Efficient Based Estimation of Evolutionary Distance when Substitution Rates Vary Across Sites

Stéphane Guindon, Olivier Gascuel

Efficient Biased Estimation of Evolutionary Distances When Substitution Rates Vary Across Sites

Stéphane Guindon and Olivier Gascuel

LIRMM, UMR 9928 Université Montpellier II/CNRS

This paper deals with phylogenetic inference when the variability of substitution rates across sites (VRAS) is modeled by a gamma distribution. We show that underestimating VRAS, which results in underestimates for the evolutionary distances between sequences, usually improves the topological accuracy of phylogenetic tree inference by distance-based methods, especially when the molecular clock holds. We propose a method to estimate the gamma shape parameter value which is most suited for tree topology inference, given the sequences at hand. This method is based on the pairwise evolutionary distances between sequences and allows one to reconstruct the phylogeny of a high number of taxa ($>$1,000). Simulation results show that the topological accuracy is highly improved when using the gamma shape parameter value given by our method, compared with the true (unknown) value which was used to generate the data. Furthermore, when VRAS is high, the topological accuracy of our distance-based method is better than that of a maximum likelihood approach. Finally, a data set of Maoricicada species sequences is analyzed, which confirms the advantage of our method.

Introduction

Most of the phylogenetic inference methods use an explicit model of sequence evolution. Such a model includes parameters whose values must be estimated. Among these parameters, the variability of substitution rates across sites (VRAS) has been widely studied in the past and remains an important subject in the phylogenetic tree inference domain. Indeed, VRAS is widespread among biological sequences. For example, Sullivan, Holsinger, and Simon (1995) and Yang and Kumar (1996) provided evidence that VRAS occurs in rodent 12S RNA and the D-loop sequences in mitochondrial genomes of many different vertebrates. Rzhetsky, Kumar, and Nei (1995) also built a specific model to describe VRAS among 16S-like ribosomal RNAs. Furthermore, VRAS has a strong effect on tree inference. Yang (1993) and Yang, Goldman, and Friday (1994) have demonstrated a significant improvement of the maximum likelihood (ML) approach (Felsenstein 1981) when the model of sequence evolution incorporates VRAS. The distance-based methods also suffer from this phenomenon. Tateno, Takezaki, and Nei (1994), using simulations with 4-taxon trees, demonstrated a poor robustness of the neighbor-joining method (Saitou and Nei 1987) when VRAS occurs but is not taken into account.

The gamma distribution is most commonly used for modeling rate variation across sites. The shape of this distribution is related to a parameter denoted as a in the text that follows. When a is less than 1, the density function is exponential-like and VRAS is high. Higher values of a (say $>$2) represent weak variations of substitution rates across sites. When a tends to infinity, all sites evolve at the same rate.

Key words: phylogenetic reconstruction, varying rates of substitution, distance methods, maximum likelihood, computer simulations, Maoricicada.

Distances between sequences can be analytically expressed for certain models of sequence evolution, depending on the gamma shape parameter. For the Kimura two-parameter model (K80) (Kimura 1980), the evolutionary distance between two sequences is given by (Jin and Nei 1990):

$$d = \frac{a}{2}\left[(1 - 2P - Q)^{-1/a} + \frac{1}{2}(1 - 2Q)^{-1/a} - \frac{3}{2}\right], \quad (1)$$

where P and Q are the probabilities to observe a transition and a transversion, respectively. An estimate of d is obtained by replacing P and Q by the frequencies of observed transitions and transversions and a by an estimate denoted as α. The expression given previously shows that, for any fixed values of P and Q, d is a decreasing function of a. Hence, when α overestimates a ($\alpha > a$), the evolutionary distance is underestimated.

Both likelihood and parsimony methods have been used to estimate the value of a. Yang (1993) extended the method of Felsenstein (1981) and included VRAS in the ML framework. The estimation of a is usually performed given a specific tree topology. However, when the correct topology is unknown, it is possible to alternate the estimation of a and the tree topology reconstruction, given the value of a. The procedure is stopped when the tree topology does not change between two steps. Unfortunately, this approach involves intensive computation and is only feasible for small data sets (say 30–40 taxa).

The estimation of a in the maximum parsimony framework also relies on a given tree topology, which is supposed to be correct. The computational burden is clearly less than that of ML. Unfortunately, the values of α obtained with this method are not reliable. Indeed, as the number of substitutions between taxa is underestimated, VRAS is underestimated too, and the value of a is overestimated.

The present paper is organized into two parts. The first deals with the best value of α for tree inference using distances. The best or optimal value of α is the value which minimizes the difference between the inferred tree topology and the true topology. Using sim-

ulations we show that evolutionary distances estimated from the true value of the gamma shape parameter are not optimal; underestimated distances provide a better topological accuracy and outperform usual unbiased distances.

In the second part of the paper, we present a method to estimate the optimal value of α. This approach is based on distance algorithms and allows one to deal with numerous taxa (say >1,000). We use simulations and real data to test the accuracy of the method. The results are presented, and finally, we discuss our approach and directions for future research.

The True Value of a is not Optimal

In this section we focus on the topological distance between the true tree and the inferred tree and how this depends on the value of α. We first describe our simulations and the results thereafter.

Simulations

A true phylogeny, denoted as T, was first generated using the stochastic speciation process described by Kuhner and Felsenstein (1994). The number of taxa was set to 20 and the branch length expectation to 0.03 mutations per site. Using this generating process makes T ultrametric (or molecular clock–like). This hypothesis does not hold in most biological data sets, so we created a deviation from the molecular clock. Every branch length of T was multiplied by a gamma distributed factor. The mean of the gamma distribution used was equal to 1.0 and the shape parameter, denoted as η, was set to 0.5 or 2.0. The ratio between the mutation rate in the fastest evolving lineage and the rate in the slowest evolving lineage was equal to 3.6 and 2.0, respectively. Therefore, $\eta = 0.5$ corresponds to a strong departure from the molecular clock, and $\eta = 2.0$ to a mild departure. The mean distance between two taxa in such phylogenies is not related to η and is approximately equal to 0.2.

For each T thus obtained, a unique set of 1,000-bp sequences was produced, given the pattern of speciation events and branch lengths described by the tree. The K80 model was used, with site to site rate variation following a gamma distribution. The sequences were generated using Seq-Gen (Rambaut and Grassly 1997), with a transition-transversion ratio (TS/TV) of 2.0 and equal base frequencies. Two values for a have been tested: 0.1 and 0.7. These values correspond to the first and the third quartiles of the distribution of a series of ML estimates of a, which were obtained from the analysis of 16 data sets by Yang (1996). Therefore, 0.1 represents a rather high VRAS, whereas 0.7 corresponds to a medium-low VRAS.

For each sequence set so obtained, several matrices (δ_{ij}^{α}) were computed, depending on the α value used to correct the distances. The values of α flanked the true value a. For $a = 0.1$, the values of α lay between 0.09 and 2.0, whereas for $a = 0.7$ the values of α lay between 0.6 and 4.0.

For each distance matrix (δ_{ij}^{α}), a phylogeny, denoted as $\mathcal{T}^{\alpha}$, was inferred using BIONJ (Gascuel 1997). Simulations have been done with other tree building methods, but the results were similar to those presented in this paper. The topology of $\mathcal{T}^{\alpha}$ was then compared with that of the true tree T using a topological distance equivalent to that of Robinson and Foulds (1979). It is defined by the proportion of internal branches (or bipartitions) that are found in one tree and not in the other one. This distance varies between 0.0 (both topologies are identical) and 1.0 (they do not share any internal branch). The Robinson and Foulds distance between T and $\mathcal{T}^{\alpha}$ is denoted as $RF(T, \mathcal{T}^{\alpha})$ in the text that follows.

We then defined the optimal value of α as the value that minimizes the mean of $RF(T, \mathcal{T}^{\alpha})$, denoted as $\overline{RF}(T, \mathcal{T}^{\alpha})$, given the experimental condition at hand (corresponding here to the values of η and a). This optimal value is denoted as α^{opt} and is formally defined as:

$$\alpha^{opt} = \underset{\alpha \in \mathbb{R}^+}{\mathrm{argmin}} \, (\overline{RF}(T, \mathcal{T}^{\alpha})). \qquad (2)$$

Therefore, α^{opt} corresponds to the value that ensures the lowest average topological distance between the true tree T and the inferred tree $\mathcal{T}^{\alpha}$, given the conditions at hand.

Results

Figure 1 shows the mean topological distance between the true tree and the inferred tree ($\overline{RF}(T, \mathcal{T}^{\alpha})$) as a function of the value of α. When the deviation from the molecular clock is strong ($\eta = 0.5$), α^{opt} is close to a but remains systematically higher. The difference between α^{opt} and a increases when the molecular clock is better satisfied ($\eta = 2.0$). When the molecular clock holds (results not shown), $\overline{RF}(T, \mathcal{T}^{\alpha})$ is a monotonic decreasing function of α, and α^{opt} tends to infinity. In this case, the best topological accuracy is obtained using noncorrected distances, even if VRAS occurs in sequences.

Therefore, underestimated distances outperform unbiased distances when the molecular clock holds. Steel and Penny (2000) showed that, in this case, the correct topology is induced by any monotonic increasing function of the true distances, in particular the Hamming distance between infinite length sequences. Hence, when the noise affecting distance estimates is sufficiently low, the true topology can be retrieved with a high probability even if the distances are not corrected or underestimated. However, this property does not hold when the true distances are not ultrametric, i.e., when the molecular clock is not satisfied.

Such a demonstration explains why correct tree topologies can be retrieved with biased distances. However, it does not explain why, when the molecular clock holds, underestimated distances provide a better topological accuracy than unbiased distances. A widespread idea is that this phenomenon is caused by a decrease in the variance of the distance estimates (Saitou and Nei 1987; Sourdis and Nei 1988; Zharkikh and Li 1993; Schöniger and von Haesler 1993; Tajima and Takezaki 1994; Takahashi and Nei 2000). Because overestimating
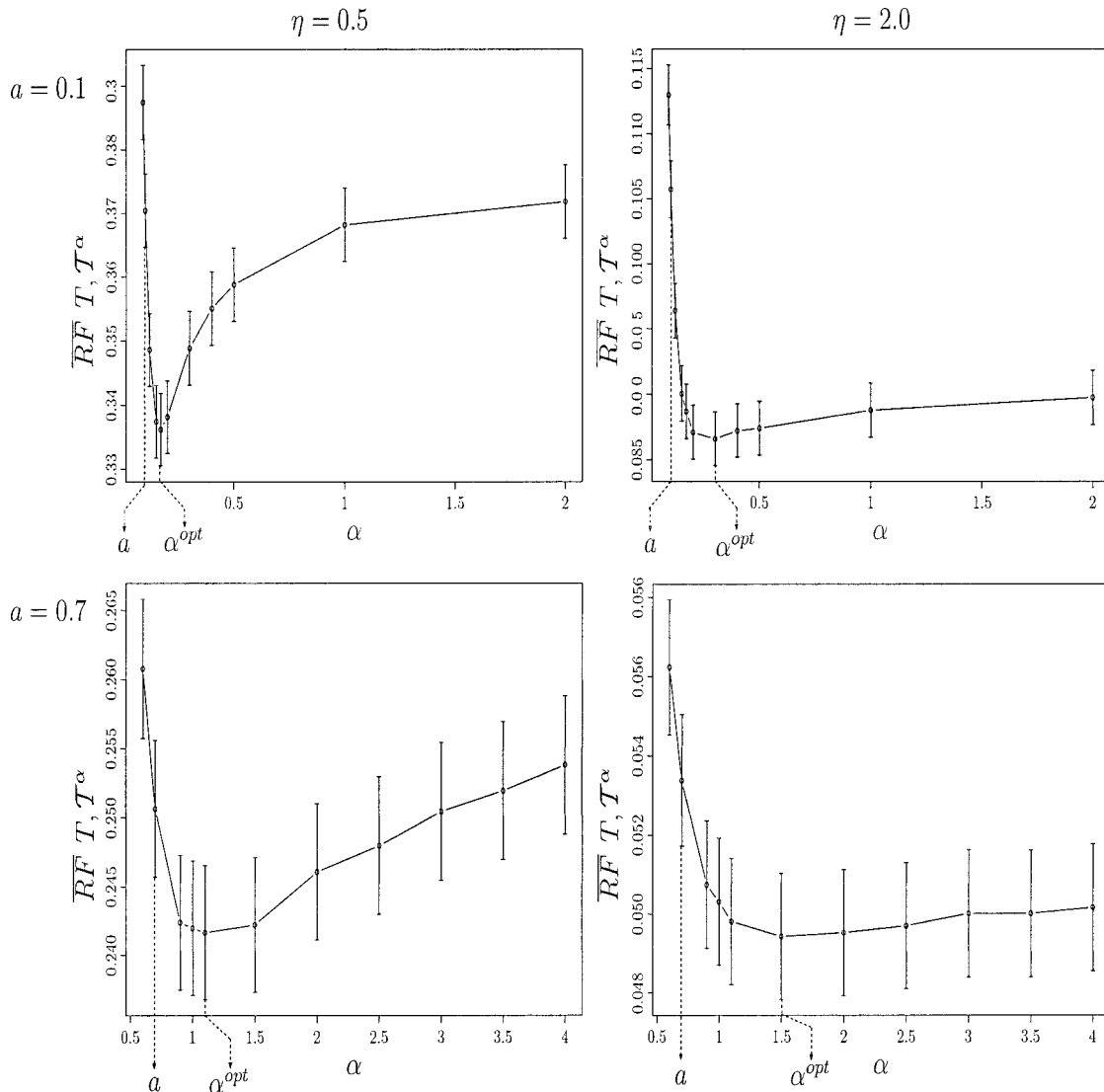
FIG. 1.—Topological distance between the true and inferred trees as a function of the α value used to estimate the distances. $\overline{RF}(T, \mathcal{T}^\alpha)$, average Robinson and Foulds distances between T and $\mathcal{T}^\alpha$; a, true value of the gamma distribution parameter, which is used to generate the data; η, parameter measuring the deviation from molecular clock; α^{opt}, value of α which minimizes $\overline{RF}(T, \mathcal{T}^\alpha)$. Each value of $\overline{RF}(T, \mathcal{T}^\alpha)$ is found by averaging over 1,000 simulated 20-taxon data sets.

a leads to underestimating distances, hence, to a decrease in the variances of the estimates, this explanation could hold there. However, this point remains to be formally demonstrated.

Another interesting point is the comparison between curves for $a = 0.1$ and $a = 0.7$. The region surrounding α^{opt} is indeed much flatter for $a = 0.7$ than for $a = 0.1$. This phenomenon is caused by a shape property of the gamma distribution. When a is small (e.g., near 0.1), the variation of α around a induces a strong variation of distance estimates, and perturbations of tree topologies follow. When a is higher (e.g., =0.7), the variation of α around a produces a small variation of distance estimates, and tree topologies remain more stable. In this case, a large range of values of α around α^{opt} give the same topology as the one obtained with α^{opt}.

In conclusion, the optimal value of the gamma distribution parameter is always higher than the real value

of this parameter, and this deviation is the largest when the molecular clock holds.

Approximating α^{opt}

As the topology of T is unknown and represents what is searched for, the value of α^{opt} cannot be estimated from equation (2). We propose in this section a criterion, denoted as Q, to approximate α^{opt}. Q measures the reliability of the inferred tree. The approximation of α^{opt} is denoted as α^* and corresponds to the most reliable tree in the sense of Q. The formal definition of α^* is analogous to equation (2), that is,

$$\alpha^* = \underset{\alpha \in \mathbb{R}^+}{\operatorname{argmin}} \, (Q((\delta_{ij}^\alpha), \mathcal{T}^\alpha)). \qquad (3)$$

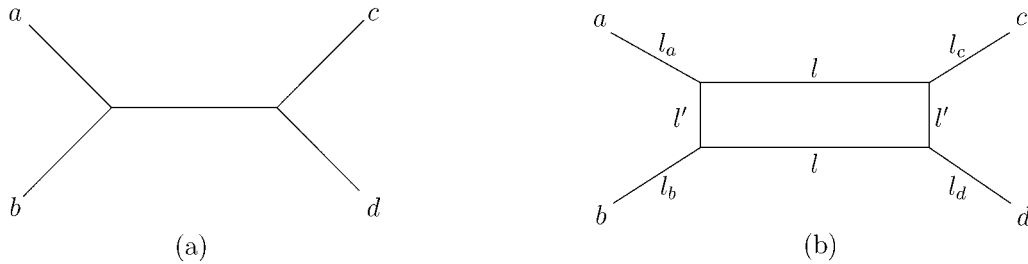We first describe the computation of Q with four taxa and then for a higher number of taxa. The average ac-

Fig. 2.—Exact representation of estimated distances. *a*, inferred tree. *b*, exact graphical representation of the six estimated distances δ_{ab}, δ_{cd}, δ_{ac}, δ_{bd}, δ_{ad}, and δ_{bc}; we have $L = \delta_{ad} + \delta_{bc} = (l_a + l + l' + l_d) + (l_b + l' + l + l_c)$, $M = \delta_{ac} + \delta_{bd} = (l_a + l + l_c) + (l_b + l + l_d)$, and $Q = L - M = 2l'$.

curacy of Q is found using simulations. Finally, we present a new tree inference method based on this criterion.

Definition of the Criterion with 4-Taxon Trees

Take four taxa denoted as *a, b, c,* and *d,* and the six distances δ_{ab}, δ_{ac}, δ_{ad}, δ_{bc}, δ_{bd}, and δ_{cd}. Assume: $(\delta_{ab} + \delta_{cd}) < (\delta_{ac} + \delta_{bd}) \leq (\delta_{ad} + \delta_{bc})$. The three terms of this inequality are denoted as S (Small), M (Median) and L (Large), respectively. Given this inequality, most of the distance-based methods (in particular BIONJ that is used here) infer the same unrooted topology, denoted as {*a, b*}/{*c, d*} and shown in figure 2*a*. In this case, S can also be defined as the sum of the distances between the two external pairs (external pairs are made of two taxa separated by a single node).

Because of random noise, the fit of the distance estimates to a tree distance is almost always imperfect. However, the graph (Bandelt and Dress 1992) of figure 2*b* provides an exact representation of the six distance estimates. In this graph, the distance between two taxa is equal to the length of the path that separates them, e.g., $\delta_{ad} = l_a + l + l' + l_d$. The set of equations, in which each estimated distance is expressed as a sum of edge lengths, has six degrees of freedom corresponding to l, l', l_a, l_b, l_c, and l_d. Hence, one can express the edge lengths as linear combinations of the distances. In particular, $l = (L - S)/2$ and $l' = (L - M)/2$.

When the fit of the distance estimates to the tree {*a, b*}/{*c, d*} is perfect, $l' = 0$ and the graph of figure 2 becomes a tree. In this case, the 4-point condition (Zaretskii 1965; Buneman 1971) holds, and $L = M$. As explained previously, this situation is not encountered in most real data sets and the edge l' has a positive length.
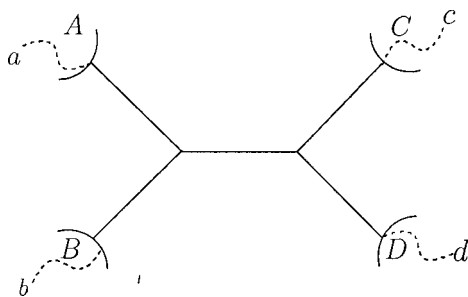
If l' is small compared with l, the support for the topology {*a, b*}/{*c, d*} is higher than that for {*a, c*}/{*b, d*}. If l and l' are close, one cannot clearly choose between {*a, b*}/{*c, d*} and {*a, c*}/{*b, d*}. Note that this uncertainly is not necessarily translated into a small internal branch length in the inferred tree (at least, when using least squares branch length estimates). If $l \simeq l' \simeq 0$, the internal edge of the inferred tree is close to zero, and the data support a star tree.

The Q criterion is then:

$$Q = 2l' = L - M. \tag{4}$$

Hence, Q assesses the reliability of the inferred internal edge. This criterion also measures the fit of the distance estimates to a tree distance: the larger the value of Q, the more the distance estimates differ from a tree distance.

Definition of the Criterion with *n*-Taxon Trees

Let *n*, the number of taxa, be larger than four. Each of the $n - 3$ internal branches of the inferred tree defines four subtrees, denoted as *A, B, C,* and *D* (fig. 3). Let $\overline{\delta_{AB}}$ be the mean of the estimated distances between subtree *A* and subtree *B*, i.e.,

$$\overline{\delta_{AB}} = \frac{\sum\limits_{a \in A} \sum\limits_{b \in B} \delta_{ab}}{n_A \cdot n_B},$$

where n_A and n_B are the numbers of taxa in subtrees *A* and *B*, respectively (fig. 3). Let $\overline{\delta_{AC}}$, $\overline{\delta_{AD}}$, $\overline{\delta_{BC}}$, $\overline{\delta_{BD}}$, and $\overline{\delta_{CD}}$ be defined in the same way. S, M, and L now correspond to $(\overline{\delta_{AB}} + \overline{\delta_{CD}})$, $(\overline{\delta_{AC}} + \overline{\delta_{BD}})$, and $(\overline{\delta_{BC}} + \overline{\delta_{AD}})$, respectively. S is then defined by both external pairs. S is also the smallest of the three sums in most practical cases (99% of cases with the data sets used in the previous section, when inferring the trees with BIONJ).

The value of the criterion for the focused edge is then obtained using equation (4). The value of the criterion for the whole tree is equal to its mean value for every internal branch. However, as the criterion only makes sense when branches have positive length, the negative or null branches are not taken into account. Q is then a global measure of internal branch reliability. The value of Q is null when the distances are tree-like. Therefore, assuming that the evolutionary model used to estimate the distances is satisfied, α^* converges to the true value a of the gamma shape parameter when the sequence length increases.
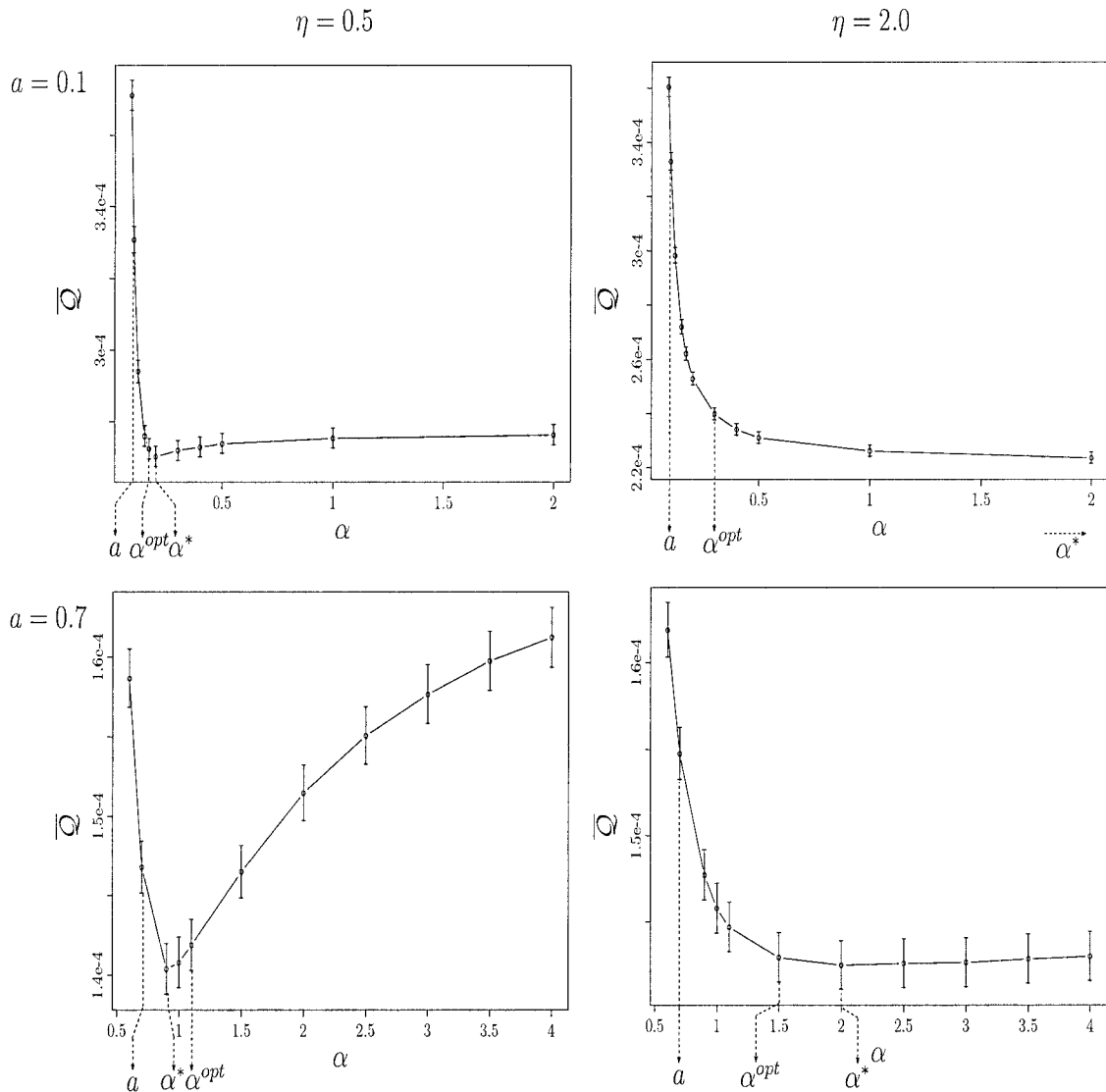

Fig. 3.—Subtrees associated with an internal edge. Each internal edge is associated with four subtrees denoted *A, B, C,* and *D; a, b, c,* and *d* are taxa belonging to these subtrees.

FIG. 4.—Mean value of the Q criterion depending on the α value used to estimate the distances. α^* is the value of α that minimizes $\overline{Q}$. α^{opt} is the value of α that minimizes $\overline{RF}(T, \ \mathcal{T}^\alpha)$. Results are based on 1,000 simulated 20-taxon data sets for each combination of η and a.

The time complexity of the computation of Q for one branch is equal to $O(n^2)$ in the worst case ($n_A = n_B = n_C = n_D = n/4$). The worst case complexity for n taxa is then $O(n^3)$, but in practice it is often lower. This worst-case time complexity is equal to that of NJ-like tree building algorithms, so the Q criterion can be used with large data sets. For example, with $n = 500$, the computing time to build a tree using BIONJ is equal to 11.43 s, whereas the time to compute Q is equal to 2.21 s (PentiumIII, 750 MHz).

Mean Performance of Q in Approximating α^{opt} Using Simulations

The performance of Q is shown in figure 4. The curves are obtained in the same manner as the ones in figure 1; but instead of the Robinson and Foulds distance, the ordinate reports now the value of the Q criterion. This value is averaged over 1,000 data sets for each experimental condition, and α^* is obtained by con-

sidering the mean values of Q and not a single value as used in equation (4). Therefore, figure 4 provides a view on the mean accuracy of Q in approximating α^{opt}.

The curves of figures 4 and 1 are similar, and $\overline{Q}$ appears to be relatively accurate in approximating α^{opt}. However, when $a = 0.1$ and $\eta = 2.0$, which corresponds to a strong VRAS and a moderate deviation from the molecular clock, the curve of figure 4 is a monotonic decreasing function and α^* tends to infinity, whereas $\alpha^{opt} \simeq 0.3$. In spite of this difference, $\overline{RF}(T, \ \mathcal{T}^{\alpha^*}) = 0.1893$ is close to $\overline{RF}(T, \ \mathcal{T}^{\alpha^{opt}}) = 0.1951$, whereas the accuracy obtained with a is much lower: $\overline{RF}(T, \ \mathcal{T}^a) = 0.2296$. Indeed, table 1 indicates that $\overline{RF}(T, \ \mathcal{T}^{\alpha^*})$ is always inferior to $\overline{RF}(T, \ \mathcal{T}^a)$ and very close to $\overline{RF}(T, \ \mathcal{T}^{\alpha^{opt}})$. Therefore, the performance of α^* in reconstructing T is similar to that of α^{opt}, even when α^* is remote from α^{opt}. However, results in table 1 have to be interpreted carefully as the values of α^* are obtained from 1,000 data sets which are generated under the same evolutionary conditions, whereas parameter estimation in

Table 1.
Topological Accuracy of the Inferred Tree when Distances are Corrected with a, α^{opt} or α^*

Γ param.		$\eta = 0.5$	$\eta = 2.0$
$a = 0.1$	$\overline{RF}\,(T, \mathcal{T}^a)$	0.3704	0.2296
	$\overline{RF}\,(T, \mathcal{T}^{\alpha^{opt}})$	0.3362	0.1893
	$\overline{RF}\,(T, \mathcal{T}^{\alpha^*})$	0.3381	0.1951
$a = 0.7$	$\overline{RF}\,(T, \mathcal{T}^a)$	0.2506	0.1134
	$\overline{RF}\,(T, \mathcal{T}^{\alpha^{opt}})$	0.2416	0.1122
	$\overline{RF}\,(T, \mathcal{T}^{\alpha^*})$	0.2424	0.1127

NOTE.—η, parameter measuring the deviation from molecular clock; a, true value of the gamma distribution parameter, which is used to generate the data; α^{opt}, optimal value of α; α^*, our approximation of α^{opt}; $\overline{RF}\,(T, \mathcal{T}^a)$, $\overline{RF}\,(T, \mathcal{T}^{\alpha^{opt}})$, and $\overline{RF}\,(T, \mathcal{T}^{\alpha^*})$, average topological accuracies that are obtained with a, α^{opt}, and α^*, respectively. Results are based on 1,000 simulated 20-taxon data sets for each combination of η and a.

the frame of phylogenetic inference is done from a single data set. A better view of the performance of Q is given in the results section.

It must be underlined that numerous other criteria have been tested in this study (e.g., Eigen and Winkler-Oswatitsch 1981; Vach 1992; Guénoche and Garreta 2001), but none of these performed as well as Q.

Using Q for Phylogenetic Inference

Given a set of homologous sequences, several (δ_{ij}^{α}) distance matrices are computed. The α values are obtained from a predefined sample with size r. In this study, the r values of α ranged from 0.1 to 5,000. Between 0.1 and 3.0, the step was equal to 0.02, between 3.0 and 10, to 0.1, whereas the remaining α values were 10, 50, 100, 500, 1,000 and 5,000. These increasing steps are explained by the necessity to concentrate on the area where a small variation of α likely involves some perturbations in the inferred topology. The calculation of the different (δ_{ij}^{α}) matrices is very fast. Indeed, the transition, transversion, and identity frequencies are computed only once, which requires $O(n^2l)$ computing time where l is the sequence length. The (δ_{ij}^{α}) distances matrices are obtained by correcting these three frequencies with the corresponding α values using equation (1) in the case of K80 model; the computational burden for the r matrices is then equal to $O(n^2r)$. The $\mathcal{T}^{\alpha}$ phylogenies are inferred from the (δ_{ij}^{α}) distance matrices using BIONJ (Gascuel 1997). The values of Q for the various values of α are then computed using both the (δ_{ij}^{α})'s and $\mathcal{T}^{\alpha}$'s. Finally, we select the tree $\mathcal{T}^{\alpha^*}$ that minimizes $Q((\delta_{ij}^{\alpha}), \mathcal{T}^{\alpha})$ among the r inferred trees. The whole time complexity is equal to $O(n^2l + n^2r + n^3r)$, where the three terms correspond to: (1) counting the observed mutations, (2) computing the distance matrices, and (3) inferring the trees and computing Q. Practical computing times are given in the next section, and a PHYLIP compatible program, called GAMMA, is available from http://www.lirmm.fr/~w3ifa/MAAS/.

Results

We first compare the performance which is obtained using our approximation, α^*, to the performance

that would be obtained if the true value of a was known. Then, we compare the topological accuracy of our method with the one of ML (Felsenstein 1981; Yang 1993). Finally, we illustrate our approach using sequences from Maoricicada species (Buckley, Simon, and Chambers 2001).

$\mathcal{T}^{\alpha^*}$ versus $\mathcal{T}^a$

We performed simulations in a way similar to that described previously. Three deviations from the molecular clock were used: $\eta = 0.5$ and $\eta = 2.0$, as previously, while the molecular clock (MC) held in the third case. The evolution of the sequences along the trees was simulated using three values of the a gamma shape parameter: 0.1, 0.7, and 2.0. The sequences were 300 or 1,000 bp long, and each data set contained 20 taxa. For each of these data sets, two trees were inferred. The first was built with BIONJ from the (δ_{ij}^a) matrix, where a was the value used to generate the sequences. The second was built with BIONJ by using the $(\delta_{ij}^{\alpha^*})$ matrix, where α^* was the value computed by our method. Both inferred topologies were compared with the true topology T. We then obtained the two topological distances $RF(T, \mathcal{T}^a)$ and $RF(T, \mathcal{T}^{\alpha^*})$ and computed the average (denoted as $\overline{RF}$) of these distances over 4,000 data sets with a and η being fixed. For each of the experimental conditions, we also computed the relative error decrease induced by the use of α^* instead of the (unknown) true value a. This corresponds to the ratio $[\overline{RF}(T, \mathcal{T}^{\alpha^*}) - \overline{RF}(T, \mathcal{T}^a)]/\overline{RF}(T, \mathcal{T}^a)$, which is negative when α^* performs better than a. Finally, a sign test was used to check the statistical significance of our findings.

The results are displayed in table 2. With 300-bp sequences, the three topologies inferred using α^* present less errors than those inferred using a, whatever the values of a and η. The best results occur when VRAS is strong ($a = 0.1$) and when the molecular clock holds. In this case, the relative decrease in topological error is close to 30%, which is highly significant and corresponds to much better inferred topologies.

For 1,000-bp sequences, the results are similar. However, the relative decrease in topological error is lower than before for seven of the nine experimental conditions. When $a = 2.0$ and $\eta = 2.0$, the topological accuracy is better using a than α^*, but the difference is not statistically significant. On the other hand, we still obtain an error decrease of about 30% in some cases (e.g., MC and $a = 0.1$). For longer sequences, the performances of α^* and a should become close, simply because (δ_{ij}^a) tends to be tree-like; therefore, α^* becomes close to a.

In conclusion, our method is remarkably accurate because its results are better than those that would be obtained if the real value of the gamma shape parameter was known. Its relative topological accuracy increases when VRAS is strong and when the deviation from molecular clock is slight or null.

Table 2.
Topological Accuracy of *a* Versus α*

	Length		$\overline{RF}$ (RED)		
			η = 0.5	η = 2.0	MC
a = 0.1 . . .	300 bp	[BIONJ + *a*]	0.499	0.389	0.332
		[BIONJ + α*]	0.438 (−12.2%)*	0.298 (−23.2%)*	0.238 (−28.3%)*
	1,000 bp	[BIONJ + *a*]	0.367	0.226	0.165
		[BIONJ + α*]	0.351 (−4.4%)*	0.190 (−15.9%)*	0.116 (−29.8%)*
a = 0.7 . . .	300 bp	[BIONJ + *a*]	0.367	0.223	0.165
		[BIONJ + α*]	0.352 (−4.1%)*	0.204 (−8.4%)*	0.146 (−11.5%)*
	1,000 bp	[BIONJ + *a*]	0.257	0.116	0.075
		[BIONJ + α*]	0.254 (−0.9%)*	0.114 (−1.8%)*	0.066 (−11.4%)*
a = 2.0 . . .	300 bp	[BIONJ + *a*]	0.334	0.190	0.134
		[BIONJ + α*]	0.328 (−2.0%)*	0.184 (−2.8%)*	0.130 (−3.0%)*
	1,000 bp	[BIONJ + *a*]	0.227	0.096	0.061
		[BIONJ + α*]	0.226 (−0.5%)*	0.097 (+0.9%)	0.058 (−4.0%)*

NOTE.—η, parameter measuring the deviation from molecular clock; MC, the molecular clock holds in the true tree; *a*, true value of the gamma shape parameter; [BIONJ + *a*], the trees are inferred with BIONJ from distances corrected with *a*; [BIONJ + α*], the trees are inferred with BIONJ from distances corrected with α*; $\overline{RF}$, mean values of RF (*T*, $\mathcal{T}^a$) and RF (*T*, $\mathcal{T}^{\alpha^*}$) computed from 4,000 20-taxon data sets; *RED*, relative error decrease between $\overline{RF}$ (*T*, $\mathcal{T}^a$) and $\overline{RF}$ (*T*, $\mathcal{T}^{\alpha^*}$); the statistical significance of each value is checked with the sign test: * → $P \leq 0.05$.

$\mathcal{T}^{\alpha^*}$ versus $\mathcal{T}^{ML}$

The results of our approach are now compared with that of ML. We used DNAML from the PHYLIP package (Felsenstein 1989) to build the ML trees. VRAS was modeled by a four category discretized gamma distribution using the true value *a* of the gamma shape parameter. In the same way, the TS/TV ratio was set to its real value, i.e., 2.0. Under such conditions, ML likely performs better than if *a* and TS/TV were unknown and had to be estimated from the sequences.

The values of η and *a* were identical to the previous ones, the sequences were 300 bp long and each data set contained 20 taxa. We computed the mean Robinson and Foulds distance between the true tree and the ML tree, $RF(T, \mathcal{T}^{ML})$, and the relative deviation $[RF(T, \mathcal{T}^{\alpha^*}) - RF(T, \mathcal{T}^{ML})]/RF(T, \mathcal{T}^{ML})$ assessed the difference of performance between our method and ML.

The results are displayed in table 3. When VRAS is strong (*a* = 0.1), the tree topology inference is better using BIONJ with α* than ML with *a*. For example, when the molecular clock holds, the relative decrease in topological error is about 26% with our method. When

a = 0.7 and *a* = 2.0, this property does not hold anymore. For example, ML trees are better than ours by about 12%–15%, when *a* = 2.0, which corresponds to a low VRAS. However, it must be underlined that ML trees are likely less accurate in real cases where *a* and the TS/TV ratio are unknown.

As phylogenetic inference methods are sensitive to the number of taxa analyzed, we have done supplementary simulations with 10-taxon trees. The results are similar to those obtained with 20-taxon trees, that is, the tree topology inference is better using BIONJ with α* than ML with *a* when VRAS is strong (*a* = 0.1), irrespective of the deviation from the molecular clock. The mean relative error decrease averaged over the three values of η is then close to 18%, in favor of our method (17% with 20-taxon trees). On the other hand, the topological accuracy is better with ML than with our method for *a* = 0.7 and *a* = 2 with a mean relative error decrease close to 14% in favor of ML trees (8% with 20-taxon trees).

Simulations with more than 20-taxon trees have not been carried out as it takes more than 1 week to run the

Table 3.
Comparison of our Approach and Maximum Likelihood

		$\overline{RF}$ (RED)		
		η = 0.5	η = 2.0	MC
a = 0.1 . . .	[ML + *a*]	0.471	0.365	0.325
	[BIONJ + α*]	0.433 (−7.92%)*	0.303 (−16.78%)*	0.238 (−26.74%)*
a = 0.7 . . .	[ML + *a*]	0.317	0.184	0.139
	[BIONJ + α*]	0.344 (+8.54%)*	0.192 (+4.38%)	0.141 (+1.84%)
a = 2	[ML + *a*]	0.311	0.171	0.118
	[BIONJ + α*]	0.337 (+8.45%)*	0.197 (+14.76%)*	0.132 (+11.63%)*

NOTE.—[ML + *a*], trees are inferred by maximum likelihood using the true value *a* of the gamma shape parameter; [BIONJ + α*], trees are inferred with BIONJ from distances corrected with α*; $\overline{RF}$, averages of RF (*T*, $\mathcal{T}^{ML}$) and RF (*T*, $\mathcal{T}^{\alpha^*}$); *RED*, relative error decrease between RF (*T*, $\mathcal{T}^{ML}$) and RF (*T*, $\mathcal{T}^{\alpha^*}$); the statistical significance of these values are checked with sign tests (* → $P < 0.05$). Results are based on 300 simulated 20-taxon data sets for each combination of η and *a*.

Table 4.
Computing Times Required by our Method and by DNAML

	[BIONJ + α*]	[ML + a]
$n = 100$	38.6 s	>3 days
$n = 50$	5.7 s	≃6 h
$n = 20$	0.8 s	≃15 min

NOTE.—n, number of taxa. The given values represent the time needed to infer one phylogeny with n taxa. These experiments have been performed with a PentiumIII, 750 Mhz computer.

tests with 20-taxon trees. Most of this computational time amount is caused by the building of ML trees. We have done supplementary simulations to compare more precisely the computational time required by both methods. We measured the time needed on a PentiumIII, 750 MHz computer by both methods to infer 20-, 50-, or 100-taxon trees from data sets being generated as described previously. Results are given in table 4. Our method is clearly more efficient than ML. For example, with 50 taxa, our method requires ≃6 s, whereas ML requires ≃6 h. This clearly precludes to bootstrap the data in the case of ML, whereas this task is easily achieved when using our method. Moreover, 3 days of computation are needed by ML with 100-taxon trees, which make its use rather unrealistic, whereas our method only requires ≃40 s.

We did not use fastDNAml (Olsen et al. 1994) (which is faster than DNAML) because it does not have the ability to handle the gamma distribution. However, refined implementations of DNAML, for example based on ideas from fastDNAml, would significantly reduce the computing times given here (although remaining much slower than our distance-based method).

Application to Maoricicadas Sequences

To illustrate our approach, we analyzed 25 orthologous sequences of the Maoricicada species (Buckley, Simon, and Chambers 2001). These sequences are 1,520 bp long and contain two mitochondrial regions which have been concatenated. The first is the COI gene, the second is the region from the tRNA^{Asp}, A8 and A6 genes. This data set was previously collected and analyzed by Buckley, Simon, and Chambers (2001) and Buckley et al. (2001). These authors used and compared different models of substitution and rate heterogeneity. All the variants of the Jukes and Cantor (1969), Kimura (1980), and Hasegawa, Kishino, and Yano (1985) models were rejected against the variants of the general-time reversible (GTR) model (Yang 1994). The rate heterogeneity model with best fit was obtained when partitioning the characters into first, second, and third codon positions and all tRNA^{Asp} sites and then estimating the gamma shape parameter separately for each of the four categories (Γ_4 model). The ML estimate of α was equal to 0.168 when considering all sites together. Hence, Maoricicada sequences seem to follow a more sophisticated pattern of evolution than simple models, such as Jukes and Cantor's or Kimura's, and VRAS is relatively strong in these sequences. Moreover, the ML tree that

is inferred presents a moderate deviation from molecular clock (figure 6 in Buckley et al. 2001).

Our method was used in the same way as previously described (i.e., K80 model and $0.1 < \alpha < 5,000$). We obtained for α* a value of 5,000 ($\simeq\infty$) which implies that the fit of the estimated distances to a tree distance is optimal when VRAS is not taken into account. The phylogeny inferred with BIONJ, given the ($\delta_{ij}^{\alpha*}$) matrix is shown in figure 5. The topology of this tree is similar to the one inferred with ML using the GTR + Γ_4 model, but three differences appear. The first difference concerns the position of the two *M. cassiope* species. These two sequences and both of *M. tenuis* constitute a monophyletic clade in the tree of Buckley et al. (2001). However, this clade is not well supported by the data, so the position of *M. cassiope* in our tree is also a plausible one (T. Buckley, personal communication). In the same manner, the position of *M. phaeoptera* differs in the two trees, but neither of these two positions is well supported. The third difference is more interesting and concerns the monophyly of the three *M. campbelli* sequences. This monophyly is retrieved in our tree but not by the tree of Buckley et al. (2001), despite it being very likely for several biological reasons (T. Buckley, personal communication). Note that this monophyletic clade is not recovered by BIONJ when using K80-distances and $\alpha = 0.168$, the ML value of a. Even if the bootstrap proportion corresponding to this clade is not very high (0.478, against 0.384 for Buckley et al.'s clade), it is worth noting that this biologically likely fact is retrieved, despite an apparently low amount of information in the data.

In summary, the Maoricicada tree inferred using our method is close to the ML tree but also proposes an original and biologically relevant group of taxa. It should be noted that the sequences present a strong VRAS and a low deviation from *MC,* which likely explains our good results (see the previous comparison between our method and ML).

Conclusions

This paper contains two main parts. In the first part, we show that the best value of α (α^{opt}) for tree inference from evolutionary distances is not equal to its true value (a). The lower the deviation from the molecular clock, the larger α^{opt} is relative to a and the more the optimal distances underestimate the true distances. This finding corroborates the observations from many authors (e.g., Saitou and Nei 1987; Sourdis and Krimbas 1987; Tajima and Takezaki 1994), established under many different experimental conditions without VRAS, where uncorrected/corrected for multiple substitutions distances were compared.

Given these observations, we propose a method to approximate the optimal value of α. We use a criterion that measures the reliability of the inferred tree, and our approximation (α*) corresponds to the value which optimizes this criterion. Simulation results demonstrate the topological accuracy of our method because performance is better using α* than using the (unknown) true
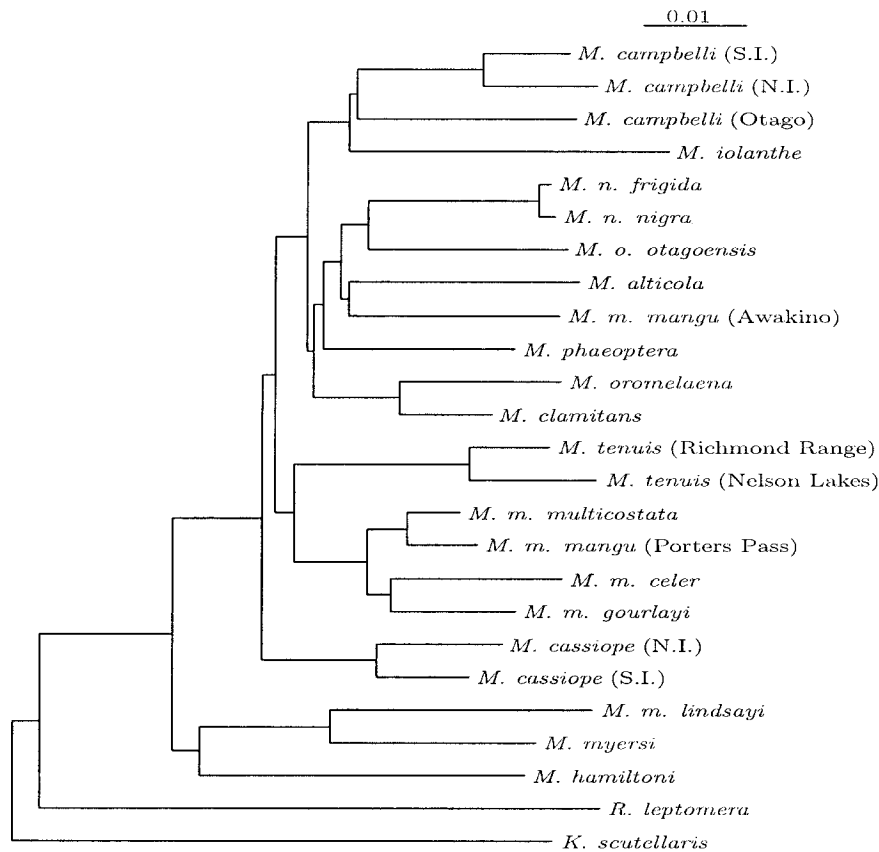
FIG. 5.—Phylogeny of Maoricicada species. This tree has been built with BIONJ from distances estimated with the K80 model and our $\alpha^*(\simeq\infty)$value.

value a. In numerous realistic experimental conditions, we obtain a relative decrease in topological error of about 30%. The comparison with the ML approach leads to unexpected results. Indeed, when VRAS is strong, our method seems to be more efficient than ML. This result is of importance because the always increasing amount of biological data confirms that VRAS is widespread and often very strong, notably in the first and second codon positions (Buckley et al. 2001). Moreover, our analysis of the Maoricicada sequences shows that correcting the distances by α^* yields a plausible topology with biologically likely clades which are not retrieved by ML and more sophisticated models.

As pointed out before, different authors have already described the improvement of topology inference induced by underestimating evolutionary distances when the molecular clock holds. However, no fully convincing explanation of this phenomenon has been given so far. A line of approach could be to extend some of the ideas presented by Rzhetsky and Sitnikova (1996).

In this study we compared various criteria to estimate α^*, and we selected the criterion that best performed in simulations. However, other criteria and other tree building algorithms could be combined to achieve better performance. Moreover, the approach presented here could likely be used to estimate other parameters involved in sequence evolution models.

Acknowledgments

LITERATURE CITED

BANDELT, H.-J., and A. DRESS. 1992. Split decomposition: a new and useful approach to phylogenetic analysis of distance data. Mol. Biol. Evol. **1**:242–252.

BUCKLEY, T. R., C. SIMON, and G. K. CHAMBERS. 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. **50**:67–86.

BUCKLEY, T. R., C. SIMON, H. SHIMODAIRA, and G. K. CHAMBERS. 2001. Evaluating hypotheses on the origin and evolution of the New Zealand Alpine Cicadas (Maoricicada) using multiple-comparison tests of tree topology. Mol. Biol. Evol. **18**:223–234.

BUNEMAN, P. 1971. The recovery of trees from measures of dissimilarity. Pp. 387–395 *in* F. R. HODSON, D. G. KENDALL, and P. TAUTA, eds. Mathematics in archeological and historical sciences. University Press, Edinburgh.

EIGEN, M., and R. WINKLER-OSWATITSCH. 1981. Transfer-RNA: the early adaptor. Die Naturwissenschaften **68**:217–228.

FELSENSTEIN, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17**:368–376.

———. 1989. PHYLIP (phylogeny inference package). Version 3.2. Cladistics **5**:164–166.

GASCUEL, O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. **14**:685–695.

GUÉNOCHE, A., and H. GARRETA. 2000. Can we have confidence in a tree representation? Pp. 45–56 *in* O. GASCUEL and M.-F. SAGOT, eds. Computational biology, LNCS 2066. Springer, Berlin.

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.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* H. N. MUNRO, ed. Mammalian protein metabolism, Vol. III, Chap. 24. Academic Press, New York.

KIMURA, M. 1980. A simple method for estimating evolutionary rates base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KUHNER, M., and J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

OLSEN, G. J., H. MATSUDA, R. HAGSTROM, and R. OVERBEEK. 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10**:41–48.

RAMBAUT, A., and N. GRASSLY. 1997. Seq-Gen: an application for the Monte-Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. **13**:235–238.

ROBINSON, D. F., and L. R. FOULDS. 1981. Comparison of phylogenetic trees. Math. Biosci. **53**:131–147.

RZHETSKY, A., S. KUMAR, and M. NEI. 1995. Four-cluster analysis: a simple method to test phylogenetic hypotheses. Mol. Biol. Evol. **12**:163–167.

RZHETSKY, A., and T. SITNIKOVA. 1996. When is it safe to use an oversimplified substitution model in tree-making? Mol. Biol. Evol. **13**:1255–1265.

SAITOU, N., and M. NEI. 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4**:406–425.

SCHÖNIGER, M., and A. VON HAESLER. 1993. A simple method to improve the reliability of tree reconstruction. Mol. Biol. Evol. **10**:471–483.

SOURDIS, J., and C. KRIMBAS. 1987. Accuracy of phylogenetic trees estimated from DNA sequence data. Mol. Biol. Evol. **4**:159–166.

SOURDIS, J., and M. NEI. 1988. Relative efficiencies of the maximum parsimony and distance-based methods in obtaining the correct phylogenetic tree. Mol. Biol. Evol. **5**:298–311.

STEEL, M., and D. PENNY. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Mol. Biol. Evol. **17**:839–850.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1995. Among-site variation and phylogenetic analysis of 12s rRNA in sigmontine rodents. Mol. Biol. Evol. **12**:988–1001.

TAJIMA, F., and N. TAKEZAKI. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11**:278–286.

TAKAHASHI, K., and M. NEI. 2000. Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol. Biol. Evol. **17**:1251–1258.

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

VACH, V. 1992. The Jukes-Cantor transformation and additivity of estimated genetic distances. Pp. 141–150 *in* M. SHADER, eds. Analysing and modeling data and knowledge. Springer, Berlin.

YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol. Biol. Evol. **10**:1396–1401.

———. 1994. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **10**:105–111.

———. 1996. Among-site rate variation and its impact on phylogenetic analyses. TREE **11**:367–372.

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 S. KUMAR. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. Mol. Biol. Evol. **13**:650–659.

ZARETSKII, K. 1965. Construction d'un arbre sur la base d'un ensemble de distances entre ses feuilles. USpekHi Math. Nauk. **20**:90–92 [in Russian].

ZHARKIKH, A., and W.-H. LI. 1993. Inconsistency of the maximum parsimony method: the case of five taxa with a molecular clock. Syst. Biol. **42**:113–125.