).

Determining how best to incorporate gene rates in a maximum likelihood (ML) context is a relatively unexplored area of phylogenetics research. When incorporating gene rates into maximum likelihood phylogeny estimation there are two approaches that can be taken. The first approach involves using a single rate for each gene (hereafter referred to as the *n*-parameter method) as initially proposed for DNA sequences by Yang (1996) and extended to protein data by Pupko et al. (2002). This approach to accounting for gene rate heterogeneity has been shown to lead to both an improved model fit according to the AIC (Yang, 1996; Pupko et al., 2002; Bevan et al., 2005) and better topologies (Bevan et al., 2005). The second approach involves integrating out over all possible rates for a given gene using a discrete approximation to a continuous distribution (hereafter referred to as the α-parameter method; Felsenstein, 2001, 2004).

Both approaches to accounting for gene rate heterogeneity assume that a gene evolves at a particular rate of evolution. However, the *n*-parameter method does not allow for any uncertainty in the rate for a particular gene but assumes that it is valid to use ML estimates of the rate to account for gene rate heterogeneity. Conversely, the α-parameter method does account for uncertainty in the rate estimate for a gene through integration over all possible values that a rate might take. Although assuming a single rate for each gene is computationally faster than integrating, it could potentially suffer from the difficulty of infinite parameterization when many genes are used in the analysis (thus overfitting the data). Although the *n*-parameter method has been tested and found to lead to significant improvement in maximum likelihood phylogeny estimation (Yang, 1996; Pupko et al., 2002; Bevan et al., 2005), the α-parameter method has yet to be investigated.

In addition to determining how best to incorporate gene rate parameters, the question remains of which features of the data correlate with improved fit over the concatenated model. This paper has two goals: (i) to determine whether the computational effort required by the α-parameter method is justified according to the Akaike information criterion (AIC) and cross-validation information criterion (CVIC); and (ii) to determine the properties of the data that lead to an improved fit when accounting for gene rate heterogeneity in phylogenetic models.

MATERIALS AND METHODS

The n-Parameter Method

The *n*-parameter method is well studied and has been shown to lead to higher likelihood for a particular topology than the concatenated model (Yang, 1996; Pupko et al., 2002; Bevan et al., 2005). Consider *n* genes $G_1, \ldots, G_n$. Let $R_1, \ldots, R_n$ denote the relative rates of evolution of $G_1, \ldots, G_n$, let θ denote the pair $\{T, \lambda\}$ where T is a tree topology and λ a set of branch lengths. Also let α_s be the parameter for the distribution

accounting for rates across sites. Here α_s is used instead of α in order to differentiate between rates across sites and rates across genes. The likelihood is computed as:

$$\begin{aligned} &L_n(\theta, \alpha_s, R_1, \ldots, R_n \mid G_1, \ldots, G_n) \\ &= P(G_1, G_2, .., G_n \mid \theta, \alpha_s, R_1, \ldots, R_n) \\ &= P(G_1 \mid \theta, \alpha_s, R_1)P(G_2 \mid \theta, \alpha_s, R_2) \cdots P(G_n \mid \theta, \alpha_s, R_n) \end{aligned} \quad (1)$$

The rates $R_1, \ldots, R_n$ have mean 1.0. The parameter θ may also include other parameters (such as the proportion of invariant sites). The formula in Equation 1 makes the assumption that the rates of evolution of distinct genes are independent.

Let $G_{g,i}$ denote site i in gene g. For this site the likelihood is

$$\begin{aligned} L_{n_{g,i}}(\theta, \alpha_s, R_g \mid G_{g,i}) &= P(G_{g,i} \mid \theta, \alpha_s, R_g) \\ &= \int_0^\infty f(r \mid \alpha_s)P(G_{g,i} \mid \theta, r, R_g)dr \\ &\approx \sum_{j=1}^{S} p(\hat{r}_j \mid \alpha_s)P(G_{g,i} \mid \theta, \hat{r}_j, R_g) \\ &= \sum_{j=1}^{S} p(\hat{r}_j \mid \alpha_s)P(G_{g,i} \mid \theta, \hat{r}_j \times R_g) \end{aligned} \quad (2)$$

where S is the number of categories used to approximate the probability density function of the Γ distribution $[f = \Gamma(\alpha_s, \frac{1}{\alpha_s})]$ for site rate heterogeneity, $\hat{r}_j$ is the site rate for category j, and $p(\hat{r}_j)$ is the probability of this site rate category. It is possible to have one α_s over all sites in all genes. It is also possible to have one α_s for each gene (or $\alpha_{s_1}, \ldots, \alpha_{s_n}$). In both cases the likelihood of a site is calculated in the same way.

This model assumes that branch lengths for different genes vary only by a constant scale factor. In effect, the branch lengths are multiplied by a value proportional to the evolutionary rate of the gene and the evolutionary rate of a site.

Because the sites are assumed to be independent, the likelihood for an entire gene $L_{n,g}$ is calculated from the product of the site likelihoods (2) for all sites i in gene g (i.e., sites $i \in g$) as:

$$\begin{aligned} &L_{n_g}(\theta, \alpha_s, R_g \mid G_g) \\ &= \prod_{\text{sites } i \in g} L_{n_{g,i}}(\theta, \alpha_s, R_g \mid G_{g,i}) \\ &= \prod_{\text{sites } i \in g} \sum_{j=1}^{S} p(\hat{r}_j \mid \alpha_s)P(G_{g,i} \mid \theta, \hat{r}_j \times R_g) \end{aligned} \quad (3)$$

Combining (1) and (3) with the assumption of independence between genes, the log–likelihood over all sites is calculated as:

$$\begin{aligned} &\log[L_n(\theta, \alpha_s, R_1, \ldots, R_n \mid G_1, \ldots, G_n)] \\ &= \log\left[\prod_{\text{genes } g} L_{n_g}(\theta, \alpha_s, R_g \mid G_g)\right] \\ &= \log\left[\prod_{\text{genes } g} \prod_{\text{sites } i \in g} L_{n_{g,i}}(\theta, \alpha_s, R_g \mid G_{g,i})\right] \\ &= \sum_{\text{genes } g} \sum_{\text{sites } i \in g} \log\left[L_{n_{g,i}}(\theta, \alpha_s, R_g \mid G_{g,i})\right] \\ &= \sum_{\text{genes } g} \sum_{\text{sites } i \in g} \log\left[\sum_{j=1}^{S} p(\hat{r}_j \mid \alpha_s)P(G_{g,i} \mid \theta, \hat{r}_j \times R_g)\right] \end{aligned}$$

Thus, no time or computational complexity is added when calculating the likelihood versus computing the likelihood of the concatenated data set with no gene rates (or equivalently computing the likelihood with $R_g = 1.0$ for all genes g). The only additional computational time required is optimizing over the gene rates $R_1, \ldots, R_n$. However, with good starting estimates, such as those found with the DistR method (Bevan et al., 2005), this time is not too significant.

The α-Parameter Method

Define $\theta = \{\lambda, T\}$ for branch lengths λ and a topology T and α_s as the parameter for the rates across sites distribution. Also let ω be the parameter for the rates across genes distribution h. Then the likelihood of gene g under the α-parameter method is calculated as

$$\begin{aligned} &L_{\alpha_g}(\theta, \alpha_s, \omega \mid G_g) \\ &= P(G_g \mid \theta, \alpha_s, \omega) \\ &= \int_0^\infty h(R \mid \omega)P(G_g \mid \theta, \alpha_s, R)dR \quad (4) \\ &\approx \sum_{k=1}^{C} p(\hat{R}_k \mid \omega)P(G_g \mid \theta, \alpha_s, \hat{R}_k) \quad (5) \\ &= \sum_{k=1}^{C} p(\hat{R}_k \mid \omega) \prod_{\text{sites } i \in g} P(G_{g,i} \mid \theta, \alpha_s, \hat{R}_k) \quad (6) \\ &\approx \sum_{k=1}^{C} p(\hat{R}_k \mid \omega) \prod_{\text{sites } i \in g} \sum_{j=1}^{S} p(\hat{r}_j \mid \alpha_s)P(G_{g,i} \mid \theta, \hat{r}_j \times \hat{R}_k) \end{aligned} \quad (7)$$

where C is the number of categories used to approximate the probability density function h with parameter ω. Probabilities $p(\hat{R}_k)$ are used to approximate h with

rates $\hat{R}_k$, where h is a density function that describes the distribution of gene rates. The best choice of distribution h will be discussed later; however, the mean of the distribution must be 1.0.

In (4) we integrate over the parameter ω, thus computing the likelihood of the data G_g for infinitely many gene rates, weighted by the probability of the gene rate. The approximation (5) of the integral by a summation is made in order to reduce the number of computations involved in integrating. This involves approximating h with a discrete version of the distribution with C categories and rates $\hat{R}_1, \ldots, \hat{R}_C$. Without such an approximation the integration would not be computationally feasible. The equivalence between (5) and (6) is obtained because all sites i in gene g are assumed to be independently evolving. The equality of (6) and (7) exists because site rate heterogeneity is accounted for as in Equation 2. As with the n-parameter method, it is possible to have one Γ distribution describing site rate heterogeneity, or it is possible to have n Γ distributions, one describing the site rate heterogeneity in each gene.

Because the genes are independent, the overall likelihood is

$$L_\alpha(\theta, \alpha_s, \omega \mid G_1, \ldots, G_n) = L_{\alpha_1}(\theta, \alpha_s, \omega \mid G_1) \cdots L_{\alpha_n}(\theta, \alpha_s, \omega \mid G_n)$$

Computing the log-likelihood $\log[L_\alpha(\theta \mid G_1, G_2, \ldots, G_n)]$ for the α-parameter method is more complex, due to the summation over a product. See Appendix 1 for details. Under the α-parameter method it is possible to approximately compute the probability of gene g evolving at a particular rate $\hat{R}_k$, using the Γ distribution as a prior over the possible rates as $P(\hat{R}_k \mid G_g, \theta, \alpha_s) \approx P(G_g \mid R_k, \theta, \alpha_s)P(R_k)$ (which does not account for the probability of the data $P(G_g)$). This can provide a sense of whether the ML rate estimate is a meaningful parameter to describe the data. Additionally, if the unnormalized probabilities are uniform, accounting for rate heterogeneity for the gene of interest may not provide an improved fit over the concatenated model.

The Γ distribution.—The Γ distribution is used to describe gene rate heterogeneity. Under the α-parameter method, reasonable starting values are chosen based upon the ML fit of the Γ distribution to initial gene rate estimates. In phylogenetic analyses, the expected rate over multiple genes is 1.0. Under the $\Gamma(\alpha, \beta)$ distribution this is accomplished by setting $\beta = \frac{1}{\alpha}$, since the expectation of the distribution is then $\alpha\frac{1}{\alpha}$ or 1. A log-normal distribution could also be used here (Felsenstein, 2001).

The DistR Approach

Under both the n-parameter and α-parameter methods, using good initial estimates for the gene rates will help reduce the computation time to find maximal likelihood estimates of the gene rate parameter(s) in each method. In the case of the n-parameter method, initial estimates of the gene rates can be used directly. In the case of the α-parameter method, initial estimates of the gene rates can be used to find a maximum likelihood estimate of the α parameter of the Γ distribution. These initial parameter estimates (either the gene rates or the initial ML estimate of α) are then further optimized to determine the maximum likelihood values.

Here, initial estimates of the gene rates $R_1, \ldots, R_n$ are computed beforehand using the DistR method (Bevan et al., 2005). Initial pairwise distances for the method were estimated using ML distances from PHYML (Guindon and Gascuel, 2003), with the JTT model of evolution, a proportion of invariant sites, and Γ distribution for site rate heterogeneity with 8 categories.

Improved Fit over the Concatenated Model

To determine the improvement, if any, of the α-parameter and n-parameter methods over the concatenated model, the Akaike information criterion (AIC) (Akaike, 1974) and cross-validation information criterion (CVIC) (Smyth, 2000) were used. The AIC provides a measure of the expected Kullback Leibler distance between the model of interest and the actual true model. The CVIC does not rely upon data independence like the AIC (Smyth, 2000). It applies the cross-validation principle to obtain a penalized likelihood. However, it is much more computationally demanding and thus was only used on two of the smaller data sets to validate the results.

The LRT was not used because the concatenated model is not nested within either gene rate heterogeneity model (when gene rates are accounted for using the n-parameter or α-parameter method) when each gene has a separate Γ distribution for site rate heterogeneity. The concatenated model is nested within the gene rate heterogeneity model (both n-parameter and α-parameter methods) with one-Γ for site rate heterogeneity. However, the LRT does not follow a χ^2 distribution because the alternative and null models are equivalent when some parameters are fixed at the boundary of parameter space (i.e., when the value of α in the α-parameter method tends towards a large value such as 100).

Calculating the AIC and CVIC.—The AIC is calculated based upon correcting the log-likelihood by some function of the number of parameters in the model of interest. Under the n-parameter method, the parameters are the gene rates for each gene, the site rate heterogeneity parameter(s), the tree topology, and the proportion of invariant sites. The α-parameter method has a similar set of parameters. However, rather than one rate parameter for each gene, it has a parameter for the distribution that describes gene rate heterogeneity. The concatenated model has the same set of parameters but no gene rate parameters and it does not allow for each gene to have separate parameters for site rate heterogeneity.

The AIC is the sum of the negative log-likelihood of the model plus the difference in a function of the number of parameters used in each model multiplied by two. Thus, the difference in AIC between the rates

based model and concatenated model is calculated as: $\Delta AIC = 2L_r - 2L_c + 2(\Delta p)$ where L_c and L_r are the log-likelihoods of the concatenated and gene rates heterogeneity models, respectively. Here Δp is the difference in a function of the number of parameters in the concatenated model and the gene rates model (e.g., either the n-parameter method or α-parameter method) and thus will be a negative number. The first-order AIC does not account for sequence length and thus a second-order AIC was used where the number of parameters is defined as: $\frac{Kn}{(n-K-1)}$ (Burnham and Anderson, 2003) where n is the sequence length and K the number of parameters in the model of interest.

The cross-validation information criterion is useful to confirm the results of the AIC since the AIC makes the assumption of data independence. Although the concatenated model conforms to this assumption, the gene rates models do not. Under the gene rates model, each site in a gene is assumed to be under the same rate of evolution, which violates the independence assumptions of the AIC. The CVIC was designed to determine the correct number of clusters to use in a probabilistic clustering framework (i.e., components in finite mixture models) (Smyth, 2000). Thus, the CVIC does not rely upon the assumption of data independence.

The CVIC for a data set is calculated by dividing the data into two subsets. The model of interest (concatenated or gene rates) is evaluated on one subset, obtaining ML estimates of all parameters of interest. These ML estimates are used to evaluate the likelihood of the data on the second subset under the same model. This process is repeated b times (in this case 50), and the CVIC for model m is calculated as $CVIC_m = \frac{1}{50}\sum_{b=1}^{50} L_{2,m}$. Here $L_{2,m}$ is the likelihood of the second subset of data, evaluated under the ML estimates obtained from the first subset of data. Thus the $\Delta CVIC$ is defined as $\Delta CVIC = CVIC_r - CVIC_c$ where r and c denote the gene rates and concatenated models, respectively.

If the model accounting for rate heterogeneity is preferred as a better fit to the data (versus the concatenated model), the change in AIC or CVIC between the two models (or ΔAIC, Δ $CVIC$) will be positive; otherwise, it will be negative.

Data Analyzed: Empirical Investigation of Gene Rates

Under the α-parameter method it is important to choose a distribution that accurately reflects the gene rates found in empirical data. To determine if the Γ distribution accurately reflects empirical rate estimates, gene rates were calculated over a number of data sets using the DistR method (Bevan et al., 2005). The data sets used for analysis consist of: 41 data sets of size 20 to 40 species per gene (Harlow et al., 2004); a multigene data set consisting of 133 genes over 44 species (Brinkmann et al., 2005); another multigene data set over 37 species with 146 genes; and a 14 species data set with 106 genes (Rokas and Carroll, 2005). The first data set was prepared using automatic homology testing over 144 species, which is an extension of the analysis from Harlow et al. (2004). The other data sets were hand curated (i.e., proteins were hand selected for analysis). In both cases initial distance estimates provided to the DistR procedure were estimated using pairwise ML distances, with eight categories for the Γ distribution, a proportion of invariant sites, and the JTT model of evolution.

Data Analyzed with n*-Parameter and* α*-Parameter Methods*

Six protein data sets were used for analysis: a fungal mitochondrial data set with 29 species and 15 genes (Bevan et al., 2005); a eukaryotic data set with 44 species and 133 genes (Brinkmann et al., 2005); the modified Madsen alignment of placental mammals with 4 genes and 28 species (Madsen et al., 2001; Pupko et al., 2002); the modified Murphy alignment of placental mammals with 6 nuclear genes and 46 species (Murphy et al., 2001; Pupko et al., 2002); an animal mitochondrial data set with 12 genes over 56 species (Pupko et al., 2002); and a fungal nuclear data set with 8 species and 106 genes (Rokas et al., 2003). The data sets used here are available form the first author's website (http://www.mcb.mcgill.ca/rachel/RatesIntegrate/) and as supplementary material (http://systematicbiology.org).

For each data set a modified version of PHYML was run with the default BIONJ starting tree. The JTT model of evolution was used with an estimated parameter for the proportion of invariant sites. Site rate heterogeneity was accounted for using either one Γ distribution for all sites (hereafter denoted one-Γ) or a separate Γ distribution to describe site rate heterogeneity for each gene (hereafter denoted gene-Γ). In both cases, four categories were used in the discrete approximation to the distribution. Gene rate heterogeneity was accounted for using either the n-parameter or the α-parameter method as outlined above. Six equiprobable categories were used in the discrete approximation to the gene rates distribution in the α-parameter method. Gene resampling was performed on the data set over 8 fungal species and 106 genes by randomly selecting 50 gene sets of size 3, 50 gene sets of size 5, and 50 gene sets of size 10.

RESULTS AND DISCUSSION

The n*-Parameter Method Versus the* α*-Parameter Method*

Five diverse data sets with differing numbers of genes and species were analyzed to determine which approach to gene rate heterogeneity results in the greatest improvement over the concatenated model based on the ΔAIC. Table 1 and Figure 2 indicate that with more genes under analysis, there is a greater average ΔAIC favoring a model that accounts for rate heterogeneity. However, based upon these data there is no clear correlation between the spread of the data (i.e., the 1st and 3rd quartiles, or α value under the α-parameter model) and improved model fit over the concatenated model. This is likely complicated by the fact that the data sets are of varying size in terms of the number of genes and species under analysis.

It is evident that there is no advantage to using the α-parameter method over the n-parameter method to

TABLE 1. ΔAIC values for five data sets with differing numbers of genes and species. For the Madsen and Murphy data sets, the $\Delta CVIC$ was calculated. It is given on the second line, after the ΔAIC values. One-Γ and gene-Γ refer to the number of gamma distributions used to account for site rate heterogeneity: either one for the entire data set or one for each gene, respectively. Q refers to the first and third quartiles and α to the value of the α parameter for gene rate heterogeneity under the α-parameter method with one-Γ. ΔAIC and $\Delta CVIC$ values are calculated with respect to the concatenated model. Madsen-PT refers to analysis of the Madsen data set on the preferred topology. In this case, all the parameters were optimized over, except for the topology which was held constant. Madsen-nT refers to analysis of the Madsen data set on the best topology found under the n-parameter method with gene-Γ (the best topologies differ when searching tree space when one-Γ versus gene-Γ are used with the n-parameter method). Madsen-αT refers to analysis of the Madsen data set on the best topology found under the α-parameter method (the topology for with one-Γ and gene-Γ is the same under the α-parameter method when searching topology space).

			n-Parameter			α-Parameter		
Data set	No. genes	No. species	Q	One-Γ	Gene-Γ	α	One-Γ	Gene-Γ
Fungal ΔAIC	15	29	0.75–1.07	1027.77	1152.25	6.284	893.06	1010.45
Eukaryotic ΔAIC	133	44	0.83–1.14	1529.21	2474.07	8.707	1298.84	2199.74
Madsen ΔAIC	4	28	0.81–1.16	154.57	427.80	4.408	149.80	423.33
$\Delta CVIC$				49.86	119.83		13.72	80.09
Madsen–PT ΔAIC	4	28	0.82–1.16	163.77	436.82	4.473	152.64	426.62
$\Delta CVIC$				49.79	121.41		50.16	119.03
Madsen–nTΔ AIC	4	28	0.82–1.17	149.32	422.10	4.465	140.73	414.45
Madsen–αT ΔAIC	4	28	0.81–1.17	153.33	426.29	4.457	142.37	415.87
Animal ΔAIC	12	56	0.81–1.21	248.87	378.14	3.587	221.21	321.0
Murphy ΔAIC	6	46	0.39–1.23	188.88	293.71	1.187	186.48	281.90
$\Delta CVIC$				28.58	55.72		21.60	42.83

find a better fit to the data. According to the ΔAIC (Table 1), the n-parameter method has the best fit compared to the concatenated method for all data sets analyzed. This is true for both one-Γ and gene-Γ analyses. Thus there is no reason to prefer the n-parameter model or α-parameter model as a better fit to the data according to the AIC. When the CVIC was calculated on the two smallest data sets (Madsen and Murphy), the results obtained under the AIC were confirmed (Table 1). This provides independent corroboration that the α-parameter method does not provide a better fit to the data when compared to the n-parameter method. Differences in CVIC for the gene rates model versus the concatenated model are not expected to be as large as the ΔAIC because of the way the CVIC is calculated.

This is especially interesting considering the time to find the tree under each method (Table 2). The n-parameter method takes longer than the concatenated model primarily due to optimization of the ML gene rates. Notably, the α-parameter method takes 2 to 3 times longer than the n-parameter method (Table 2).

When the inferred maximum likelihood (ML) topologies of the α-parameter and n-parameter methods (with gene-Γ) were compared, four out of the five data sets had different topologies. The ML trees from the eukaryotic data set did not have different topologies; however, it is known to be a problematic data set in terms of long branch artifacts and heterotachy (Brinkmann et al., 2005). Thus, even when the ΔAIC indicates that there is little difference between the model fit (Table 1), it is possible that the α-parameter and n-parameter methods find different ML topologies (Fig. 1).

Further investigation of the Madsen data set with gene-Γ (Table 1) shows that for both methods much of the topology agrees with the topology of Murphy et al. (2001), a topology currently supported by molecular data (Fig. 1; Springer et al., 2004) However, the grouping within the Laurasiatheria does not correspond to the currently supported molecular hypothesis (Figs. 1a and b; Springer et al., 2004). The α-parameter method gives the topology for the Laurasiatheria that is closest to the Murphy topology (in terms of SPR moves), only grouping Pangolin incorrectly with flying fox/round eared bat rather than cat/dog. The n-parameter method incorrectly groups horse/rhino and dog/cat into a monophyletic group with flying fox/round eared bat an in-group. Pangolin is also grouped incorrectly in this topology (Fig. 1a).

Although the n-parameter method finds a slightly better fit according to the AIC, care must be taken when evaluating which method finds the best tree topology. Neither method finds the preferred Murphy topology, but this is likely because only four genes were under analysis for 28 species. More data are needed to correctly resolve the phylogeny. Furthermore, when the topology found under the α-parameter method is used to evaluate the data under the n-parameter method (and vice versa), according to the AIC the α-parameter method finds a better tree (Fig. 1, Table 1, Madsen–αT and Madsen analyses under n-parameter method; Madsen–nT and Madsen under α-parameter method).

When the Madsen data were analyzed on the Murphy topology, optimizing for all other parameters, the

TABLE 2. Time for analysis using the gene rate heterogeneity and concatenated models, with the one-Γ to account for site rate heterogeneity. For each data set the analysis for the different models was performed on the same desktop machine. However, times across data sets are not comparable because the different data sets were all analyzed on different computers.

Data set	n-Parameter	α-Parameter	Concatenated
Fungal mtDNA	62 min 25 s	220 min 10 s	12 min 4 s
Eukaryotic	29,866 min 57 s	6996 min 18 s	337 min 0 s
Madsen	23 min 14 s	112 min 44 s	7 min 32 s
Animal mtDNA	120 min 59 s	347 min 53 s	50 min 0 s
Murphy	33 min 24 s	88 min 10 s	13 min 41 s

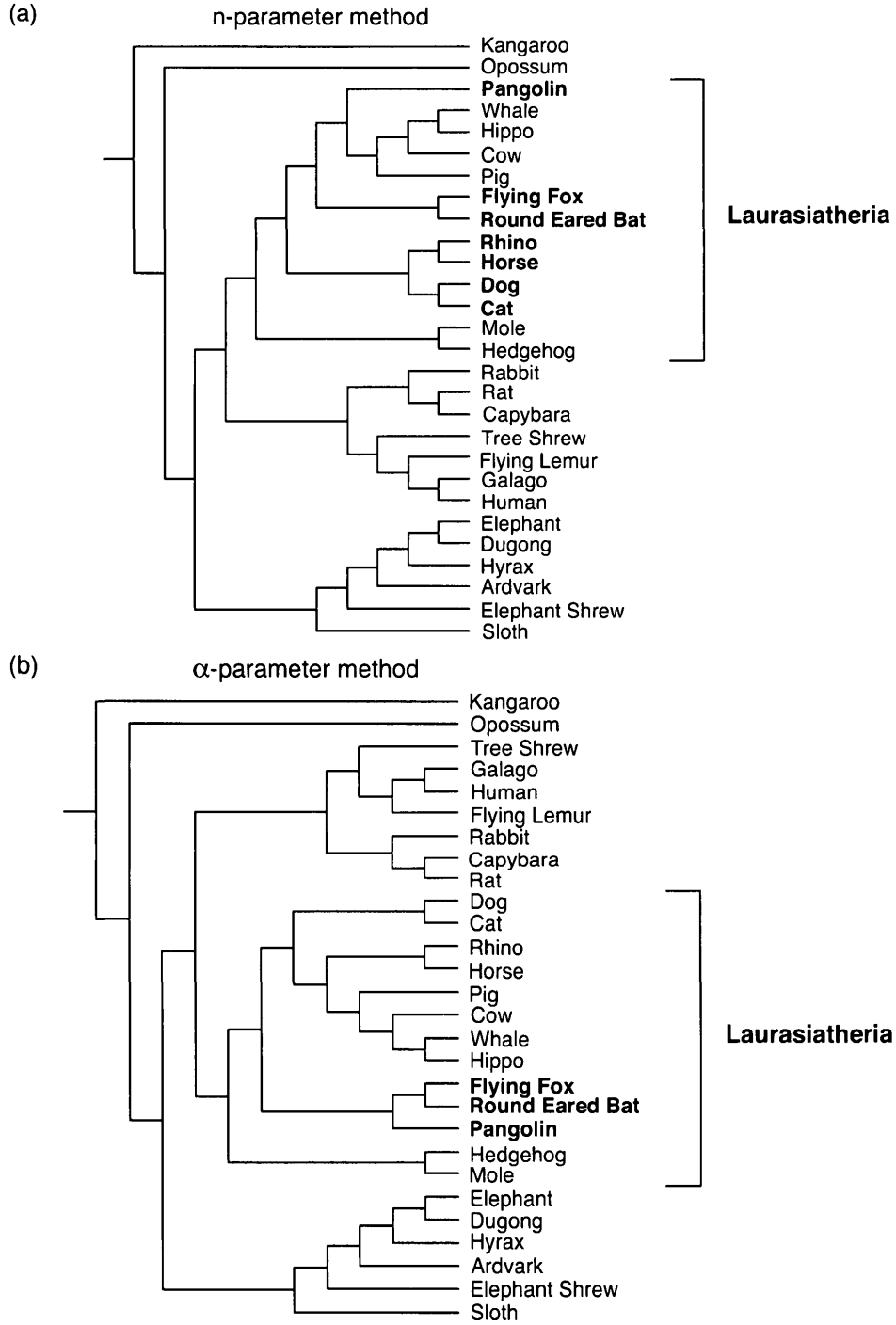

FIGURE 1. Final topologies found for the Madsen data set (Table 1), where branch lengths are not depicted. (a) ML topology found using the n-parameter method with gene-Γ. (b) ML topology found using the α-parameter method. The two methods find different groupings for the Laurasiatheria species.

α-parameter method does not find a better fit to the data than the n-parameter method (Table 1, Madsen-PT) according to both the ΔAIC and $\Delta CVIC$. Thus, although PHYML searches the topology space of trees differently under the α-parameter and n-parameter methods, neither method is preferred as a better fit to the data under the Murphy topology.

Figure 2 gives the ΔAIC values for the resampled genes data sets. These results demonstrate that (i) both methods of accounting for gene rate heterogeneity find approximately equivalent improvement over the concatenated model; (ii) there are some data sets for which accounting for gene rate heterogeneity does not lead to an improved fit (Fig. 2). Figure 2a is particularly important because there is one data set for which accounting for gene rate heterogeneity using the α-parameter method gives a worse fit than the concatenated model ($\Delta AIC = -161.319$, one-Γ for site rate heterogeneity), whereas the

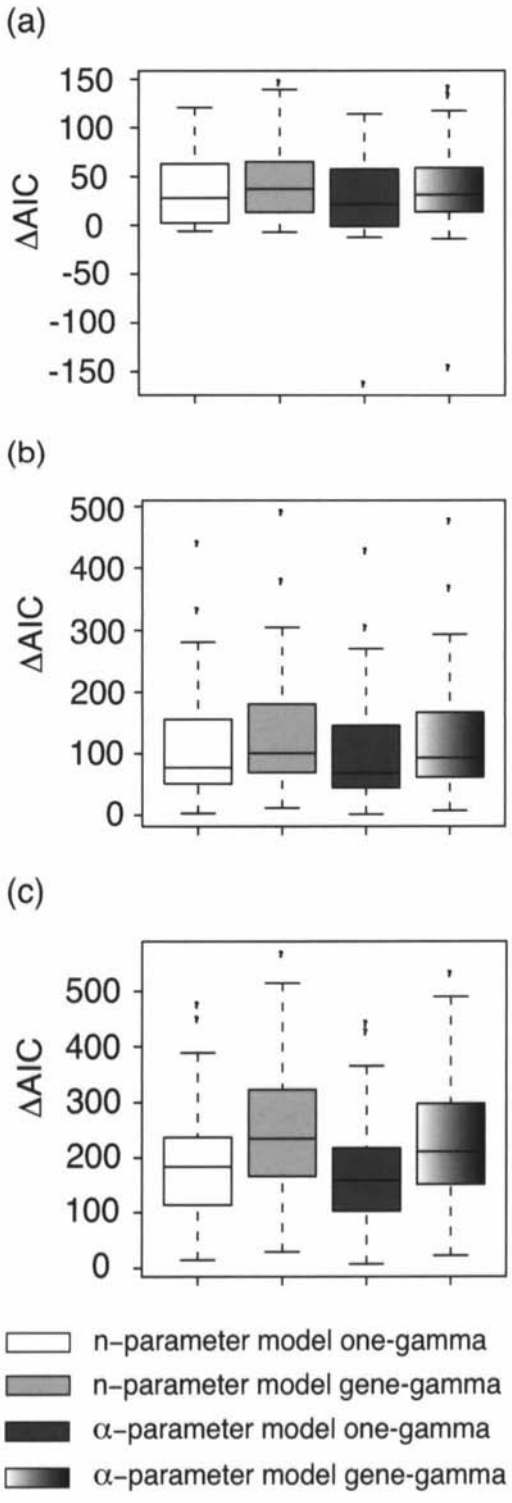

FIGURE 2. Box plots of ΔAIC values comparing the gene rates incorporated model to the concatenated model. ΔAIC values were calculated for (a) 3 genes; (b) 5 genes; (c) 10 genes. In each case the genes were sampled randomly from 106 genes, 50 times.

n-parameter method gives an improved fit (ΔAIC = 63.132, one-Γ for site rate heterogeneity). This is because under the α-parameter method *C. albicans* is incorrectly grouped as a sister taxa to *S. mikatae*, whereas under the n-parameter method the preferred tree topology (see Fig. 5a) is found. This indicates that in some cases the α-parameter method has difficulty converging to the preferred topology in PHYML. Furthermore, both Phillips et al. (2004) and Collins et al. (2005) have shown that deviations from stationary base composition cause trouble for this data set at the nucleotide level, and there is evidence from Foster (2004) that nucleotide base compositional effects translate to the amino acid level. Thus there are other properties of the data that may lead to difficulty in phylogenetic inference in addition to rate effects.

Empirical Rate Distribution—Does the Γ Distribution Describe the Empirical Distribution of Gene Rates?

The α-parameter method does not find a better fit to the data than the n-parameter method. Thus, it is important to determine if it is valid to assume that the gene rates are distributed according to the unimodal Γ distribution. To test this assumption, the distribution of gene rates across many multigene data sets was determined in order to avoid the problem of sampling error. Because DistR estimates have been shown to approximate ML gene rate estimates, a large number of DistR estimates taken from multiple data sets are likely to approximate the true distribution of gene rates.

Figure 3a shows the distribution of all the gene rates estimated over a number of data sets. The rates were estimated using the DistR method (Bevan et al., 2005), with varied size data sets in terms of number of species, number of genes, and number of missing distances. The maximum number of missing pairwise distances was about 50%, which is fairly substantial.

The Γ distribution provides an excellent fit to the data over many data sets (Fig. 3a). Thus, it is reasonable to assume that the rate of gene evolution is distributed according to a Γ distribution. It should be noted, however, that even in the case of large data sets with many genes it is possible that the Γ approximation will not be accurate (Fig. 3b). In such cases it might be better to use a mixture of Γ distributions over gene rates (as has been done for site rates in Mayrose et al., 2005). It is possible that using a better distribution will cause the α-parameter method to find a better fit to the data than the n-parameter method. However, this option was not explored in the current analysis. It is also possible that using a better method to approximate the Γ distribution will lead to more improvement of the fit under the α-parameter method. Laguerre integration was also used to approximate the distribution; however, in some cases this method had difficulty optimizing (data not shown). For the data sets that were successfully analyzed, this approach did not cause the α-parameter method to find a better fit to the data than the n-parameter method (data not shown).

DistR estimates versus ML estimates.—DistR estimates are used as initial approximations to the gene rates in the n-parameter method and to find an initial estimate of α in the α-parameter method. Thus it is important to determine how accurate these initial estimates are. Figure 4 shows the initial DistR estimates versus the final ML

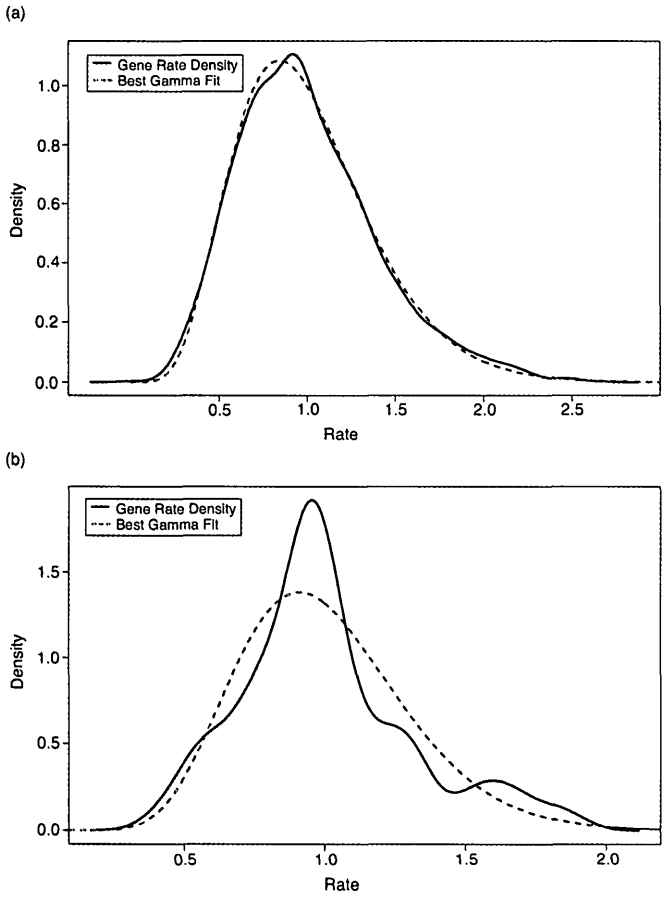

FIGURE 3. Density of estimated gene rates versus best fit of Γ distribution. (a) Density of gene rates estimated using DistR (solid line) versus best fit of Γ distribution (dashed line) for data described in Methods. (b) Density of gene rates estimated using DistR (solid line) versus best fit of Γ distribution (dashed line) for 133 genes over 44 species.

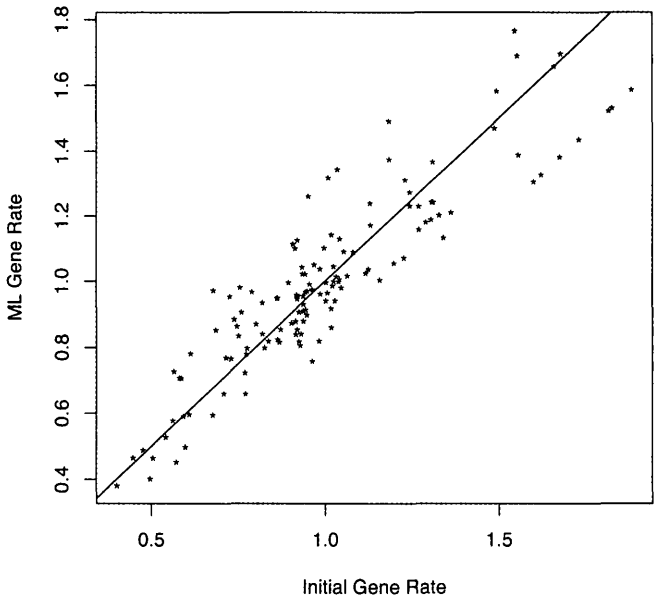

FIGURE 4. Correlation of the initial DistR gene rate estimate with the ML gene rate estimates. Maximum likelihood gene rates were estimated using the JTT model of evolution with a site rate Γ distribution for each gene and a proportion of invariant sites. Estimates are based on data sets in Table 1.

estimates from the five data sets in Table 1, estimated using the *n*-parameter method with gene-Γ. There is strong correlation between the two (Pearson's one-tailed correlation 0.904, $P < 2.2e^{-16}$), and the final ML parameter estimates are quite close to the starting DistR estimates. Furthermore, the initial DistR estimates appear to be unbiased with respect to the final ML estimates (i.e., the DistR rates are both over- and underestimates of the ML rates). Hence the DistR estimates provide an excellent starting point for the *n*-parameter method. This is especially important when many genes are involved in the analysis because less time needs to be spent searching the parameter space when good starting estimates are used.

In general, the initial α-parameter estimates were also quite close to the final estimates under the α-parameter method with gene-Γ. As expected, when more genes were present in the analysis, the initial estimate was more accurate. For example, both the fungal and animal mtDNA data sets have more than 10 genes, with relative errors of 0.0354 and 0.0263, respectively, between the initial and ML estimates of the α parameter. Conversely, both the Murphy and Madsen data sets have fewer genes (Table 1). The respective relative errors of the initial α estimates are 0.1871 and 0.1898. The relative error does not seem to be affected by the number of species since the animal mtDNA data set has the greatest number of species but the smallest relative error.

Topology Resolution under n*-Parameter and* α*-Parameter Methods*

Given that the α-parameter and *n*-parameter methods give potentially different ML topologies, even when there is little difference in the improved fit over the concatenated model, it is important to determine at what point the two methods provide congruent answers. Figure 5 shows the best (or favored) topology along with the branchings that prove difficult to resolve (in the data set of Rokas et al., 2003).

The different methods of accounting for gene rates leads to different bootstrap support. Additionally, adding more genes leads to an increase in support as shown by Rokas et al. (2003). The recovery of *S. bayanus* as sister to the in-group and *S. castellii* as sister to *S. bayanus* and the in-group was the most consistent in terms of improved bootstrap support with more genes for both the *n*-parameter and α-parameter methods. Conversely, the other two branches had inconsistent results across the two methods (Fig. 5). For example, when using 3 genes, the *S. mikatae–S. paradoxus–S. cerevisiae* clade starts with low support under the *n*-parameter method and then quickly reaches 90% support at 5 genes and

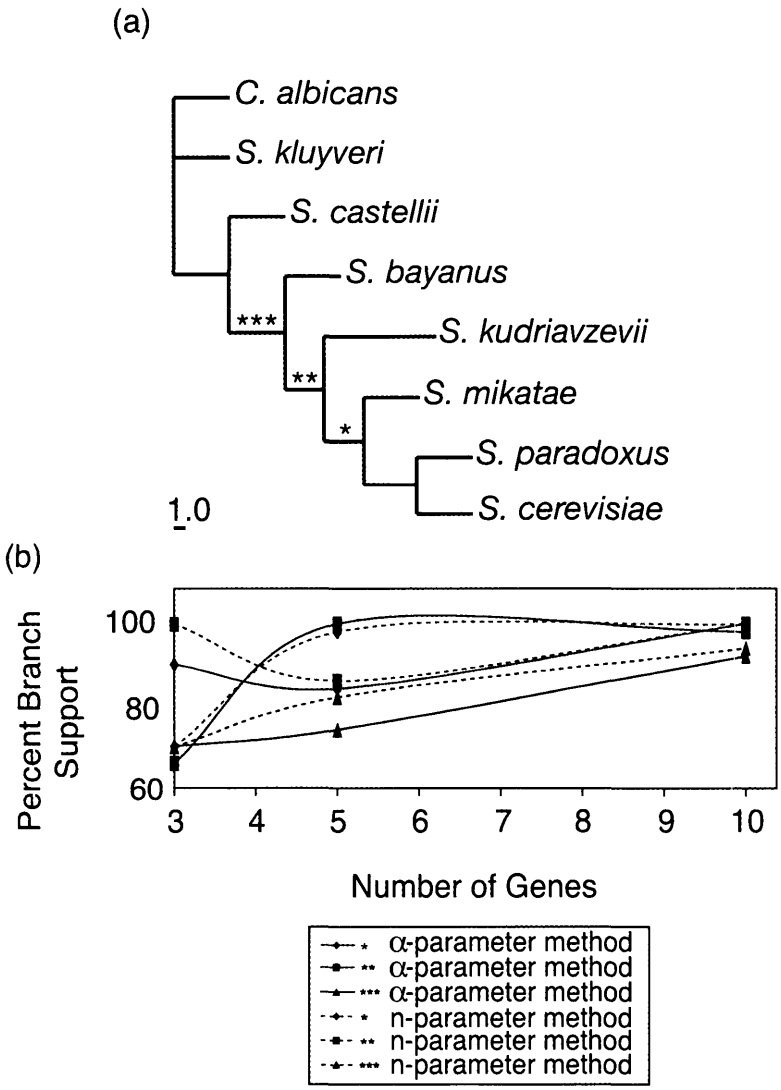

FIGURE 5. Bootstrap support for 8 fungal species under the n-parameter and α-parameter methods. (a) The correct tree topology. The branches that are difficult to resolve are labeled with *, **, and ***. (b) Bootstrap support for three branches for 3 genes, 5 genes, and 10 genes sampled 50 times each from 106 genes in total.

100% support at 10 genes. For the α-parameter method the opposite is true: at 3 genes the clade has 90% bootstrap support, which drops to just above 80% support with 5 genes and increases to 100% support with 10 genes. It should be noted, however, that these results may be due to model violations unrelated to gene rates, such as amino acid composition deviations from stationarity.

Evidently, in order to obtain a consistent ML tree between the two methods, more data are necessary. Thus, lack of data is one explanation for the inconsistent topologies found by the two methods (e.g., for the data in Table 1). This problem is exacerbated when more species are under analysis (as opposed to the 8 species used in this experiment). However, even with sufficient data, if model assumptions of the gene rates models are not valid (i.e., the model is misspecified with respect to true sequence evolution) then topology resolution artifacts can occur, even with sufficient data (see for instance Philippe et al., 2005).

Correlation of Gene Rates with Improved Fit under n-Parameter and α-Parameter Methods

The gene resampling experiment on the 106 gene Rokas data set not only provides information on how much data are necessary for both methods to provide congruent ML topologies, but it also demonstrates that for some data sets the gene rates methods have little or no improvement over the concatenated model (Fig. 2).

To determine what leads to an improved model fit of the gene rates model over the concatenated model, we calculated the correlation between the ΔAIC and three values: the rate of the slowest evolving gene in the data set; the rate of the fastest evolving gene in the data set; the difference between the rates of the fastest and slowest evolving genes in the data set. In order to compare gene rates properly across all resampled gene data sets, the gene rates were estimated over all 106 genes.

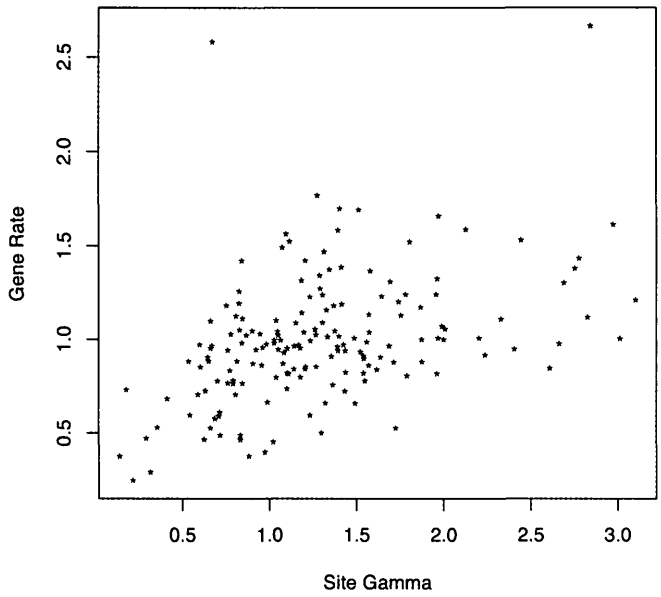

FIGURE 6. Correlation of the gene-Γ site rate heterogeneity parameter with the maximum likelihood rate of the gene. Maximum likelihood gene rates were estimated using the JTT model of evolution with a site rate Γ distribution for each gene and a proportion of invariant sites. Estimates are based on data sets in Table 1.

Correlation was tested only on the *n*-parameter and α-parameter methods with one-Γ distribution for site rates. This allows for the influence of the gene rates on the ΔAIC to be tested without the influence of separate Γ distributions for site rate heterogeneity for each gene. Results are given for the *n*-parameter method. Results based upon the α-parameter method were identical except in one data set where the concatenated model was preferred (Fig. 2).

Results (Table 3) demonstrate that it is not only the number of genes in the analysis which affects the improvement of the rates incorporated model over the concatenated model. Both the minimum rate in the analysis and the difference between the maximum and minimum rates also have a strong effect. Correlation values show that with fewer genes both the difference between the maximum and minimum rates, and the maximum rate are positively correlated with ΔAIC. Conversely, the minimum rate is negatively correlated with ΔAIC (Table 3). However, as the number of genes increases, correlation of ΔAIC with the maximum gene rate decreases and becomes statistically insignificant (Table 3). Correlation of the ΔAIC with the difference between the maximum and minimum rates also decreases, as does the statistical significance. Interestingly, the negative correlation of the ΔAIC with the minimum gene rate increases, as does the significance of the correlation (Table 3). Thus, although the difference between maximum and minimum rate (i.e., the degree of rate heterogeneity) is important for improved fit, it is not as important as the minimum rate of the gene under analysis.

TABLE 3. Correlation of ΔAIC of the *n*-parameter method of accounting for gene rate heterogeneity with: the minimum gene rate (Min(GR)); maximum gene rate (Max(GR)); the difference between the minimum and maximum gene rates (Max(GR) − Min(GR)). Gene rates were estimated globally over the 106 genes from which the data sets were sampled. As the number of genes under analysis increases, so does the correlation of the ΔAIC with the minimum gene rate. Conversely, both the maximum gene rate and the difference between the two become less correlated with the ΔAIC. ΔAIC values are based upon accounting for gene rate heterogeneity using the *n*-parameter with one-Γ distribution for site rate heterogeneity.

	3 Genes		5 Genes		10 Genes	
Statistic	ρ	*P*-value	ρ	*P*-value	ρ	*P*-value
Max(GR)− Min(GR)	0.812	$4.008e^{-13}$	0.662	$8.407e^{-8}$	0.485	$1.185e^{-4}$
Max(GR)	0.464	$3.398e^{-4}$	0.244	$4.376e^{-2}$	0.144	0.1595
Min(GR)	−0.682	$2.537e^{-8}$	−0.723	$1.055e^{-9}$	−0.774	$2.111e^{-11}$

The results indicate that it is the minimum gene rate that is the primary variable that determines whether there is improved model fit when using a model that accounts for gene rate heterogeneity. Indeed, a slower global minimum rate indicates that a higher improvement in the fit of the model to the data are likely when accounting for gene rates. This is partially due to the fact that a slower global rate will likely lead to a slower relative rate in the data set under analysis, and thus greater gene rate heterogeneity. However, if gene rate heterogeneity were the only factor influencing improved fit, we would expect to see that the correlation of improved fit with maximum rate would remain high (or at least significant) with more genes under analysis. This is because a faster global rate should also lead to greater gene rate heterogeneity. However, the maximum rate does not correlate with improved fit when there are more genes under analysis. Conversely, the minimum rate has a higher correlation with improved model fit when more genes are under analysis. Thus, the minimum rate of the gene has an effect upon the improved model fit, independent of the fact that a slower gene will likely lead to greater gene rate heterogeneity.

When the relative rates of the genes are used to test for correlation, the slowest evolving gene under analysis has an even more significant negative correlation with ΔAIC (−0.857, *P*-value of $1.064e^{-15}$ for data sets with 10 resampled genes). This correlation indicates that the DistR method can be run to test initial gene rates, and if there are very slow rates a much higher improved fit under the gene rates model can be expected.

Some analyses focus on eliminating fast sites/genes from phylogenetic analysis since these sites typically violate model assumptions or lead to long branch attraction (LBA; Hirt et al., 1999; Brinkmann and Philippe, 1999; Dacks et al., 2002; Brinkmann et al., 2005). It has also been noted that invariant sites can cause problems in phylogenetic reconstruction (Lockhart et al., 1996; Hirt et al., 1999; Dacks et al., 2002), leading to the removal of these sites from the analysis. This analysis indicates that properly accounting for the slow genes is quite important. Perhaps accounting for the slow genes correctly causes the invariant sites to no longer violate model assumptions by shortening the branch lengths, and thus increasing the

probability of no change over the branches. Conversely, given the low correlation of fast genes with improved model fit, fast sites which violate model assumptions (i.e., are saturated) probably continue to violate model assumptions.

Although correlations were tested over only one resampled data set, with few species, these results provide a preliminary indication that the more heterogeneous the data, the more likely an improvement will occur when accounting for the heterogeneity. This is especially true with few genes under analysis. However, as the number of genes increases this becomes less important than the evolutionary rate of the slowest gene.

Correlation of Gene Rate with Site Rate Heterogeneity

Given that accounting for site rate heterogeneity separately for each gene leads to a much better model fit, the question arises of whether or not there is any correlation between the rate of evolution of a gene and the ML estimate of the α parameter accounting for site rate heterogeneity. Figure 6 shows the gene rate versus the ML estimate of the α parameter estimated over the data sets in Table 1. The positive correlation (Pearson's one-tailed correlation 0.432, $P = 1.887e^{-14}$) is significant.

This result demonstrates that it is not evident that independently accounting for both gene rates and site rates within a gene is the best way to model the rate heterogeneity of all the sites. The rate of a site is here modeled based on both the site and gene rate heterogeneity. Yet there is only one absolute rate at which a given site evolves, ignoring rate heterogeneity through time. Clearly modeling this rate separately through site rate heterogeneity and gene rate heterogeneity is not completely correct. The correlation between the α parameter for site rate heterogeneity with the rate of evolution of the gene supports this conclusion. The gene rate parameter and the α parameter of the Γ distribution are dependent. Thus, to a certain extent the different parameters are modeling the same information in the data, even though the parameters are estimated independently of one another. Perhaps it is possible to use a model that will account for the correlation between the two, in order to find even better improvement of the model fit to the data.

CONCLUSIONS

In conclusion, given the current analysis, there is no reason to prefer the α-parameter method over the n-parameter method in phylogenetic inference. This is a promising result since it means that it is not necessary to use much additional computation time to find a good fit of a model with gene rates to the data. However, these analyses also suggest that there is further work to be done in improving rate heterogeneity modeling in maximum likelihood methods. Because there is no guarantee of an improved model fit, even with an increasing number of genes, and there is high correlation between α estimates of site rate heterogeneity and gene rate estimates, clearly there are problems with current approaches.

ACKNOWLEDGMENTS

We thank Trevor Bruen, Stéphane Guindon, Nicolas Lartillot, and Tim Collins for helpful comments on the manuscript. Thanks to Stéphane Guindon for kindly providing the source code of PHYML v2.2 for our use. Thanks to Joe Felsenstein for initially proposing the α-parameter approach to account for gene rate heterogeneity. Salary and support from the Canadian Institutes of Health Research (MOP 42475; BFL), the Canadian Institute for Advanced Research (CIAR; BFL), National Science and Engineering Research Council (NSERC grant 238975-01; DB), New Zealand Marsden Grant (DB), and supply of laboratory equipment and informatics infrastructure by Genome Quebec/Canada (BFL) is gratefully acknowledged. RBB is supported by an NSERC PGS-B scholarship and Genome Quebec.

REFERENCES

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

Bapteste, E., H. Brinkmann, J. A. Lee, D. V. Moore, C. W. Sensen, P. Gordon, L. Duruflé, T. Gaasterland, P. Lopez, M. Müller, and H. Philippe. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium, Entamoeba, and Mastigamoeba*. Proc. Nat. Acad. Sci. 99:1414–1419.

Bevan, R. B., B. F. Lang, and D. Bryant. 2005. Calculating the evolutionary rates of different genes: A fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. Syst. Biol. 54:900–915.

Brinkmann, H., and H. Philippe. 1999. Archaea sister group of bacteria? Indications from tree reconstruction aftifacts in ancient phylogenies. Mol. Biol. Evol. 16(6):817–825.

Brinkmann, H., M. van der Giezen, Y. Zhou, G. P. de Raucourt, and H. Philippe. 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics. Syst. Biol. 54(5):743–757.

Bull, J., J. P. Huelsenbeck, C. W. Cunningham, D. L. Swofford, and P. J. Waddell. 1993. Partitioning and combining data in phylogenetic analysis. Syst. Biol. 42(3):384–397.

Burnham, K. P., and D. R. Anderson. 2003. Model selection and multimodel inference: A practical information–theoretic approach. Springer-Verlag, New York.

Collins, T. M., O. Fedrigo, and G. J. Naylor. 2005. Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics. Syst. Biol. 54:493–500.

Cranston, K., and B. Rannala. 2005. Closing the gap between rocks and clocks. Heredity 94:461–462.

Dacks, J. B., A. Marinets, W. F. Doolittle, T. Cavalier-Smith, and J. M. Logsdon Jr. 2002. Analyses of RNA polymerase II genes from free–living protists: Phylogeny, long branch attraction and the eukaryotic big bang. Mol. Biol. Evol. 19:830–840.

Felsenstein, J. 2001 Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. 53:447–455.

Felsenstein, J. 2004. Pages 537–538 *in* Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Foster, P. G. 2004. Modeling compositional heterogeneity. Syst. Biol. 53:485–495.

Gontcharov, A. A., B. Marin, and M. Melkonian. 2004. Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and *rbc*L sequence comparisons in the Zygnematophyceae (Streptophyta). Mol. Biol. Evol. 21:612–624.

Guindon, S., and O. Gascuel. 2003. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst. Biol. 52:696–704.

Harlow, T. J., J. P. Gogarten, and M. A. Ragan. 2004. A hybrid clustering approach to recognition of protein families in 114 microbial genomes. BMC Bioinformat. 5:1–14.

Hirt, R. P., J. M. Logsdon Jr., B. Healy, M. W. Dorey, W. F. Doolittle, and T. M. Embley. 1999. Microsporidia are related to fungi: Evidence from the largest subunit of RNA polymerase II and other proteins. Proc. Nat. Acad. Sci. 96:580–585.

Huelsenbeck, J. P., J. Bull, and C. W. Cunningham. 1996. Combining data in phylogenetic analysis. Tree 11:152–158.

Lockhart, P. J., A. W. D. Larkum, M. A. Steel, P. J. Waddell, and D. Penny. 1996. Evolution of chlorophyll and bacteriochlorophyll: The

problem of invariant sites in sequence analysis. Proc. Nat. Acad. Sci. 93:1930–1934.
Madsen, O., M. Scally, C. J. Douady, D. J. Kao, R. W. DeBry, R. Adkins, H. M. Amrine, M. J. Stanhope, W. W. de Jong, and M. S. Springer. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature 409:610–614.
Mayrose, I., N. Friedman, and T. Pupko. 2005. A Gamma mixture model better accounts for among site rate heterogeneity. Bioinformatics 21:ii151–ii158.
Murphy, W. J., E. Eizirik, W. E. Johnson, Y. P. Zhang, O. A. Ryder, and S. J. O'Brien. 2001. Molecular phylogenetics and the origins of placental mammals. Nature 409:614–618.
Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, and J. L. Nieves-Aldrey. 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53:47–67.
Philippe, H., Y. Zhou, H. Brinkmann, N. Rodrigue, and F. Delsuc. 2005. Heterotachy and long-branch attraction in phylogenetics. BMC Evol. Biol. 5:1–8.
Phillips, M., F. Delsuc, and D. Penny. 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Biol. Evol. 21:1455–1458.
Pupko, T., D. Huchon, Y. Cao, N. Okada, and M. Hasegawa. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Mol. Biol. Evol. 19:2294–2307.
Rokas, A., and S. B. Carroll. 2005. More genes of more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22:1337–1344.
Rokas, A., B. L. Williams, N. King, and S. B. Carroll. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.
Smyth, P. 2000. Model selection for probabilistic clustering using cross-validated likelihood. Stat. Comput. 10:63–72.
Springer, M. S., M. J. Stanhope, O. Madsen, and W. W. de Jong. 2004. Molecules consolidate the placental mammal tree. Trends Ecol. Evol. 19:430–438.
Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42:587–596.

First submitted 4 April 2006; reviews returned 20 June 2006; final acceptance 15 October 2006
Associate Editor: Tim Collins

APPENDIX

Calculating the Log-Likelihood of a Gene when Integrating over Gene Rates

The computation of likelihood in the α-parameter approach is critically sensitive to numerical error. Here we outline the approach taken to avoid round-off error. All calculations are for gene g. Let $LogL_{\alpha,k}$ be the log-likelihood of gene g for category k of the probability density h over gene rates from (7). There are C categories that approximate distribution h. Scale factor SF is used to prevent overflow and underflow errors. SF is the maximum of the log-likelihoods ($LogL_{\alpha,k}$) over all categories $k \in C$. $SLogL_{\alpha,k}$ is the $LogL_{\alpha,k}$ of gene g for category k, scaled by both the scale factor SF and the log of the probability of rate category $\hat{R}_k$. SL_g is the total scaled likelihood of gene g and $LogSL_g$ is the total scaled log-likelihood of gene g. Overall, the likelihood of gene g is computed as follows:

Compute $LogL_{\alpha,k} = Log(L_g)$ where $R_g = \hat{R}_k$ is 3 for each category C, using all sites i in gene g. This results in $LogL_{\alpha,1}, \ldots, LogL_{\alpha,C}$. Next calculate $SF = \max_k LogL_{\alpha,k}$ and $SLogL_{\alpha,k} = LogL_{\alpha,k} - SF\text{-}1 + \log p(\hat{R}_k)$ for $k = 1, \ldots, C$. This scaling is performed in order prevent over- and underflow errors. Thus,

$$SLogL_{\alpha,k} = \log\left[\frac{p(\hat{R}_k)\prod_{\text{site } i \in g}\sum_{j=1}^{S} p(\hat{r}_j)P(g_i \mid T, \hat{r}_j, \hat{R}_k)}{e^{SF+1}}\right] \quad (8)$$

from Equation 7.

From (8) compute $e^{SLogL_{\alpha,k}}$ in order to calculate the scaled likelihood

$$\frac{p(\hat{R}_k)\prod_{\text{site } i \in g}\sum_{j=1}^{S} p(\hat{r}_j)P(g_i \mid T, \hat{r}_j, \hat{R}_k)}{e^{SF+1}}$$

for every category $k = 1, \ldots, C$. Thus, the total scaled likelihood is

$$SL_g = \sum_{k=1}^{C} \frac{p(\hat{R}_k)\prod_{\text{site } i \in g}\sum_{j=1}^{S} p(\hat{r}_j)P(g_i \mid T, \hat{r}_j, \hat{R}_k)}{e^{SF+1}} \quad (9)$$

The scaled log-likelihood for gene g is computed from (9) as

$$LogSL_g = \log\left(\sum_{k=1}^{C} e^{SLogL_{\alpha,k}}\right) - \log(e^{SF+1}) \quad (10)$$

Solve for the log-likelihood of gene g from Equation 10 as

$$log(L_g) = LogSL_g + SF + 1$$

Note that the smallest scaled log-likelihood $LogSL_{\alpha,k}$ value possible that will not result in over- or underflow is approximately −707 (where the smallest signed number than can be expressed with a double is $2.225074e^{-308}$). Thus when the scaled log-likelihood is less than −707, it is set to −707, essentially setting the likelihood for this rate category to 0. This means that for that particular gene rate category, the probability of the data, given the rate and other parameters, approached 0.