Maximum-Likelihood Phylogenetic Analysis Under a Covarion-like Model

Nicolas Galtier

Centre National de la Recherche Scientifique UMR 5000—Génome, Populations, Interactions, Université Montpellier 2, Montpellier, France

Here, a model allowing covarion-like evolution of DNA sequences is introduced. In contrast to standard representation of the distribution of evolutionary rates, this model allows the site-specific rate to vary between lineages. This is achieved by adding as few as two parameters to the widely used among-site rate variation model, namely, (1) the proportion of sites undergoing rate changes and (2) the rate of rate change. This model is implemented in the likelihood framework, allowing parameter estimation, comparison of models, and tree reconstruction. An application to ribosomal RNA sequences suggests that covarions (i.e., site-specific rate changes) play an important role in the evolution of these molecules. Neglecting them results in a severe underestimate of the variance of rates across sites. It has, however, little influence on the estimation of ancestral G+C contents obtained from a nonhomogeneous model, or on the resulting inferences about the evolution of thermophyly. This theoretical effort should be useful for the study of protein adaptation, which presumably proceeds in a typical covarion-like manner.

Introduction

Markov models of DNA sequence evolution are widely used for reconstructing phylogenetic trees and studying the processes of molecular evolution from genomic sequence data. Considerable progress has been made since the precursor works of Jukes and Cantor (1969) and Kimura (1980): models have been built to account for unbalanced base composition (Hasegawa, Kishino, and Yano 1985; Tamura 1992), variable G+C content between sequences (Galtier and Gouy 1998), and synonymous versus nonsynonymous changes (Goldman and Yang 1994), to mention only a few. A significant advance occurred when Yang (1993, 1994) introduced a model allowing variable substitution rates across sites within the likelihood framework. Yang showed that adding a single parameter—namely, the shape parameter of an assumed Gamma distribution of rates—could increase the likelihood by many orders of magnitude. This is presumably because selective constraints vary across sites. Some sites in a protein or an RNA evolve more or less freely, whereas others can hardly be substituted without a significant drop in fitness. Accounting for this effect greatly improves our representation of molecular evolution.

This picture, however, is an instantaneous one. It is quite likely that the selective constraints applying to a particular site also vary in time and between lineages. As far as long periods of time are concerned, critical sites with respect to the function of a macromolecule may change, making the evolutionary rate of a particular site variable across the phylogeny. This notion was introduced as early as 30 years ago by Fitch and Markowitz (1970) and Fitch (1971) and was called the ''covarion'' process. The terminology comes from the idea that the rate of a site might be modified by a substitution arising at a distinct site of the molecule, with which it therefore covaries. Site-specific rate variation, however, can as well be caused by external factors such as environmental changes. In this paper, I refer to ''true covarions'' when two (several) sites of a sequence are undergoing nonindependent evolution, and I use the term ''site-specific rate variation'' (SSRV) or ''covarion-like evolution'' in the more general case.

There are at least two good reasons for being interested in modeling SSRV. First, it might improve phylogenetic reconstructions. Philippe and colleagues have shown that SSRV is common in genes widely used for the recovery of deep phylogenies and have and suggested that it induces tree-building biases (Germot and Philippe 1999; Lopez, Forterre, and Philippe 1999; Philippe et al. 2000). Second, modeling SSRV should help to obtain an understanding of the way natural selection applies at the molecular level. In particular, episodic adaptive evolution presumably proceeds with frequent changes of rates at various sites.

Little theoretical work was achieved after Fitch's early reports about SSRV until Tuffley and Steel (1998) proposed an explicit Markov model that resulted in a distance-based test for detecting SSRV effects (Lockhart et al. 1998, 2000). In this paper, I introduce a more general Markov model of site-specific rate variation and devise a maximum-likelihood implementation of this model. This method is applied to ribosomal RNA sequences to assess the amount of covarion-like evolution for this kind of data and to check the robustness of previously reported inferences about the early evolution of life.

Abbreviations: ASRV, among-site rate variation; SSRV, site-specific rate variation; USSRV, unequal site-specific rate variation.

Key words: covarion, Markov model, thermophyly, LUCA, positive selection.

Address for correspondence and reprints: Centre National de la Recherche Scientifique UMR 5000—Génome, Populations, Interactions, Université Montpellier 2, Place E. Bataillon, 34095 Montpellier cedex, France. E-mail: galtier@crit1.univ-montp2.fr.

Mol. Biol. Evol. 18(5):866–873. 2001

Methods

The Model

Yang (1994) proposed a discrete-Gamma model of among-site rate variation where the evolutionary rate of a particular site is one out of an arbitrary finite number g of possible relative rates ($r_1, r_2, \ldots, r_g$). These possible relative rates and their probabilities are obtained by discretizing a Gamma distribution with mean unity; one can ensure equal probabilities for the r_i's by using

appropriate cutting points, which was done by Yang and in this study. The probability distribution of rates across sites is therefore determined by a single parameter, namely, the shape parameter α of the Gamma distribution. Under this model, simulating the evolution of a site on a given tree using a given Markov process **M** of nucleotide substitution involves three steps: (1) randomly draw a rate from the discretized Gamma distribution, (2) randomly draw an ancestral nucleotide state at the root of the tree, and (3) make this state evolve along the branches of the tree according to process **M** and the rate that was drawn at step 1. A site therefore belongs to a ''category'' (following Felsenstein's terminology) which is fixed during the entire simulation process.

I now generalize Yang's model by allowing site-specific rates to vary in time. It is assumed that the rate of a particular site can switch from one category to another according to a continuous Markov process **R**. With rate ν, the current evolutionary rate moves to a new category, obtained by randomly drawing from the distribution of r_i's. Switches between distant categories of rates are therefore assumed to be as probable as short-range changes. The parameter ν could be called the ''rate variation rate.'' It determines the amount of site-specific rate variation. It is assumed constant over sites and in time. Note that the process **R** of rate change is continuous: switches can occur anywhere in the tree, not specifically at nodes.

Simulating the evolution of a site under the SSRV model now involves four steps: (1) randomly draw an ancestral rate at the root of the tree from the discretized Gamma distribution; (2) make this rate evolve along the branches of the tree according to process **R**, and record the rate assigned to each segment of the tree (i.e., between switching points); (3) randomly draw an ancestral nucleotide state at the root of the tree; and (4) make this state evolve along the branches of the tree according to process **M** and the local rate determined at step 2 (i.e., scaling the length of each segment according to its relative rate). Nucleotide process **M** is therefore compounded with (i.e., superimposed on) rate process **R**. **R** is time-reversible, making the compound process time-reversible if **M** is so. Note that sites are independent and identically distributed under this model: there is no ''true covariation'' here. The equal-rates (ER), among-site rate variation (ASRV), and site-specific rate variation (SSRV) models are graphically compared in figure 1. SSRV reduces to ASRV when $\nu = 0$ (no change of rate) and to ER when ν tends to infinity (permanent change of rate results in constant rate).

In the formulation above, the rate of a site can be reassigned its current category when a switch occurs. This has no biological relevance: process **R** is identical to a process where switches would occur at rate $\nu' = \nu \cdot (g - 1)/g$ without self-reassignment. The chosen parameterization simplifies some of the calculations below.

Maximum-Likelihood Implementation

This section aims at computing the probability of a particular DNA sequence data set under the SSRV model given a tree topology, a set of branch lengths $\{\lambda_i\}$, a time-reversible nucleotide Markov process (or substitution matrix) **M**, a Gamma shape parameter α, and an SSRV rate ν.

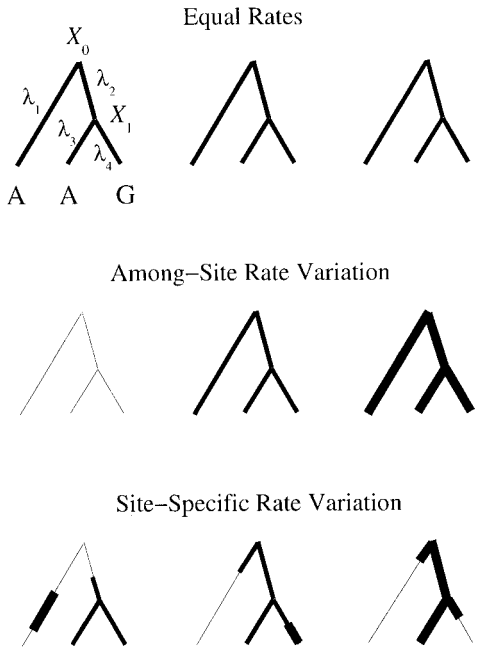

FIG. 1.—Distribution of rates across sites and lineages under three models of evolution. Each tree plot describes the distribution of rates across lineages for a particular site under the considered model. Three categories of rate are assumed, represented by different line thicknesses. Under the equal-rates (ER) model, all sites evolve at a constant, unique, moderate rate. Under the among-site rate variation (ASRV) model, each site has its own rate (low, moderate, or high), which is constant between lineages. Under the site-specific rate variation (SSRV) model, the rate of a site can switch between categories; a site has distinct rates in distinct lineages.

Felsenstein (1981) introduced likelihood calculation across trees under the ER model ($\alpha = \infty$, $\nu = 0$). Since independent evolution of sites is assumed, the probability of a data set is the product of the probabilities of observing each site. The probability of a particular site is computed by conditioning over all possible nucleotide states at internal nodes of the tree. For example, the probability of site (AAG) under ER in the three-species tree of figure 1 (top left) is

$$\Pr(\mathrm{AAG}) = \sum_{X_0} \sum_{X_1} \Pr(R = X_0) \times \Pr(X_0 \to \mathrm{A}/\lambda_1) \times \Pr(X_0 \to X_1/\lambda_2) \times \Pr(X_1 \to \mathrm{A}/\lambda_3) \times \Pr(X_1 \to \mathrm{G}/\lambda_4), \quad (1)$$

where R is the nucleotide state at the root node, X_0 and X_1 belong to {A, C, G, T}, and $\Pr(X \to Y/\lambda)$ is the probability of reaching state Y when evolving from state X along a branch of length λ according to process **M**. These transition probabilities are derived from the theory of stochastic processes (e.g., see Yang 1995). With time-reversible **M**, the likelihood does not depend on the location of the root as soon as the probabilities of nucleotide states at the root are taken from the stationary distribution of **M**—the so-called pulley principle (Fel-

senstein 1981). Equation (1) can be generalized to any tree topology using a recurrent formula. Let y be a site;

$$\Pr(y) = \sum_X \Pr(R = X) \times \Pr(y/R = X) \quad (2)$$

$$\Pr(y/N = X) = \sum_{X_1} \Pr(X \to X_1/\lambda_1) \times \Pr(y/N_1 = X_1) \times \sum_{X_2} \Pr(X \to X_2/\lambda_2) \times \Pr(y/N_2 = X_2), \quad (3)$$

where R is the nucleotide state at the root node, N is the state at any internal node, N_1 and N_2 are the states at the son nodes of N, and λ_1 and λ_2 are the lengths of branches leading from N to N_1 and from N to N_2, respectively. The summation for X_1 and X_2 is over {A, C, G, T}. Equations (2) and (3) show that the likelihood of a site under ER can be computed by a single pass on the tree, in time linear in the number of leaves and in the square of the number of states (four for nucleotide sequences).

I now extend this method to the above SSRV model. The probability of a site can be computed by conditioning on both states and rates at ancestral nodes. Equations (2) and (3) become

$$\Pr(y) = \sum_X \sum_r \Pr(R = X, r_R = r) \times \Pr(y/R = X, r_R = r) \quad (4)$$

$$\Pr(y/N = X, r_N = r) = \sum_{X_1} \sum_s \Pr(X \to X_1, r \to s/\lambda_1, \nu) \times \Pr(y/N_1 = X_1, r_{N_1} = s) \times \sum_{X_2} \sum_t \Pr(X \to X_2, r \to t/\lambda_2, \nu) \times \Pr(y/N_2 = X_2, r_{N_2} = t), \quad (5)$$

where r_R, r_N, r_{N_1}, and r_{N_2} are the rates at nodes root, N, N_1, and N_2, and the summations for s and t are over the range $\{r_1, \ldots, r_g\}$. Equations (4) and (5) show that computing the likelihood under the SSRV model with four states (nucleotides) has the same complexity as computing the likelihood under equal rates with $4{\cdot}g$ states. The time required is therefore g^2 times as great as that required under the ER model, and g times as great as that required under the ASRV model (Yang 1994).

Now comes the problem of calculating transition probabilities in equation (5), namely, the probability of evolving from rate r_i to rate r_j and from state X to state Y during length λ under processes **R** and **M**. This can be done following the standards of stochastic process theory. The combined **R** × **M** process can be viewed as a single Markov process **Q** in which states are defined as (X, r_i) pairs, where X is a nucleotide state and r_i is a rate. There are $4{\cdot}g$ possible states. Transitions between states occur at rate

$$\begin{aligned}
&\mathbf{Q}(X, r_i \to Y, r_i) = r_i \times \mathbf{M}(X \to Y) && \text{(nucleotide change)}\\
&\mathbf{Q}(X, r_i \to X, r_j) = \nu/g && \text{(rate changes)}\\
&\mathbf{Q}(X, r_i \to Y, r_j) = 0 && \text{(simultaneous changes neglected).}
\end{aligned} \quad (6)$$

The probability matrix **P** of final state (Y, r_j) given initial state (X, r_i) after evolution according to **Q** along branch length λ is given by taking the exponent of matrix $\mathbf{Q}{\cdot}\lambda$ (Yang 1995):

$$\mathbf{P} = \exp(\mathbf{Q}{\cdot}\lambda). \quad (7)$$

This approach allows exact calculation of the likelihood. It is not optimal, however, for technical purposes, because equation (7) requires numerical diagonalization of $4{\cdot}g \times 4{\cdot}g$ matrix **Q**, which can be time-consuming. This numerical step, moreover, precludes analytical calculation of the derivatives of the likelihood with respect to parameters of the model, quite useful for likelihood maximization. This is especially problematic when a complex model of nucleotide change is used (e.g., see below). I now derive an approximate calculation of the transition probabilities that does not require any matrix diagonalization. This is achieved by writing

$$\Pr(X \to Y, r_i \to r_j) = \Pr(r_i \to r_j/\lambda, \nu) \times \Pr(X \to Y/r_i, r_j, \lambda, \nu). \quad (8)$$

Equation (8) holds because rate changes do not depend on nucleotide states. The left-hand factor depends only on the rate process **R**. From stochastic process theory, it equals $(1 - \exp(-\nu{\cdot}\lambda))/g$ if $j \neq i$, and $\exp(-\nu{\cdot}\lambda) + (1 - \exp(-\nu{\cdot}\lambda))/g$ if $j = i$. The right-hand factor is the probability of evolving from nucleotide state X to nucleotide state Y along branch length λ given that the initial rate was r_i and the final rate is r_j. This can be approximated by

$$\Pr(X \to Y/r_i, r_j, \lambda, \nu) \approx \Pr(X \to Y/\lambda{\cdot}\bar{r}_{ij}(\lambda, \nu)), \quad (9)$$

where $\bar{r}_{ij}(\lambda, \nu)$ is the mean of the relative rate across its evolution along a branch of length λ at rate ν given the initial and final rates r_i and r_j. Equation (9) is an approximation because the nucleotide process is assumed to apply along the average net branch length, rather than to be integrated over the distribution of the net length (where the net length of a branch is the length obtained after every segment has been scaled according to its rate). The conditional average rate is

$$\bar{r}_{ij}(\lambda, \nu) = \frac{(a + b{\cdot}\lambda{\cdot}\nu){\cdot}e^{-\lambda{\cdot}\nu} + \lambda{\cdot}\nu - a}{\lambda{\cdot}\nu{\cdot}[1 + (c - 1){\cdot}e^{-\lambda{\cdot}\nu}]}, \quad (10)$$

where $a = 2{\cdot}(1 - r_i)$, $b = 1 + (g - 2){\cdot}r_i$, and $c = g$ if $j = i$, and where $a = 2 - r_i - r_j$, $b = 1 - r_i - r_j$, and $c = 0$ if $j \neq i$. The derivation of equation (10) is given in the appendix. The approximation was found to be quite good when applied to the data sets used in this study. The approximate transition probabilities can be differentiated analytically, allowing the use of efficient maximization algorithms.

Unequal SSRV Rates Among Sites

A constant rate of rate change ν is assumed in the above SSRV model, which might be unrealistic. It is likely that for many proteins or RNAs the rate of some sites remains more or less constant for long periods, while other sites switch more often. Some sites might remain critical for the function of the macromolecule and evolve at a slow rate in all of the branches of the tree as a consequence of strong purifying selection pressure. Some may escape any selective pressure and evolve at a (fast) neutral rate in the long run. Other sites, involved in episodic adaptive events, might recurrently switch between a slow and a fast rate, in agreement with the above-described SSRV process. I now generalize the SSRV model to account for this possibility. It is assumed that a proportion π of the sites evolve according to SSRV (with SSRV rate ν), while the remaining sites evolve according to ASRV (with SSRV rate 0). This general model is called unequal site-specific rate variation (USSRV).

The probability of site y under USSRV is given by

$$\Pr_{\text{USSRV}}(y) = \pi \cdot \Pr_{\text{SSRV}}(y) + (1 - \pi) \cdot \Pr_{\text{ASRV}}(y), \quad (11)$$

where $\Pr_{\text{SSRV}}(y)$ is given by equations (4) and (5) and $\Pr_{\text{ASRV}}(y)$ is obtained by setting $\nu = 0$ in equations (4) and (5) (or see Yang 1994). The USSRV model acknowledges $g + 1$ categories of sites: sites with constant rate r_1 (proportion $[1 - \pi]/g$), sites with constant rate r_2 (proportion $[1 - \pi]/g$), . . . , sites with constant rate r_g (proportion $[1 - \pi]/g$), and sites with variable rates (proportion π). The likelihood of a site under USSRV is computed by averaging likelihoods conditional on that site belonging to every category. USSRV reduces to SSRV when $\pi = 1$ and to ASRV when $\pi = 0$.

The approximate likelihood is maximized using the Newton-Raphson method. Then, the exact calculation is performed using the near-optimal parameter values sought in the previous step. Finally, the exact likelihood is maximized using the downhill-simplex method. These algorithms are available as part of the NHML package (ftp://pbil.univ-lyon1.fr/pub/mol_phylogeny).

Data Analysis

Ribosomal RNA (rRNA) sequences are widely used for reconstructing ancient evolutionary events. Recently, an improved model of nucleotide substitutions was applied to data from Eukaryotes, Bacteria, and Archaea for the purpose of estimating ancestral rRNA base compositions (Galtier, Tourasse, and Gouy 1999). This model, however, did not allow for site-specific rate variation. In this section, I investigate the importance of SSRV effects in rRNA evolution. The relevance of the above-mentioned study is checked in the light of the SSRV model. Ribosomal RNA data are also used to examine the influence of the number of taxa on SSRV estimation.

Two data sets, each including 40 species (10 Archaea, 17 Bacteria or chloroplasts, and 13 Eukaryotes), were used: small-subunit (SSU) rRNA (695 unambiguously aligned, complete sites) and large-subunit (LSU) rRNA (1,409 sites). Applying a nonhomogeneous, nonstationary model of nucleotide substitution (Galtier and Gouy 1998) to these data, Galtier, Tourasse, and Gouy (1999) estimated the G+C content of the most recent common ancestor (MRCA) to extant life forms. They found that the moderate estimated G+C content (56.1% $\pm$ 5% for SSU, 54.0% $\pm$ 2.5% for LSU) is not compatible with life at very high temperatures: high ribosomal G+C content is a necessary condition for survival of present-day species in hot environments (Galtier and Lobry 1997). This result therefore questions the hypothesis of a thermophilic common ancestor (e.g., see Woese 1987; Forterre 1996).

Galtier, Tourasse, and Gouy (1999) assumed a Gamma distribution of rates among sites. I conducted the analysis again under the more general SSRV and USSRV models, allowing covarion-like effects. This involved combining the above piece of theory with Galtier and Gouy's (1998) nonhomogeneous model of DNA evolution. The latter allows G+C content to vary in time and between lineages. The combined model accounts for unequal transition/transversion ratio, variable G+C content between lineages, variable rates among sites and, site-specific rate changes. The assumed Gamma distribution was discretized in $g = 4$ equally probable classes of rates. Table 1 displays the log likelihoods and the details of parameter estimates for four models of rate variation, namely, ER, ASRV, SSRV, and USSRV (assumed tree topology: fig. 1 of Galtier, Tourasse, and Gouy 1999).

This analysis suggests that site-specific changes of evolutionary rates (i.e., covarion-like evolution) is a major feature of rRNA evolution. Allowing site-specific rate changes resulted in a vast increase in log likelihood ($\ln L_{\text{SSRV}} - \ln L_{\text{ASRV}} = 105.1$ for SSU, 230.2 for LSU). Allowing unequal ν among sites (USSRV) also significantly improved the fit (23.6 more log likelihood units for SSU, 37.7 for LSU). The difference in log likelihood was highly significant according to likelihood ratio tests. The estimated proportion π of sites undergoing rate changes under USSRV was remarkably high. Although the increase of log likelihood was lower than than between ER and ASRV, accounting for site-specific changes of rates seemed to significantly improve our representation of rRNA evolution.

Allowing covarion-like evolution made a difference with respect to the estimation of parameters of the evolutionary model. Although the general shape of the tree was unchanged (data not shown), branches were slightly longer under (U)SSRV than under ASRV (total tree length was increased by roughly 10%). This suggests that some saturation might be overlooked when site-specific rate variation exists and is not taken into account. The transition/transversion ratio κ was also higher when estimated under (U)SSRV. Again, this is reminiscent of the previously reported (and confirmed here) underestimate of κ when among-site rate variation is not accounted for (e.g., Wakeley 1996). Underestimating κ is a consequence of overlooking multiple substitutions, since transitions saturate more quickly than transversions. The shape parameter α of the Gamma dis-

Table 1
Likelihood Analysis of Two 40-Species rRNA Data Sets Under Four Models of Site-Specific Rate Distribution

	SSU				LSU			
	ER	ARSV[a]	SSRV	USSRV	ER	ASRV[a]	SSRV	USSRV
No. of parameters . . .	158	159	160	161	158	159	160	161
Log likelihood	−10,380.4	−9,762.6	−9,657.5	−9,633.9	−21,488.3	−20,302.6	−20,072.4	−20,034.7
κ	2.70	3.22	3.57	3.86	2.52	2.81	3.07	3.18
α	—	0.603	0.279	0.113	—	0.650	0.247	0.120
ν	—	—	1.118	6.864	—	—	1.825	6.337
π	—	—	—	0.639	—	—	—	0.689
Total tree length.	3.257	4.448	4.804	5.208	3.211	3.825	4.142	4.334
GC (%).	59.8	56.1	55.7	55.4	55.7	53.8	53.0	53.6

NOTE.—SSU = small-subunit rRNA data set; LSU = large-subunit rRNA data set; ER = equal-rates model; ASRV = among-site rate variation model; SSRV = site-specific rate variation model; USSRV = unequal site-specific rate variation model; κ = transition/transversion ratio; α = Gamma shape; ν = SSRV rate; π = proportion of SSRV sites; GC = ancestral G+C-content.

[a] The small differences between these results and those of Galtier, Tourasse, and Gouy (1999) are a consequence of a difference in the number of classes of the discretized Gamma distribution (eight in the above reference, four in this analysis).

tribution is the most sensitive to covarion-like effects. It is dramatically decreased when site-specific rate variation is allowed—remember that a decrease in α means a higher variance of rates across categories. This is presumably because the mean evolutionary rate of a rate-changing site is not extreme: fast and slow periods average in the long run. These sites are ''seen'' as moderately fast when considered from the point of view of ASRV, while they are actually a mixture of slow and fast rates. The variance of the overall distribution of rates is therefore underestimated.

The use of a nonhomogeneous, nonstationary model of evolution allows one to estimate ancestral base compositions. The SSRV and USSRV estimates of the MRCA's rRNA G+C content are very close to (and even slightly lower than) the ASRV estimate. Galtier, Tourasse, and Gouy's (1999) result is therefore confirmed when site-specific rate variation is taken into account. Their claim of a nonhyperthermophilic common ancestor did not result from a biased analysis—as far as covarion-like effects are concerned.

LSU rRNA data were used to assess the sensitivity of SSRV detection to the number of analyzed sequences. Fifteen subsets of sequences including 20, 10, or 5 species (5 of each category) were randomly drawn from the total of 40 sequences, making sure that four domains (namely, Eukaryotes, Bacteria, Euryarchaea, and Crenarchaea) were represented in proportions roughly identical to those of the total data set. Each subset was analyzed using ASRV, SSRV, and USSRV. Results are shown in table 2 (the approximate likelihood calculation was used here). For all three models, the estimated variance of rate across sites was decreased (parameter α increases) when the number of species decreased, as previously reported (e.g., Tourasse and Gouy 1997). A similar effect was found for parameters ν and π. Twenty-species data sets were quite consistent with the total 40-species data set but 10- and 5-species data sets contained little information with respect to site-specific rate variation, as indicated by the small difference in log likelihood between models and the high variance of parameter estimates. With 10 species, a significant increase in log likelihood was found when moving from the ASRV to the SSRV model, but not when moving from SSRV to USSRV (excepting one data set out of five). With five species, no (U)SSRV effect was detected. This again underlines the importance of species sampling in molecular phylogeny and evolution studies. There is little hope of detecting any SSRV effect using fewer than 20 or 30 sequences.

Discussion

The SSRV and USSRV models introduced in this paper are more general than Fitch and Markowitz's (1970) original description of the so-called ''covarion''

Table 2
Influence of the Number of Analyzed Sequences on Site-Specific Rate Variation Detection and Parameter Estimates

No. of Species	ASRV	SSRV			USSRV			
	α	α	ν	Δln *L*	α	ν	π	Δln *L*
40	0.65	0.25	1.83	229.9	0.12	6.34	0.69	41.2
20	0.68[a]	0.29	1.70	72.4	0.14	6.81	0.68	12.7
	(0.66, 0.70)[a]	(0.25, 0.31)	(1.52, 1.99)	(64.3, 77.3)	(0.06, 0.25)	(3.99, 9.05)	(0.65, 0.70)	(4.2, 20.3)
10	0.75	0.39	1.47	11.9	0.29	4.70	0.70	2.0
	(0.64, 0.85)	(0.32, 0.46)	(1.22, 1.84)	(7.1, 15.3)	(0.05, 0.39)	(1.38, 10.0)	(0.61, 0.91)	(0.0, 6.6)
5	0.82	0.74	0.36	0.3	0.69	110.5	0.27	0.0
	(0.74, 0.98)	(0.43, 0.98)	(0.0, 1.52)	(0.0, 1.5)	(0.49, 0.96)	(0.0, 551.4)	(0.0, 0.97)	(0.0, 0.0)

NOTE.—ASRV = among-site rate variation model; SSRV = site-specific rate variation model; USSRV = unequal site-specific rate variation model; Δln *L* = increase in log likelihood obtained by adding parameters ν (SSRV) and π (USSRV).

[a] Mean and (minimal, maximal) estimates out of five data sets.

process, recently formalized by Tuffley and Steel (1998). In the original model, sites can be either ''on'' or ''off'': there are two classes of rates, one of which is rate 0. A common rate of switch between the two categories is assumed for all sites. The proposed SSRV model allows an arbitrary number of classes but involves the additional assumption of discrete-Gamma distribution (i.e., constraints on the relative rates of distinct categories). The advantages of (discrete-)Gamma versus discrete rate class models are discussed by Yang (1996) in the context of among-site rate variation. Discrete-Gamma models provide a good fit to many data sets at the cost of few, easily interpretable parameters. Yang's arguments presumably also apply to SSRV. The USSRV model is original in allowing a proportion of sites not to experience rate changes, relaxing a dubious assumption of Tuffley and Steel's (1998) and SSRV models. It is quite unlikely that every site of a molecule undergoes rate changes at a common rate.

A maximum-likelihood implementation was achieved by making use of the properties of the compound process of rate and nucleotide changes. A desirable property of the SSRV and USSRV models in the likelihood framework is their generalization of the widely used ER and ASRV models. (U)SSRV reduces to ASRV when $\nu = 0$ and/or $\pi = 0$, and to ER when α or ν tends to infinity. This means that the parameters are easily interpretable. They directly measure the importance of site-specific rate variation and can be compared between data sets. Furthermore, the nested relationship between the four models allows relevant comparisons of likelihoods and election of the most appropriate model through likelihood ratio tests.

The new models revealed a significant amount of site-specific rate variation when applied to ribosomal RNA data. This analysis suggested that neglecting SSRV when it exists has at least two important consequences. First, the variance of the distribution of rates among sites is underestimated (Gamma shape parameter overestimated). Second, correction of multiple substitutions is less efficient (total tree length and transition/transversion ratio underestimated). The two effects presumably result from a unique cause: highly switching sites have a moderate average rate in the long run. Said simply, these sites are considered moderately fast when analyzed under ASRV, so the number of multiple substitutions they experience during fast-rate episodes is underestimated. The ability of (U)SSRV models to detect some saturation that is hidden to ASRV suggests that these models might improve phylogenetic reconstructions. Lockhart et al. (1998) feel the same: they argue that the internal edge that separates plastid and cyanobacterial 16S rRNA sequences is mainly the consequence of overlooked covarion effects. The large number of taxa required to properly account for SSRV effects and the resulting extensive running time preclude any attempt to search the tree space with reasonable efficiency, however. If these models have to be used for phylogenetic purposes, it should be in the context of evaluating a small number of competing topologies previously sought using faster algorithms.

Another promising application field is the study of protein adaptation. An important and popular goal of molecular evolution is the detection of positive selection at the sequence level. Classically, this was achieved by comparing nonsynonymous (K_a) and synonymous (K_s) rates of evolution (e.g., Hughes and Nei 1988). Positive selection was invoked when K_a was higher than K_s, which was found for a very small fraction of proteins (Endo, Ikeo, and Gojobori 1996). This approach, however, is limited by averaging of nonsynonymous and synonymous rates over all sites. It would not detect positive selection acting on a few sites. Yang et al. (2000) improved this strategy by applying the ASRV model, therefore separating rather than averaging fast and slow nonsynonymous rates across sites. Yang et al. (2000) found that a larger number of proteins than expected included sites evolving according to a positive-selection-like process. Following their comment, I argue that the importance of positive selection might still be underestimated when data are analyzed under ASRV. This is because the nonsynonymous/synonymous rate of each site is averaged over the whole tree. ASRV cannot detect short episodes of positive selection involving sites which are constrained (i.e., slow) in other parts of the tree. It is quite likely, however, that protein adaptation involves short adaptive episodes, followed by stasis when a new local optimum of fitness has been reached. Yang, Swanson, and Vacquier (2000) dealt with this problem by allowing a distinct K_a/K_s ratio (their ω parameter) in each branch of the tree. This was done at the cost of a large number of additional parameters and of again averaging K_a/K_s over sites.

SSRV models might be the right approach to account for episodic evolution of proteins. Both lineage and site effects are automatically separated at the cost of only one or two parameters. Perspectives of this work therefore include generalization to codon-based models of evolution and the use of data analysis tools for characterizing those sites and lineages involved in positive selection in the context of (U)SSRV models.

Acknowledgments

Many thanks to Olivier Gascuel and Ziheng Yang for helpful comments and suggestions. This work was supported by the Génopole Montpellier-Perpignan.

LITERATURE CITED

ENDO, T., K. IKEO, and T. GOJOBORI. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **5**:685–690.

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

FITCH, W. M. 1971. Rate of change of concomitantly variable codons. J. Mol. Evol. **1**:84–96.

FITCH, W. M., 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.

FORTERRE, P. 1996. A hot topic: the origin of hyperthermophiles. Cell **85**:789–792.

GALTIER, N., and M. GOUY. 1998. Inferring pattern and process: maximum likelihood implementation of a non-homogeneous model of DNA sequence evolution for phylogenetic analysis. Mol. Biol. Evol. **15**:871–879.

GALTIER, N., and J. LOBRY. 1997. Relationships between genomic G+C content, RNA secondary structures and optimal growth temperature in prokaryotes. J. Mol. Evol. **44**: 632–636.

GALTIER, N., N. J. TOURASSE, and M. GOUY. 1999. A nonhyperthermophilic ancestor to extant life forms. Science **283**:220–221.

GERMOT, A., and H. PHILIPPE. 1999. Critical analysis of eukaryotic phylogeny: a case study based on the HSP70 family. J. Eukaryot. Microbiol. **46**:116–124.

GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11**:725–736.

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

HUGHES, A. L., and M. NEI. 1988. Nucleotide substitution at major histocompatibility complex loci reveals overdominant selection. Nature **335**:167–170.

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

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

LOCKHART, P. J., D. H. HUSON, U. MAIER, M. J. FRAUNHOLZ, Y. VAN DE PEER, A. C. BARBROOK, C. J. HOWE, and M. A. STEEL. 2000. How molecules evolves in eubacteria. Mol. Biol. Evol. **17**:835–838.

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

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

PHILIPPE, H., P. LOPEZ, H. BRINKMAN, K. BUDIN, A. GERMOT, J. LAURENT, D. MOREIRA, M. MÜLLER, and H. LE GUYADER. 2000. Early branching or fast evolving eukaryotes? An answer based on slowly evolving positions. Proc. R. Soc. Lond. B Biol. Sci. **267**:1213–1221.

TAMURA, K. 1992. Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. Mol. Biol. Evol. **9**:678–687.

TOURASSE, N. J., and M. GOUY. 1997. Evolutionary distances between nucleotide sequences based on the distribution of substitution rates among sites as estimated by parsimony. Mol. Biol. Evol. **14**:287–298.

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

WAKELEY, J. 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol. Evol. **11**: 158–163.

WOESE, C. R. 1987. Bacterial evolution. Microbiol. Rev. **51**: 221–271.

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

———. 1994. Maximum-likelihood phylogenetic estimation of from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

———. 1995. On the general reversible Markov process model of nucleotide substitution: a reply to Saccone et al. J. Mol. Evol. **41**:254–255.

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

YANG, Z., R. NIELSEN, N. GOLDMAN, and A.-M. K. PEDERSEN. 2000. Codon-substitution-models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

YANG, Z., W. J. SWANSON, and V. D. VACQUIER. 2000. Maximum likelihood analysis of molecular adaptation in abalone sperm lysin reveals variable selective pressures among lineages and sites. Mol. Biol. Evol. **17**:1446–1455.

APPENDIX

Derivation of the Average Relative Rate $\bar{r}_{ij}$ Along a Branch of Length λ Conditional on Initial Rate r_i and Final Rate r_j

With rate ν, the relative rate changes from its current category to any of the g possible categories with equal probabilities. The mean rate along a pathway of length λ starting from r_i and ending with r_j is computed by conditioning on the number of changes C:

$$\bar{r}_{ij} = \sum_{k=0}^{\infty} r(C = k) \times \Pr(C = k/r_i, r_j, \lambda, \nu), \quad \text{(A1)}$$

where $r_{ij}(C = k)$ is the mean relative rate given initial and final rates r_i and r_j, respectively, and given that k changes occurred. First, consider the $i \neq j$ case. Since changes are equiprobable, the mean rate conditional on $k > 0$ changes is

$$r(C = k) = \frac{r_i + r_j + k - 1}{k + 1}. \quad \text{(A2)}$$

Equation (A2) holds because k changes cut the branch into $k + 1$ intervals, the first and last of which have rates r_i and r_j, respectively, with the remaining $k - 1$ having average rate 1 since they are randomly sampled from a distribution of mean 1 (remember that reassignment of current rate is allowed).

The probability that k changes occurred given r_i and $r_j \neq r_i$ is 0 for $k = 0$ and

$$\Pr(C = k/r_i, r_j, \lambda, \nu) = \frac{\Pr(C = k, r_j/r_i, \lambda, \nu)}{\Pr(r_j/r_i, \lambda, \nu)} \quad \text{(A3)}$$

for $k \neq 0$. These probabilities can be computed by noting that (1) the probability of k changes irrespective of initial and final states is $(\lambda\cdot\nu)^k\cdot\exp(-\lambda\cdot\nu)/k!$ (changes occur according to a Poisson process with rate $\lambda\cdot\nu$), and (2) all g possible final states are equiprobable given that at least one change occurred. The numerator and denominator in equation (A3) are therefore

$$\Pr(C = k/r_i, r_j, \lambda, \nu) = \frac{(\lambda \cdot \nu)^k \cdot e^{-\lambda \cdot \nu}}{g \cdot k!} \quad \text{(A4)}$$

$$\Pr(r_j/r_i, \lambda, \nu) = \sum_{m=1}^{\infty} \frac{(\lambda \cdot \nu)^m \cdot e^{-\lambda \cdot \nu}}{g \cdot m!} = \frac{1 - e^{\lambda \cdot \nu}}{g}. \quad \text{(A5)}$$

The ratio (A4)/(A5) simplifies to

$$\Pr(C = k/r_i, r_j, \lambda, \nu) = \frac{(\lambda \cdot \nu)^k}{(e^{\lambda \cdot \nu} - 1) \cdot k!}. \quad \text{(A6)}$$

Substituting equations (A2) and (A6) into equation (A1) and summing (noting that the term corresponding to $k = 0$ in equation (A1) is 0 when $i \neq j$) results in equation (10). Similar reasoning can be used when $j = i$.

DAN GRAUR, reviewing editor

Accepted January 11, 2001