Testing for Differences in Rates-Across-Sites Distributions in Phylogenetic Subtrees

Edward Susko, Yuji Inagaki,† Chris Field,* Michael E. Holder,‡ and Andrew J. Roger†*

*Department of Mathematics and Statistics, Dalhousie University; †Program in Evolutionary Biology, Department of Biochemistry and Molecular Biology, Canadian Institute for Advanced Research, Dalhousie University; and ‡Xeotron Corporation, Houston

It has long been recognized that the rates of molecular evolution vary amongst sites in proteins. The usual model for rate heterogeneity assumes independent rate variation according to a rate distribution. In such models the rate at a site, although random, is assumed fixed throughout the evolutionary tree. Recent work by several groups has suggested that rates at sites often vary across subtrees of the larger tree as well as across sites. This phenomenon is not captured by most phylogenetic models but instead is more similar to the covarion model of Fitch and coworkers. In this article we present methods that can be useful in detecting whether different rates occur in two different subtrees of the larger tree and where these differences occur. Parametric bootstrapping and orthogonal regression methodologies are used to test for rate differences and to make statements about the general differences in the rates at sites. Confidence intervals based on the conditional distributions of rates at sites are then used to detect where the rate differences occur. Such methods will be helpful in studying the phylogenetic, structural, and functional bases of changes in evolutionary rates at sites, a phenomenon that has important consequences for deep phylogenetic inference.

Introduction

Recent work has suggested that the rates of molecular evolution at sites in an alignment often vary across subtrees of the larger evolutionary tree as well as across sites (Miyamoto and Fitch 1995; Lockhart et al. 1998, 2000; Lopez, Forterre, and Phillippe 1999; Galtier 2001; Gaucher, Miyamato, and Benner 2001; Gu 2001; Penny et al. 2001). This phenomenon is often viewed in the context of the covarion (concomitantly variable codons) model of Fitch and coworkers (Fitch and Markowitz 1970; Miyamoto and Fitch 1995), whereby the sites in a protein that were thought to be able to change (and perhaps the rates at which they change) were not constant over an evolutionary tree. Covarion-like evolution in molecules has three important consequences. First, most commonly used phylogenetic models assumed that the rates-across-sites sites process is fixed over the tree. Violation of this assumption by covarion processes can possibly lead to a longer persistence of phylogenetic signal over time under some conditions (Penny et al. 2001). Second, under other conditions, ignoring covarion-like evolution can lead to a special form of the long-branch attraction tree reconstruction artifact (Lockhart et al. 1998). For instance, rooting of the tree of life on the bacterial branch with ancient gene duplicates has been suggested to represent an artifact of this sort (Lopez, Forterre, and Phillippe 1999). Third, large changes in the evolutionary rates at sites may be related to major shifts in the structure, function, or interactions of protein or RNA molecules. Therefore, elucidating where rate changes have occurred in phylogenetic trees and in molecules may provide important insights into changes in the properties of the molecules during evolution along with their causes (Gaucher, Miyamato, and Benner 2001; Gu 2001; Knudsen and Miyamoto 2001). It is of interest in such cases to determine whether such changes actually are present or can be explained by sampling variation, what the general tendencies of change are, and where changes have occurred. We present three methodologies for detecting such changes. Two of these methods, regression tests with the rate estimates and a parametric bootstrap of rate distances, can be used to detect whether there are significant differences between the rates for the two subtrees. Confidence interval construction methods that can be used to detect the location of such changes are presented. Because the confidence interval methodology assumes a bivariate rates-across-sites model it differs from the likelihood methods for the detection of the rate difference of Knudsen and Miyamato (2001) which assumes a rates-across-sites model in estimating the tree but does not utilize this information in testing for rate differences.

The first type of methodology for detecting rate differences uses distances between rates estimated separately for the two subtrees of interest. Because there is a wide variety of ways in which rate variation can occur (for instance, only small rates might vary between the two subtrees or only large rates might vary) a number of different distances are considered. A parametric bootstrap is used with these distances to determine what types of distances are expected under the null hypothesis that there are no rate differences across the two trees.

An alternative approach for detecting whether rate differences exist comes through a regression analysis of the rate estimates. This is fairly easy to implement and can be used to complement the parametric bootstrap methodology. Under the null hypothesis that the rates are the same at each of the sites, a plot of the rates for one of the trees against the rates at the corresponding

Abbreviations: EF-1α, elongation factor 1α; aEF-1α, archaebacterial EF-1α; eEF-1α, eukaryotic EF-1α; eRF3, eukaryotic release factor 3.

Key words: covarion, rates-across-sites, Markov models, maximum likelihood, molecular evolution, phylogenetics.

Address for correspondence and reprints: Edward Susko, Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia B3H 3J5, Canada. E-mail: susko@mathstat.dal.ca.

Mol. Biol. Evol. 19(9):1514–1523. 2002
 ISSN: 0737-4038

site for the other tree should look like scatter around the $x = y$ line. We use orthogonal regression methodology with the rate estimates for the two trees to test whether the mean relationship between the rates satisfies this hypothesis. The estimated coefficients from the regression methods allow for statements about the general differences in the rates.

Given that rate differences exist between subtrees, it is of interest to determine at which sites in the sequences these changes occur. Confidence intervals for the rates at a site in the two subtrees based on the conditional distribution of the rates, given the data at the site, provide one way of detecting where changes that take random variation into account occur. Although confidence intervals for the rate at a site for a tree are of interest themselves, in the present context they are useful to construct confidence intervals for the differences in rates. Sites with confidence intervals for differences that do not contain 0 are likely to correspond to rate differences in the two subtrees.

Confidence intervals can be estimated separately for subtrees or based on a bivariate distribution for the rates in the separate subtrees. The bivariate model considered here is similar to Gu's (2001) model of functional divergence. Under the Gu (2001) model, the bivariate distribution of the rates in subtrees is a mixture of a distribution in which the rates are assigned independently to the two subtrees and a distribution in which the same rates are assigned to both subtrees. Here, in contrast, the bivariate distribution for rates is not restricted to any particular form. In addition, the methodology presented here provides confidence intervals for the differences in the rates at sites.

Models for Rate Variation

The usual model for rate variation treats rates at different sites as random variables that are independently distributed from a common rate distribution (cf. Yang, 1994; Felsenstein and Churchill 1996). The substitution model at a site is a continuous time Markov chain model with rate matrix Q. In our applications we use the empirically derived rate matrix of Dayhoff (Dayhoff, Schwartz, and Orcutt 1979) as implemented in PHYLIP 3.5c (Felsenstein 1993). For a site with rate r, along a branch of length t, the probability of substituting amino acid i with j is calculated as

$$M_{ij}(t) = \exp(Qrt).$$

For a given tree T, the probability of $f(x|r;T)$ can then be calculated by combining these probabilities according to the postorder tree traversal algorithm of Felsenstein (1981). Under this model, one can show that the expected number of substitutions along a branch of length t for a site with rate r is rt; thus a site with a rate of 4 is expected to have experienced four times as many substitutions as a site with a rate of 1. The total expected number of substitutions throughout a tree at a site with rate r is then r times the sum of the branch lengths for that tree. The marginal probability of data x for a rate distribution that assigns probabilities $\zeta_1, \ldots, \zeta_k$ to rates $r_1, \ldots, r_k$ is then

$$f(x; T) = \sum_{j=1}^{k} \zeta_j f(x \mid r_j; T).$$

To ensure that the branch lengths can be interpreted as the average number of substitutions per site, the expected rate coming from the rate distribution is constrained to be 1. In the independence model, the likelihood for a tree is the product of the $f(x;T)$ over all sites. The maximum likelihood methodology chooses the tree that gives the largest likelihood as an estimate of the evolutionary tree.

Given the estimated tree for a set of data, rate estimates are based on the conditional distribution of rates, given the data at a site. This distribution can be obtained through Bayes' formula:

$$p(r_j \mid x; T) = \frac{f(x \mid r_j; T)\zeta_j}{\left(\sum_j f(x \mid r_j; T)\zeta_j\right)}.$$

The most common rate estimate is the conditional mode: the rate giving the largest conditional probability $P(r_j|x;T)$. Alternative estimates can be constructed from the conditional distribution, however. For instance, the conditional mean

$$\sum_j r_j p(r_j \mid x; T) \quad (1)$$

is another reasonable estimate of the rate for a site with data x.

In cases where differences in rates in subtrees are of interest, separate rate estimates can be obtained by treating the data in the subtrees separately:

1. Use the data for subtree 1 to obtain a tree estimate T_1 for the taxa in that subtree, and similarly obtain a tree estimate T_2 for the taxa in subtree 2.
2. Use the separate conditional distributions $P(r_j|x;T_1)$, $P(r_j|x;T_2)$ and the separate data at the sites to obtain rate estimates r_{i1} and r_{i2} at site i.

Parametric Bootstrap Methodology for Detecting Rate Differences

The first method that we present for detecting whether rate differences exist between two subtrees compares the distance between rates estimated separately for the two subtrees with a parametric bootstrap distribution of the distances under the null hypothesis of no rate change. To obtain an overall measure of the change in rates in subtrees we developed three global distance measures, whereby real data could be compared with Monte-Carlo simulated data under the null hypothesis of no rate change.

The first rate-distance measure is simply the absolute value of the conditional mode rate for position i in the alignment from the first subtree subtracted from the conditional mode rate for the corresponding position in the

second subtree. The sum of these distances, summed over all sites, yields the global arsum distance measure

$$\text{arsum} = \sum_i |r_{i1} - r_{i2}|. \quad (2)$$

However, because of the uneven separation between rates implied by the standard form of the discrete gamma model (especially for $\alpha < 1$ where the distribution has a large, but low mass, tail), the arsum measure will be most strongly influenced by changes between the largest rate category in one subtree and all other rate categories in the other subtree. To deal with this problem we used a second rate-distance measure based on the absolute value of the logarithm of site-by-site rate ratios summed over all sites:

$$\text{alrsum} = \sum_i |\log(r_{i1}/r_{i2})|. \quad (3)$$

This measure intrinsically upweights large rate ratios which occur when the rate changes from the lowest rate category to the other rate categories (especially the highest) and thus emphasizes changes from the lowest rate category to all others. To achieve a slightly more balanced measure between these two extremes, a third global distance measure that scales the difference between two rates to their overall magnitude was calculated:

$$\text{abrsum} = \sum_i |r_{i1} - r_{i2}|/(r_{i1} + r_{i2}). \quad (4)$$

In addition to these three distance measures, the same measures were calculated without the absolute values (designated rsum, lrsum, and brsum, respectively). In principle, changes in rates at sites can be equally distributed across both subtrees (a homogeneous shift) or unequally distributed (nonhomogeneous shift). In the latter case, the rates could be systematically higher in one subtree versus the second subtree. The nonabsolute value measure described here can distinguish between these alternatives. If no systematic shift in rates has occurred, then the summed differences are expected to cancel out and fall within the simulated null distribution of the same measure. By contrast, a systematic shift will yield an imbalance of positive or negative distances that, through summation, will be evident as a large positive or negative value falling outside of the simulated null distribution.

The parametric bootstrapping tests were carried out as follows. For each pair of subtrees, the full data set alignment of the two subdata sets together is initially considered. From this alignment, the α parameter, α_{total}, is estimated by maximum likelihood on a neighbor-joining topology using TREE-PUZZLE version 4.02 (Strimmer and von Haeseler 1996), and maximum likelihood distances for all pairs of taxa are estimated using the PAM + Γ model (the PAM 001 model was used and the gamma distribution is approximated by an eight-category discrete rate distribution, where the eight categories have equal probability and the rate for each category is set to its mean). From this distance matrix, an optimal topology is estimated using the Fitch-Margoliash weighted least-squares method with 10 random additions with global rearrangements implemented in the PHYLIP 3.5c (Felsenstein 1993) program FITCH. Maximum likelihood branch lengths for topology are estimated with TREE-PUZZLE with the PAM + Γ model using α_{total}. From this topology, *N* data sets of equal size are simulated with the PAM + Γ model using the program PSeq-Gen (Rambaut and Grassly 1997). Each of the full data set alignments (observed or simulated) is then split into the two smaller data sets corresponding to the separate subtrees (data set–1 and data set–2) for separate analysis. For each of the smaller data sets, a maximum likelihood distance matrix is estimated with the PAM + Γ model using α_{total}. A topology is estimated from this matrix using the neighbor-joining algorithm (implemented in the PHYLIP program NEIGHBOR), and this is used as a user-defined tree for a second round of TREE-PUZZLE analysis. This analysis uses the PAM + Γ model, using α_{total} to estimate maximum likelihood branchlengths for the NJ topology and the conditional mode rates for the data set. The conditional mode rates for each site are then compared between data set–1 and data set–2, using the global distance measures described earlier. A shell-script program, COVAR, was written to automate these comparisons.

Testing for Rate Differences Using Regression Methods

The regression methods presented in this section provide a straightforward way of testing whether rate changes have occurred at all in the two subtrees of interest.

Let r_{i1}, r_{i2} denote the rate estimates at site *i* for the two trees, and let μ_{i1}, μ_{i2} denote the true unobserved rates at the site. We can then define errors in estimation through

$$r_{i1} = \mu_{i1} + \epsilon_{i1} \qquad r_{i2} = \mu_{i2} + \epsilon_{i2}.$$

If the rate estimates are reasonable estimates of the rates at sites, we expect that the mass of the distribution of the ϵ_{ij}s will be near 0. Without additional information that indicates otherwise, we assume that the error in estimation of rates will on average be similar for both of the trees so that the ϵ_{ij}s are independent 0 mean random variables having a common distribution.

Under the null hypothesis of interest H_0: $\mu_{i1} = \mu_{i2}$ for all *i*. If this null hypothesis is true, then

$$r_{i2} = \mu_{i1} + \epsilon_{i1} = \mu_{i2} + \epsilon_{i2} = r_{i1} - \epsilon_{i1} + \epsilon_{i2}.$$

Let $\epsilon_i = \epsilon_{i2} - \epsilon_{i1}$. Then

$$r_{i2} = r_{i1} + \epsilon_i.$$

Thus, if we fit the regression model

$$r_{i2} = \beta_0 + \beta_1 r_{i1} + \epsilon_i. \quad (5)$$

The estimates of the slope and intercept in the model should be close to 1 and 0.

In most regression models of the form (5), least squares estimates of (β_0, β_1) are used. These are not, however, appropriate in the current setting. This is because both of the variables in the regression equation

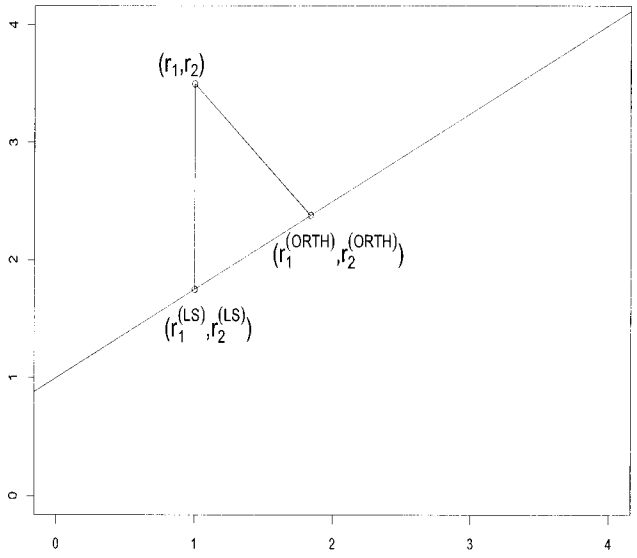

FIG. 1.—The points used for orthogonal and least squares regression.

are subject to error. As a consequence, least squares estimates of the slope are biased in the direction of 0. Because of the common error process for the two trees (the errors ϵ_{i1} and ϵ_{i2} are assumed to have the same error distribution), an adjustment is available through orthogonal regression (cf. §12.3 Casella and Berger 1990, pp. 583–594).

Orthogonal regression is best explained by contrast with least squares regression. In both forms of regression, parameters are chosen to make the sum of the distances between the observed rates (r_{i1}, r_{i2}) and points on the regression line as small as possible. The regression methods differ in the way they measure the distance from the observed (r_{i1}, r_{i2}) to the line. As illustrated in figure 1, least squares estimation uses the vertical distance, whereas orthogonal regression uses perpendicular distance.

Standard calculations give the orthogonal regression estimates of the slope and intercept of the regression line as

$$\hat{\beta} = \frac{-(S_{11} - S_{22}) + \sqrt{(S_{11} - S_{22})^2 + 4S_{12}^2}}{2S_{12}} \quad \text{and} \quad (6)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (7)$$

Here $S_{11} = \Sigma_i (r_{i1} - \bar{r}_1)^2$, $S_{22} = \Sigma_i (r_{i2} - \bar{r}_2)^2$, and $S_{12} = \Sigma_i (r_{i1} - \bar{r}_1)(r_{i2} - \bar{r}_2)$.

If the distribution of the error terms ϵ_{i1} and ϵ_{i2} is normal, orthogonal regression turns out to be equivalent to maximum likelihood estimation of the regression parameters. Even if this is not the case, it provides a reasonable methodology for estimating the parameters in the model. A standard error for the estimate $\hat{\beta}_1$ is available as

$$\mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{(1 + \hat{\beta}_1^2)^2 (S_{11}S_{22} - S_{12}^2)}{(S_{11} - S_{22})^2 + 4S_{12}^2}}. \quad (8)$$

The hypothesis that $(\beta_0, \beta_1) = (0,1)$ in (eq. 5) is equivalent to the hypothesis that $\beta_1 = 1$ and that the mean rates, averaged over all sites, for the two subtrees is the same. A paired *t*-test can be used to test the hypothesis that the mean rates are equal. A test of the hypothesis that $\beta_1 = 1$ can be constructed from the orthogonal regression estimate and standard error for $\hat{\beta}_1$. An approximate *P* value for the test is

$$1 - \Phi(|\hat{\beta}_1|/\mathrm{se}(\hat{\beta}_1))$$

where $\Phi(z)$ is the standard normal cumulative distribution function.

Confidence Bounds for Rates

The regression and parametric bootstrap methodologies allow one to draw inferences about whether a rate difference between two subtrees is present. In the event that a rate difference is detected through the methodology, a follow-up analysis of the location of the rate differences is required. In this section we present methods that can be used to construct confidence intervals for rates at individual sites in a given subtree.

Confidence bounds for the rate at a site for a given subtree is most naturally based on the conditional distribution for rates at a site:

$$p(r_j | x; T) = \frac{f(x | r_j; T)\zeta_j}{\left(\sum_j f(x | r_j; T)\zeta_j\right)}$$

A $(1 - \alpha) \times 100\%$ confidence limits for the rate at site with data x is given by $[l(x), u(x)]$, where $l(x)$ and $u(x)$ satisfy

$$P(\text{rate} < l(x)) \sum_{r_j < l(x)} p(r_j | x; T) \leq \alpha/2 \quad \text{and} \quad (9)$$

$$P(\text{rate} > u(x)) \sum_{r_j > u(x)} p(r_j | x; T) \leq \alpha/2. \quad (10)$$

The conditional probability that the true rate is contained within the interval $[l(x), u(x)]$ is at least $(1 - \alpha)$ for any $l(x)$ and $u(x)$ satisfying equations (9) and (10). To make the widths of the confidence intervals as small as possible, $l(x)$ should be chosen as the largest number satisfying equation (9) and $u(x)$ should be chosen as the smallest value satisfying equation (10). The confidence interval $[l(x), u(x)]$ must be considered an approximate confidence interval because its stated confidence properties hold when the rate distribution, branch lengths, and all other parameters involved in the calculation of $P(r_j|x; T)$ are known. In practice, estimates of these quantities are used.

The confidence bounds for the rates at individual sites in individual subtrees are of interest in their own right but also provide a means for detecting the locations of rate differences. If, for a given site, the confidence bounds for rates in two subtrees do not overlap, a rate difference is suggested. More generally, the confidence intervals for the rates at sites can be used to construct confidence intervals for the rate difference between the two sites. For a given site, suppose that the $(1 - \alpha/2) \times 100\%$ confidence bounds for the rates in two subtrees

are $[l_1, u_1]$ and $[l_2, u_2]$. Then a $(1 - \alpha) \times 100\%$ confidence interval for the rate difference is the set of all difference $r_1 - r_2$ with r_1 in $[l_1, u_1]$ and r_2 in $[l_2, u_2]$; this set is simply $[l_1 - u_2, u_1 - l_2]$.

Confidence Bounds for Rate Differences

Using individual confidence intervals to detect rate differences turns out to be a crude approach to the problem in cases where rates at sites for two subtrees appear to be positively correlated. In this case, estimation of a bivariate rate distribution can be used to construct tighter confidence limits for the rate difference at a site.

The usual model for rate variation treats rates at different sites as random variables that are independently distributed from a common rate distribution (cf. Yang 1994; Felsenstein and Churchill 1996). For all practical purposes, rate distributions are discrete; they assign probabilities $\zeta_1, \ldots, \zeta_k$ to a set of rates $r_1, \ldots, r_k$. For a given tree, and a set of rates $r_1, \ldots, r_k$, the likelihood for the rate distribution parameters $\zeta_1, \ldots, \zeta_k$ is obtained by taking the product of the probabilities of data at the individual sites. An estimate of the rate distribution can be obtained by maximizing the likelihood over all $\zeta_1, \ldots, \zeta_k$ corresponding to rate distributions that have a mean rate of 1 (cf. Susko et al. 2001).

The extension of rate distributions for a given tree to separate subtrees would be a bivariate distribution that assigns probabilities to pairs of rates (r_1, r_2) at a site for the two subtrees. If $r_1, \ldots, r_k$ is the set of possible rates for tree T_1 and $s_1, \ldots, s_k$ is the set of possible rates for tree T_2, a bivariate distribution would assign some probability ζ_{ij} to the pair of rates (r_i, s_j). To ensure that the branch lengths can be interpreted as the expected number of substitutions, the rate distribution should be chosen so that the expected rates, separately, at the two subtrees are 1. Specifically,

$$\sum_i \zeta_{ij} r_i = 1 \qquad \sum_j \zeta_{ij} s_j = 1.$$

A bivariate rate distribution allows a large range of variation in rate difference behavior. At the one extreme, if $\zeta_{ij} > 0$ only if $i = j$, then the rates for the two subtrees are always the same. Another possibility would be that $\zeta_{ij} = \tau_i \nu j$, where $\tau_1, \ldots, \tau_k$ are the probabilities of the rates $r_1, \ldots, r_k$ for subtree T_1 and $\nu_1, \ldots, \nu_k$ are the probabilities of the rates $s_1, \ldots, s_k$ for subtree T_2. Here rate assignment would be independent in the two subtrees.

When estimating rate distributions or constructing confidence intervals for a rate in a given subtree, it suffices to consider only the data in that subtree. In contrast, calculations of the probability of data at a site in a bivariate model require that the data in both subtrees be considered jointly. For a given site and given rates r and s for subtrees T_1 and T_2, let $f(x,y|r,s;T_1,T_2)$ be the joint probability of the data x in subtree T_1 and data y in subtree T_2. The probability of data x and y at a site is then obtained as

$$\sum_{ij} \zeta_{ij} f(x, y \,|\, r_i, s_j; T_1, T_2).$$

The likelihood for the rate distribution parameters ζ_{ij} is obtained by taking the product of the probabilities of data at the individual sites. An estimate of the rate distribution can be obtained by maximizing the likelihood over all ζ_{ij} corresponding to bivariate rate distributions that have a mean rate of 1 for each of the subtrees.

Given probabilities ζ_{ij} for the pairs of rates (r_i, s_j), similarly as in the case of a single subtree, the conditional distribution of the pair (r_i, s_j) is calculated as

$$p(r_i, s_j \,|\, x, y; T_1, T_2) = \frac{f(x \,|\, r_i, s_j; T_1, T_2)\zeta_{ij}}{\left(\sum_{ij} f(x \,|\, r_i, s_j; T_1, T_2)\zeta_{ij}\right)}.$$

The conditional probability of a difference d in rates is then calculated by summing over all pairs of rates giving this difference:

$$p(d \,|\, x, y; T_1, T_2) = \sum_{(r_i,s_i)|r_i - s_j = d} p(r_i, s_j \,|\, x, y; T_1, T_2).$$

Given the conditional distribution of the difference, a $(1 - \alpha) \times 100\%$ confidence interval can be calculated as for individual rates as $[l(x), u(x)]$, where $l(x)$ and $u(x)$ satisfy that

$$\sum_{d<l(x)} p(d \,|\, x, y; T_1, T_2) \le \alpha/2 \quad \text{and}$$

$$\sum_{d>u(x)} p(d \,|\, x; T) \le \alpha/2$$

The joint probability of the data $f(x,y|r,s; T_1,T_2)$ is needed in all of the aforementioned calculations. Calculation of this quantity requires an additional branch length for the branch connecting the two subtrees, branches in the two subtrees where the additional branch will join the two subtrees, and a position along the additional branch where the rate change occurs. To avoid the additional computation implied by these additional parameters we use an independence model approximation in practice, replacing $f(x,y|r,s; T_1,T_2)$ by the product of the separate probabilities of data in the two subtrees: $f(x|r;T_1)f(y|s;T_2)$. If the branch connecting the two subtrees is relatively long, this should provide a good approximation.

Results

As an illustrative example we consider amino acid data for taxa coming from four subtrees, one for the eukaryotic elongation factor 1α (eukaryotic EF-1α, the eEF-1α data set), one for archaebacterial EF-1α (the aEF-1α data set), one for Hsp70 subfamily B Suppressor 1 (HBS1), and one for the eukaryotic release factor 3 (eRF3). Although eukaryotic and archaebacterial EF-1α have the same primary biological function of delivering aminoacyl-tRNAs to the A site of the ribosome during translation elongation, they are known to have different auxiliary protein-protein interactions in the two domains of life (Inagaki and Doolittle 2000). HBS1 and eRF3 are eukaryote-specific EF-1α paralogs that are believed to have arisen from EF-1α via gene duplications before the divergence of extant eukaryotes (Inagaki and Doolittle 2000). eRF3 is one of several proteins that function in the translation termination process, whereas the function

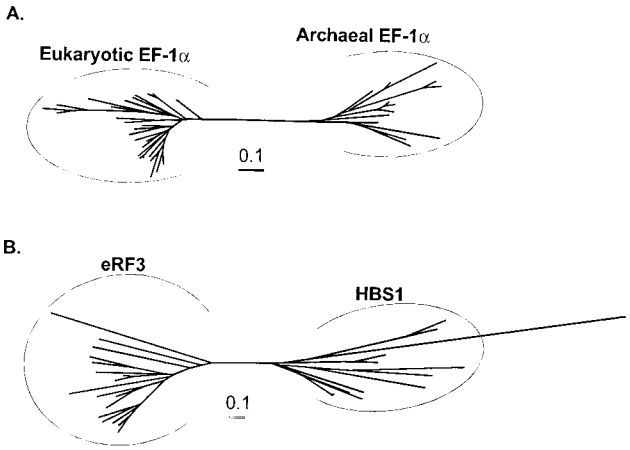

FIG. 2.—*A,* Unrooted tree based on the data set including 13 archaebacterial and 27 eukaryotic EF-1α (aEF-1α + eEF-1α data set). The tree was obtained using (PAM + Γ) TREE-PUZZLE v.4.0.2 with a maximum likelihood distance matrix employing the PAM amino acid substitution matrix, incorporating among-site rate variation (discrete gamma distribution approximated with eight categories). Subsequently the tree was reconstructed based on the maximum likelihood distance matrix using the Fitch-Margolish method with global rearrangements implemented in PHYLIP v.3.6. Branch lengths for the optimal distance tree were reestimated by the maximum likelihood method with the PAM + Γ model in TREE-PUZZLE. *B,* Unrooted tree based on the data set including 17 eRF3 and 13 HBS1 (HBS1 + eRF3 data set). The details are as described earlier. The α parameters, 0.63 and 1.03, were estimated from the aEF-1α + eEF-1α and HBS1 + eRF3 data sets, respectively.

of HBS1 is currently unknown but thought to be different from EF-1α or eRF3 (Inagaki and Doolittle 2000). It is of particular interest to compare the rates at sites for the aEF-1α and eEF-1α data sets as well as to compare the HBS1 and eRF3 data sets. The aEF-1α data set had 13 taxa, the eEF-1α data set had 28 taxa, the HBS1 data set had 13 taxa, and the eRF3 data set had 17 taxa. New sequences were added manually to the previous alignment (Inagaki and Doolittle 2000). The removal of ambiguously aligned positions left 269 sites shared between all four protein families.

Parametric Bootstrap Methodology

Maximum likelihood distance-Fitch Margoliash trees were obtained under a discrete gamma model with 16 rate categories for two combined data sets, including the aEF-1α and eEF-1α data set (aEF-1α + eEF-1α data set), and the HBS1 and eRF3 data sets (HBS1 + eRF3 data set). The estimated trees are given in figure 2. The α parameters for the gamma models were estimated from the aEF-1α + eEF-1α and HBS1 + eRF3 data sets in TREE-PUZZLE as 0.63 and 1.03, respectively. These two α parameters were used as estimates of α_{total} for the parametric bootstrap analyses.

The parametric bootstrap analyses detected a significant difference between eukaryotic and archaebacterial EF-1α data sets. The *P* values for the test statistics arsum, alrsum, and abrsum were estimated from the parametric bootstrap distribution as 0.001, 0.000, and 0.000, respectively (fig. 3*A–C*). Curiously, the nonabsolute value rate distances between the two EF-1α data sets, lrsum and brsum, fell significantly on one side of the bootstrap distribution (>99th percentile, data not shown). Their positions indicate that, at least for medium and low rate differences (emphasized by the lrsum and brsum values), sites in archaebacterial EF-1α evolved at a systematically higher rate than those in the eukaryotic homologs.

By contrast, the distances observed between HBS1 and eRF3 gave *P* values that ranged from being marginally insignificant to marginally significant at the 0.01 level of significance for the three test statistics. The *P* values for the test statistics arsum, alrsum, and abrsum were estimated from the parametric bootstrap distribution as 0.003, 0.01, and 0.018, respectively (fig. 4*A–C*). The observed differences from nonabsolute value distance measures were not significantly different from those under the null distribution (data not shown). These results suggest that the tempo and mode of the HBS1 evolution may be slightly different from those of eRF3. However, these two show less overall rate difference than that observed for the aEF-1α and eEF-1α comparison.

Regression Methodology Results

A scatter plot of the log-transform estimated rates for the aEF-1α and eEF-1α data sets is given in figure 5*A.* As an initial test of whether rate differences exist between the aEF-1α and eEF-1α data sets or the HBS1 and eRF3 data sets, we used the orthogonal regression methodology. The rate estimates used were the conditional mean rate estimates given by equation (1). The usual rate estimates are conditional mode estimates which, with a discrete gamma model, take on only a few values. Regression methods are more appropriate for rate estimates that can take on a large set of values, which is the case for conditional mean estimates. Because many of the rate estimates are small, log transformations were taken to increase the range of data and avoid difficulties with outlying values. The orthogonal regression methodology was then applied to the log-transformed conditional mean rate estimates. The results are given in table 1 and suggest very strongly that there are rate differences between the aEF-1α and eEF-1α subtrees but that there is little evidence of rate differences between the HBS1 and eRF3 subtrees.

Confidence Bounds for Rate Differences

Because the results of the orthogonal regression tests suggest that there are rate differences for the aEF-1α and eEF-1α subtrees but little evidence of rate differences for the HBS1 and eRF3 subtrees, further analysis was restricted to the aEF-1α and eEF-1α subtrees. Confidence intervals were calculated for the rates at each of the sites for each of the subtrees. A total of 13 out of the 269 sites (5%) have 95% confidence intervals that do not overlap (fig. 6).

Use of nonoverlapping confidence intervals to identify sites where significant rate differences exist is valid but can be expected not to detect some of the sites where changes have occurred because of its failure to

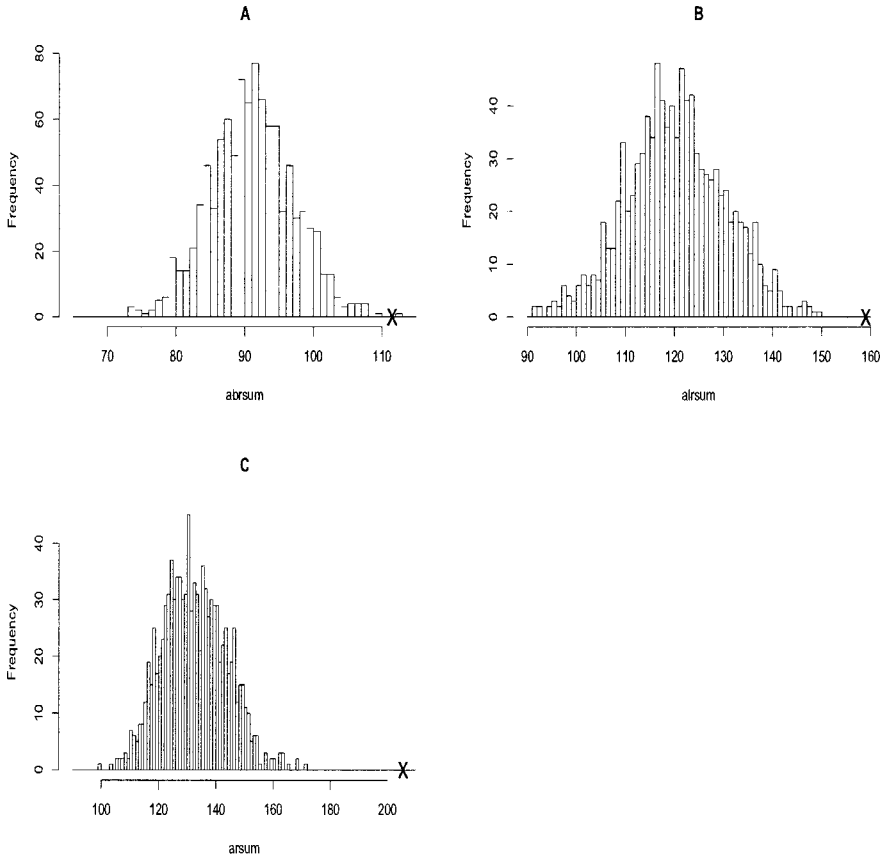

FIG. 3.—Histograms of the bootstrap values of the three test statistics abrsum (*A*), alrsum (*B*), and arsum (*C*) for the comparison of the aEF-1α and eEF-1α data sets. The locations of the observed test statistics are indicated with an *X*.

incorporate the clear correlations between rates at sites in the two subtrees evident from the orthogonal regression analysis. Modeling the bivariate distribution of the rates allows one to adjust for the correlation. For the aEF-1α and eEF-1α data sets we estimated a bivariate rate distribution, using likelihood methods and then constructed confidence intervals for the rate differences. A total of 60 of the 269 rate differences (22%) have 95% confidence intervals that do not contain 0 (fig. 7). All of the sites that were deemed to have significant differences because the individual confidence intervals were nonoverlapping are still significant, but a large number of additional sites are also significant.

Concluding Remarks

We have presented a number of different methods in this article. The first set of methods, orthogonal regression and parametric bootstrap tests, are useful for detecting whether rate changes have occurred. The orthogonal regression methodology uses the rate estimates alone as a means of determining whether there are significant differences in the means for the subtrees. The parametric bootstrap repeats the process of tree, branch length, and rate estimation under the null hypothesis. Because it incorporates additional sources of variability, it can be expected to be more sensitive in detecting significant rate differences. This is borne out in the comparison of rates between the HBS1 and eRF3 data sets. The orthogonal regression methodology failed to find significant differences where the parametric bootstrap results ranged from marginally significant to marginally insignificant. Both methodologies have value; the parametric bootstrap can be expected to be more sensitive in detecting departures from the null, whereas the orthogonal regression methods requires much less computation in large problems. In general, all of the methods will detect rate differences more easily if there are large numbers of sites and taxa; however, power investigations in simplified settings (data not shown) suggest that larger rate difference might be detected with as few as five taxa per subtree.

Failure to reject the hypothesis of significant differences suggests that further analysis is not necessary and effectively controls the type I error rate. Given the rejection of the orthogonal regression or parametric bootstrap tests, confidence intervals can be constructed for rates to locate sites that are likely to have significant rate differences. Confidence bounds for rate differences, calculated after estimating a bivariate rate distribution for the two subtrees of interest, can be expected to produce tighter bounds for the rate differences. To allow for easier computation, an independence assumption was made about the data in the two subtrees. When the branch connecting the two subtrees is relatively long, one can expect the resulting approximations to be reasonable. We expect that sites with low rates in both sub-

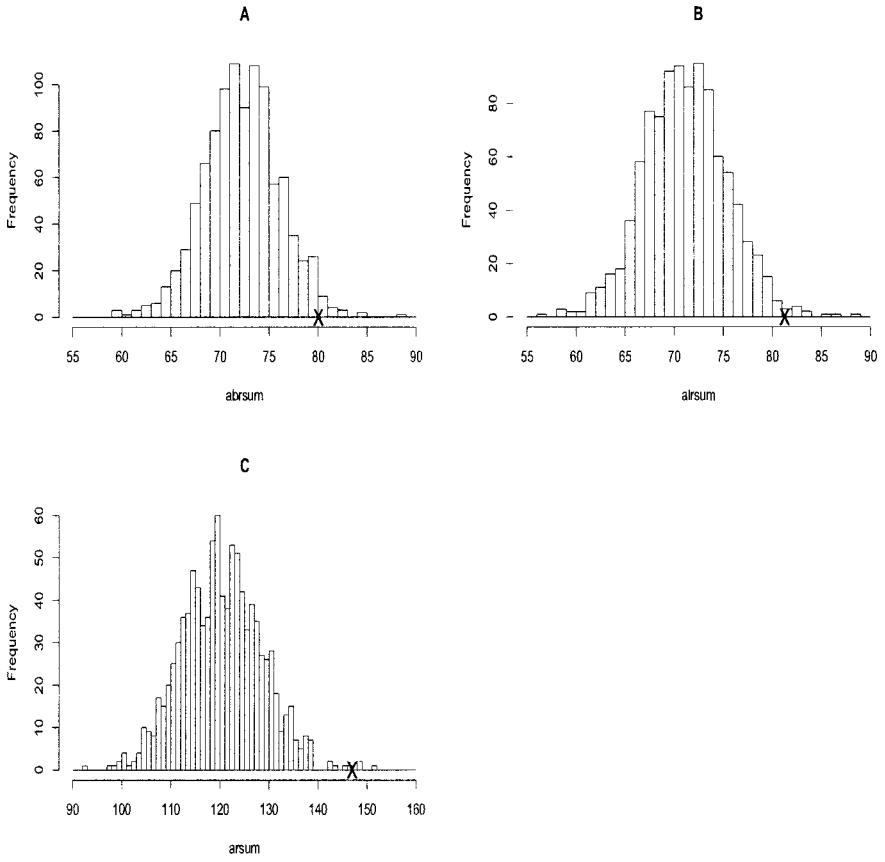

FIG. 4.—Histograms of the bootstrap values of the three test statistics abrsum (*A*), alrsum (*B*), and arsum (*C*) for the comparison of the HBS1 and eRF3 data sets. The locations of the observed test statistics are indicated with an *X*.

trees will be the ones most significantly affected by a loosening of the independence assumption. In this case, the branch length between the two subtrees for the site is effectively shortened because of the low rates so that the data in the two subtrees become more dependent. Determining how or whether this would affect the confidence interval for the rate difference will require further study.

Our methods have been applied to compare rates at sites for data sets with subtrees from the HBS1 and eRF3 data sets. Although significant rate differences were suggested in the comparison of the rates from the aEF-1α and eEF-1α data sets, the rate differences in the HBS1 versus eRF3 comparisons were marginally significant and insignificant using the parametric bootstrapping test and regression test, respectively. Further analysis is needed to determine what the effects of failing to adjust for rate differences in subtrees might be for phylogenetic estimation.

The likelihood ratio statistic test of Knudsen and Miyamato (2001) treats the rate difference at a site as a fixed parameter for estimation. It uses only the data at the site and assumes that large sample likelihood theory is applicable. In contrast, the bivariate confidence intervals treat the rate difference as a random variable. The information about the range of likely rate differences contained within the data at a site is obtained by conditioning on the data at a site, but information from other sites is obtained by using the bivariate rate distribution in the calculation of intervals. The bivariate model considered here can be viewed as an extension of the rates-across-sites model where rates are allowed to vary in subtrees of the larger tree. The model is similar to Gu's (2001) model for functional divergence but places less restrictions on the form of the bivariate distribution. The extension is also similar to the covarion models of Tuffley and Steel (1998), Galtier (2001), and Penny et al. (2001), which also allow rate variation between sub-

Table 1
The Paired *t*-Test and Orthogonal Regression Results for the Differences in the Rate Estimates for the aEF-1α and eEF-1α Data Sets as well as the HBS1 and eRF3 Data Sets

	aEF-1α–eEF-1α Data	HBS1–eRF3 Data
$\hat{\beta}_1$	2.1485	1.079
95% Confidence interval β_1	1.9874–2.3097	0.9864–1.1717
P value: $\beta_1 = 1$	0.000	0.2402
P value: paired *t*-test ($\beta_0 = 0$)	0.000	0.5281

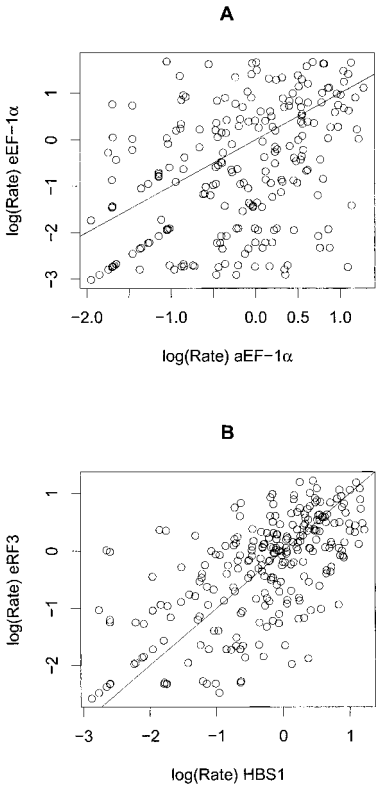

FIG. 5.—Scatter plots of the log-transformed conditional mean rates at sites with the $x = y$ line indicated.

trees. These covarion models assume a stationary process of rate variation at a site throughout the tree. This differs from the bivariate model considered here in at least two respects. First, the process of rate variation is constant throughout the tree so that the rate distribution or rates-across-sites model for one subtree should be the same as for any other subtree. The bivariate model allows different rate distributions in different subtrees. Second, in the Tuffley and Steel model, rates are allowed to vary within any branch of a subtree. In contrast, the bivariate model assumes a single rate at a site for a given subtree. Nevertheless, the bivariate models considered here can be useful in detecting covarion-type rate variation similar to that of the Tuffley and Steel model. In the Tuffley and Steel model one can think of an average rate at a site for a subtree, where the average is taken over branches of the subtree. Assuming that rates at a site vary randomly throughout a tree, by chance there should be differences in the average rates at a site in any two subtrees of the larger tree. Because the average rates will differ, with sufficient data, the orthogonal regression or parametric bootstrap tests presented here will reject the null hypothesis of a single rate distribution model.

The bivariate model can be extended to allow rate variation in smaller and smaller subtrees of the tree of interest. In the most general case, we could partition the tree into m subtrees, $T_1, \ldots, T_m$. In a multivariate model for rate variation at a site, a set of rates $r_1, \ldots, r_m$ for the subtrees would be drawn from a multivariate distribution that assigns some probability $\pi(i_1, \ldots, i_m)$ to

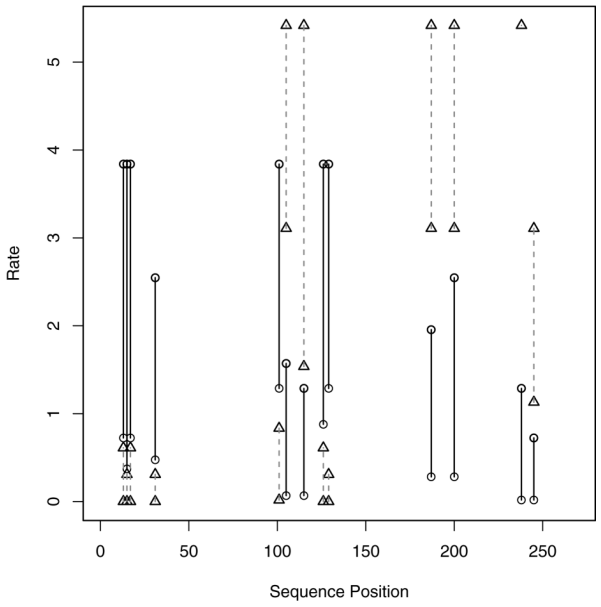

FIG. 6.—The nonoverlapping 95% confidence intervals for the rates at sites for the aEF-1α subtree (solid line) and the eEF-1α subtree (dashed line).

every possible set $r_{i1}, \ldots, ri_m$ of m rates that could arise for the subtrees.

Acknowledgments

E.S. and C.F. were supported by the Natural Sciences and Engineering Research Council of Canada. A.J.R. thanks the Canadian Institute for Advanced Research Program in Evolutionary Biology for fellowship support. A.J.R. and Y.I. were supported by the NSERC Operating Grant 227085-00 and NSERC Genomics Grant 228263-99. This collaboration is part of a Genome Atlantic/Genome Canada Large-Scale Project.

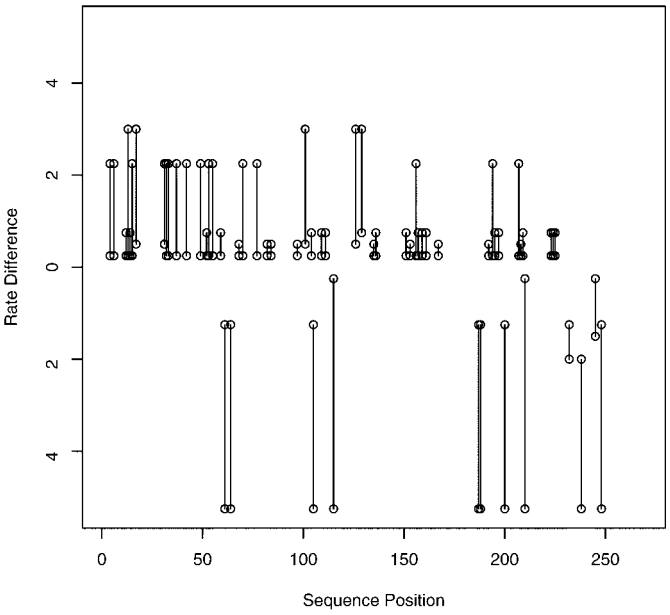

FIG. 7.—Confidence intervals for rate differences (aEF-1α and eEF-1α) that did not contain 0.

LITERATURE CITED

CASELLA, G., and R. L. BERGER. 1990. Statistical inference. Brooks/Cole, Pacific Grove, Calif.

DAYHOFF, M. O., R. M. SCHWARTZ, and B. C. ORCUTT. 1979. A model of evolutionary change in proteins. Pp. 345–352 *in* M. O. DAYHOFF, ed. Atlas of protein sequence and structure, Vol. 5(Suppl. 3). National Biomedical Research Foundation, Silver Spring, Md.

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

———. 1993. PHYLIP (phylogeny inference package). Version 3.5c. Distributed by the author, Department of Genetics, University of Washington, Seattle.

FELSENSTEIN, J., and G. A. CHURCHILL. 1996. A hidden Markov model approach to variation among sites in rate of evolution. Mol. Biol. Evol. **13**:93–104.

FITCH, W. F., and E. MARKOWITZ. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem. Genet. **4**:579–593.

GALTIER, N. 2001. Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol. Biol. Evol. **18**:866–873.

GAUCHER, E. A., M. M. MIYAMATO, and S. A. BENNER. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc. Natl. Acad. Sci. USA **98**:548–552.

GU, X. 2001. Maximum likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18**: 453–464.

INAGAKI, Y., and W. F. DOOLITTLE. 2000. Evolution of the eukaryotic translation termination system: origins of release factors. Mol. Biol. Evol. **17**:882–889.

KNUDSEN, B., and M. M. MIYAMATO. 2001. A likelihood ratio test of evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. USA **98**:14512–14517.

LOCKHART, P. J., D. H. HUSON, U. MAIRER, M. J. FRAUNHOLZ, Y. VAN DE PEER, A. C. BARBROOK, C. J. HOWE, and M. A. STEEL. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol. Biol. Evol. **15**:1183–1188.

———. 2000. How molecules evolve in eubacteria. Mol. Biol. Evol. **17**:835–838.

LOPEZ, P. FORTERRE, P., and H. PHILLIPE. 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. **49**:496–508.

MIYAMOTO, M. M., and W. M. FITCH. 1995. Testing the covarion hypothesis of molecular evolution. Mol. Biol. Evol. **12**:503–513.

PENNY, D., B. J. MCCOMISH, M. A. CHARLESTON, and M. D. HENDY. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J. Mol. Evol. **53**:711–723.

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

STRIMMER, K., and A. VON HAESELER. 1996. Quartet puzzling: a quartet maximum likelihood method for reconstructing tree topologies. Mol. Biol. Evol. **13**:964–969.

SUSKO, E., C. FIELD, C. BLOUIN, and A. R. ROGER. 2002. Estimation of rate distributions in phylogenetic models. Preprint.

TUFFLEY, C., and M. STEEL. 1998. Modeling the covarion hypothesis of nucleotide substitution. Math. Biosci. **147**:63–91.

YANG, Z. 1994. Maximum likelihood phylogentic estimation from DNA sequences when substitution rates differ over sites: approximate methods. J. Mol. Evol. **39**:306–314.

NARUYA SAITOU, reviewing editor

Accepted May 2, 2002