Testing for Covarion-like Evolution in Protein Sequences

Huai-Chun Wang,† Matthew Spencer,*†§ Edward Susko,* and Andrew J. Roger†*

*Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada; †Department of Biochemistry and Molecular Biology, Canadian Institute for Advanced Research, Program in Evolutionary Biology, Dalhousie University, Halifax, Nova Scotia, Canada; and §School of Biological Sciences, University of Liverpool, Crown Street, Liverpool, United Kingdom

The covarion hypothesis of molecular evolution proposes that selective pressures on an amino acid or nucleotide site change through time, thus causing changes of evolutionary rate along the edges of a phylogenetic tree. Several kinds of Markov models for the covarion process have been proposed. One model, proposed by Huelsenbeck (2002), has 2 substitution rate classes: the substitution process at a site can switch between a single variable rate, drawn from a discrete gamma distribution, and a zero invariable rate. A second model, suggested by Galtier (2001), assumes rate switches among an arbitrary number of rate classes but switching to and from the invariable rate class is not allowed. The latter model allows for some sites that do not participate in the rate-switching process. Here we propose a general covarion model that combines features of both models, allowing evolutionary rates not only to switch between variable and invariable classes but also to switch among different rates when they are in a variable state. We have implemented all 3 covarion models in a maximum likelihood framework for amino acid sequences and tested them on 23 protein data sets. We found significant likelihood increases for all data sets for the 3 models, compared with a model that does not allow site-specific rate switches along the tree. Furthermore, we found that the general model fit the data better than the simpler covarion models in the majority of the cases, highlighting the complexity in modeling the covarion process. The general covarion model can be used for comparing tree topologies, molecular dating studies, and the investigation of protein adaptation.

Introduction

The covarion hypothesis of molecular evolution proposes that selective pressures on a given amino acid or nucleotide site are dependent on the identity of other sites in the molecule that change throughout time, resulting in changes of evolutionary rates of sites along the edges of a phylogenetic tree (Fitch and Markowitz 1970). Covarion-like evolution is recognized as an important mode of molecular evolution in proteins, structural RNA genes, and protein-coding genes (Miyamoto and Fitch 1995; Simon et al. 1996; Lockhart et al. 2000; Galtier 2001; Huelsenbeck 2002; Misof et al. 2002; Pupko and Galtier 2002; Ané et al. 2005). The standard covarion process may be seen as a form of heterotachy, which is a general term for within-site rate variation over time (Lopez et al. 2002; Lockhart and Steel 2005). In protein sequences, the heterotachy/covarion process may relate to shifts in protein function (Naylor and Gerstein 2000; Gaucher et al. 2001; Knudsen and Miyamoto 2001; Gaucher et al. 2002; Lopez et al. 2002; Blouin et al. 2003; Inagaki et al. 2003), but for a contrasting view see Philippe et al. (2003). Failure to accommodate certain forms of heterotachy may also lead to biased tree estimation (Lockhart et al. 1998; Inagaki et al. 2004; Susko et al. 2004; Spencer et al. 2005; Lockhart et al. 2006).

The first mathematical models for a covarion process had 2 substitution rate classes: the substitution process at a site could switch between "ON" (variable) and "OFF" (invariable) (Tuffley and Steel 1998; Penny et al. 2001). Huelsenbeck (2002) implemented a version of this model, with the addition of among-site rate variation, in the phylogenetic package MrBayes (Huelsenbeck and Ronquist 2001). Huelsenbeck (2002) found that for 9 of 11 genes, this model provided a better explanation of the data than a model that does not allow rates at sites to change over time. Galtier (2001) developed a different covarion model with an arbitrary number of rate classes. In his model, the switching rates are defined by a discrete gamma distribution, similar to models of rate variation across sites (RAS) (Yang 1994). However, it does not allow rate switching to and from an invariable OFF state. Galtier's model has been implemented in the software NHML (Galtier and Gouy 1998) for nucleotide sequences.

The Huelsenbeck and Galtier models make different assumptions about the ways evolutionary rates change over time. For example, in the Huelsenbeck model, RAS is independent of the covarion process, and all sites experience a covarion process in which evolutionary rates switch between zero and a value that is fixed for the site. In the Galtier model, a variable proportion of sites have fixed rates over time, and the remaining sites switch between a set of nonzero evolutionary rates. The performance of these 2 models has not previously been compared. Furthermore, we do not know how accurately the covarion parameters can be estimated from sequence data.

In this study, we propose a general model that not only allows site rates to switch from ON to OFF and OFF to ON but also allows switching between different rates among the ON states. A different generalization was considered in Xu (2002) in a more restrictive setting. The general model contains the Galtier and Huelsenbeck models as special cases. This nesting of models allows for likelihood ratio tests (LRTs) to assess if the Galtier or Huelsenbeck models provide a sufficient fit to the data. We have implemented all 3 covarion models for amino acid sequences in a maximum likelihood framework in the software package PROCOV. We used PROCOV to test the covarion models on simulated protein sequence data and 23 empirical data sets.

Methods

Models of Covarion Evolution

One of the first models of covarion evolution was proposed by Tuffley and Steel (1998). In addition to the usual Markov model for character state changes, they assume

Key words: covarion, heterotachy, evolutionary rate, Markov model, maximum likelihood, likelihood ratio.

E-mail: hcwang@mathstat.dal.ca.

Mol. Biol. Evol. 24(1):294–305. 2007
doi:10.1093/molbev/msl155
Advance Access publication October 20, 2006

a Markov model for the rates. Rates along an edge switch from OFF to ON and from ON to OFF. When a site is OFF, no substitutions occur and when it is ON, substitutions occur at a constant rate. The model has 2 additional parameters: s_{01} and s_{10}, the rate of transition from the OFF state to the ON state and the corresponding rate for ON to OFF. In this model (and in all models discussed here), switches between classes are not allowed to occur simultaneously with substitutions. The stationary probability of being ON is $s_{01}/(s_{01} + s_{10})$. Huelsenbeck (2002) added among-site rate variation to this model. He allowed each site i to have a fixed substitution rate multiplier r_i drawn from a discrete gamma distribution with g classes, shape parameter α, and mean 1, such that the expected substitution rate per unit time at site i is r_i when the site is ON and 0 when OFF. The Huelsenbeck model is implemented in MrBayes (Huelsenbeck and Ronquist 2001) for both nucleotide and amino acid sequences. Figure 1*A* shows one way of visualizing the Huelsenbeck model.

An alternative covarion model was developed by Galtier (2001). In Galtier's model, a proportion π of sites evolves under the covarion model. The remaining proportion, $1 - \pi$, of sites has a site-specific rate drawn from a discrete gamma distribution with shape parameter α and mean 1. For sites evolving under a covarion model, rates are always elements of the set of rates in this gamma distribution. However, there is a constant rate of switching from any rate class to any other. Each rate class is equiprobable. A possible justification for using the same discrete gamma distribution for covarion and noncovarion sites is that substitution rates are being determined by the same kinds of functional constraints in both cases, even though these constraints are allowed to change over time at covarion sites. The Galtier model is implemented in NHML for nucleotides (Galtier and Gouy 1998; Galtier 2001). Figure 1*B* shows rate switching in the Galtier model.

The Huelsenbeck and Galtier covarion models are not nested, and each requires 3 parameters in addition to the usual edge length and substitution model parameters (s_{01}, s_{10}, and α in Huelsenbeck, and s_{11}, π, and α in Galtier). To help understand differences in performance between these models, we developed a general covarion model of which both the Huelsenbeck and Galtier models are special cases. We allow a covarion site to switch between an ON state with rate drawn from a discrete gamma distribution and a corresponding OFF state (with rates s_{01} and s_{10}, as in the Huelsenbeck model). The stationary probability of being ON is $s_{01}/(s_{01} + s_{10})$. We also allow a covarion site to switch between ON states (with rate s_{11}/g, as in the Galtier model). The stationary probability of each of the ON states is the same. We also allow a proportion $1 - \pi$ of noncovarion sites at which site-specific rates are drawn from the discrete gamma distribution and do not change over time. The general model is shown in figure 1*C*. We can obtain the Huelsenbeck model by setting $\pi = 1$, $s_{11} = 0$, and we can obtain the Galtier model by setting $s_{10} = 0$, $s_{01} > 0$ (so the stationary probability of any OFF state is 0). Thus, both the Huelsenbeck and Galtier models are nested within the general model, which needs 2 more parameters than either. We can then use LRTs to compare the Huelsenbeck and Galtier models with the general model. The relationships between these and other rate variation models are depicted in figure 2.

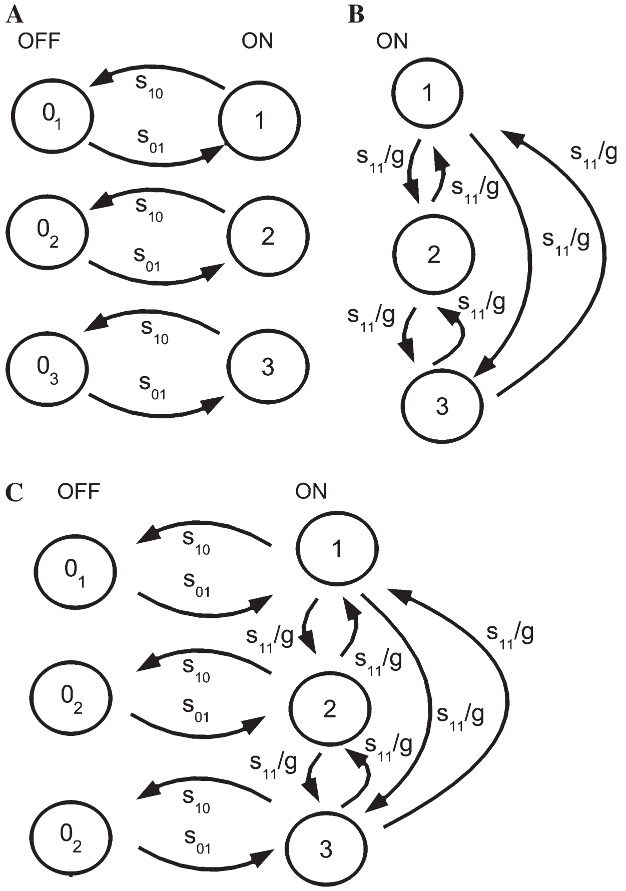

FIG. 1.—Rate switching in the covarion models. (*A*) the Huelsenbeck model. First, an overall rate is drawn for the site from a discrete gamma distribution. In the figure, these are one of the ON states labeled 1–3 with 3 rate classes. Given this overall rate, switching of rate along edges from ON to OFF occurs at rate s_{10} and from OFF to ON at rate s_{01}. Sites change only between ON states and corresponding OFF states (labeled 0_1–0_3). (*B*) The Galtier model. Each site may be in one of g rate classes (here shown with $g = 3$ and labeled 1–3) determined by a discrete gamma distribution. Each class has the same stationary probability, and switching occurs to any other class with rate s_{11}/g. (*C*) The general model. ON states have rates drawn from a discrete gamma distribution (shown here with $g = 3$ classes, labeled 1–3). Switching occurs between ON states and corresponding OFF states (labeled 0_1–0_3) with rates s_{10} and s_{01}. Switching also occurs between ON states with rate s_{11}/g. For both the Galtier and general models, there is also a proportion $1 - \pi$ of noncovarion sites, at which no rate switching occurs.

A full covarion model description consists of 2 Markov processes: a substitution process with rate matrix **M** and the rate-switching process with instantaneous rates of switching given by a rate matrix **G**. Considered jointly (r, x), where r is the rate and x is the character state, gives a Markov process with 2-dimensional state space, also referred to as a Markov-modulated Markov process (Galtier and Jean-Marie 2004). The full rate matrix, **Q**, is then of dimension $m(2g) \times m(2g)$, where m is the number of observable states (4 for nucleotides, 20 for amino acids). $Q_{(r_i,x_k),(r_j,x_l)}$ is the rate at which rate r_i and character state x_k is substituted by r_j and x_l. Because 2 events in a small period of time are unlikely under a Markov model, $Q_{(r_i,x_k),(r_j,x_l)}$ is 0 unless one of $r_i = r_j$ or $x_k = x_l$ holds. Label

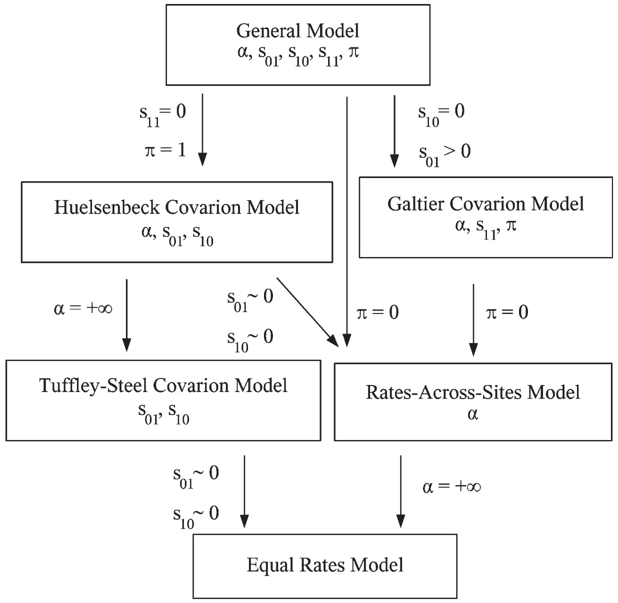

FIG. 2.—Models of amino acid substitution rate evolution implemented in PROCOV and the relationship between the models. The equal rates model assumes no variation in rates of change among sequence sites. For each model, rate variation parameters that can be estimated are shown.

the rate classes 1 … g for ON classes and 0_1 … 0_g for corresponding OFF classes, then $\mathbf{Q}$ can be expressed as follows:

$$\begin{aligned}
&\mathbf{Q}_{(r_i,x_k),(r_i,x_l)} = r_i(s_{01}+s_{10})/s_{01}\mathbf{M}_{x_k,x_l} \quad i \in 1\ldots g; k,l \in 1\ldots m \text{ (amino acid change)}\\
&\mathbf{Q}_{(r_i,x_k),(r_i,x_l)} = 0 \quad i \in 0_1\ldots 0_g \text{ (no amino acid change in OFF classes)}\\
&\mathbf{Q}_{(r_i,x_k),(r_j,x_k)} = \mathbf{G}_{r_i,r_j} \quad i,j \in 1\ldots g; k \in 1\ldots m \text{ (rate change)}\\
&\mathbf{Q}_{(r_i,x_k),(r_j,x_l)} = 0 \quad i \neq j, k \neq l \text{ (simultaneous changes not allowed)}.
\end{aligned} \tag{1}$$

For the general model,

$$\mathbf{G} = \begin{array}{c} \\ 1\\ 2\\ \vdots\\ g\\ 0_1\\ 0_2\\ \vdots\\ 0_g \end{array}\!\!\begin{array}{c} \begin{array}{cccccccc} 1 & 2 & \cdots & g & 0_1 & 0_2 & \cdots & 0_g \end{array}\\ \left(\begin{array}{cccccccc}
* & s_{11}/g & \cdots & s_{11}/g & s_{10} & 0 & \cdots & 0\\
s_{11}/g & * & \cdots & s_{11}/g & 0 & s_{10} & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\
s_{11}/g & s_{11}/g & \cdots & * & 0 & 0 & \cdots & s_{10}\\
s_{01} & 0 & \cdots & 0 & * & 0 & \cdots & 0\\
0 & s_{01} & \cdots & 0 & 0 & * & \cdots & 0\\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots\\
0 & 0 & \cdots & s_{01} & 0 & 0 & \cdots & *
\end{array}\right)\end{array},$$

where diagonal entries are determined by the constraint that rows sum to 0.

Ordering the entries of $\mathbf{Q}$ as (r_1, x_1) … (r_1, x_m), (r_2, x_1) … $(0_g, x_m)$, the rate matrix $\mathbf{Q}$ can be expressed succinctly (Galtier and Jean-Marie 2004) as follows:

$$\mathbf{Q} = \mathbf{D}_R \otimes \mathbf{M} + \mathbf{G} \otimes \mathbf{I}_m, \tag{2}$$

where ⊗ is the Kronecker product and $\mathbf{I}_m$ is an $m \times m$ identity matrix. $\mathbf{D}_R$ is a $2g \times 2g$ diagonal matrix whose (i, i)th entry is the substitution rate for class i:

$$d_r(i,i) = \begin{cases} r_i(s_{01}+s_{10})/s_{01} & i \in 1\ldots g\\ 0 & i \in 0_1\ldots 0_g \end{cases}. \tag{3}$$

In equations (1) and (3), the rates for the ON classes are rescaled by $(s_{01} + s_{10})/s_{01}$, which is required to ensure that edge lengths have the correct interpretation (i.e., the expected number of amino acid substitutions per site), as we now indicate. Similarly, as for usual Markov models of amino acid substitution, the expected number of amino acid substitutions can be shown to be

$$\sum_{i,\,k}\ \sum_{(j,\,l)|k\neq l} p_i \pi_k \mathbf{Q}_{(i,\,k),(j,\,l)} t,$$

where p_i is the equilibrium frequency of the substitution rate class i and π_k is the equilibrium frequency of residue k. Assuming the rate matrix M has been rescaled so that the edge lengths have the correct interpretation under a noncovarion model, for the Q given by (2), we obtain

$$\sum_{i,\,k}\ \sum_{(j,\,l)|k\neq l} p_i \pi_k \mathbf{Q}_{(i,\,k),(j,\,l)} = 1, \tag{4}$$

so that the expected number of substitutions is indeed t.

The computation of the likelihoods involves 2 steps. First, computing the transition probability for an edge length t by taking the matrix exponential of $\mathbf{Q}$:

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \tag{5}$$

where the entry of $\mathbf{P}(t)$ corresponding to row (r_i, x_i) and column (r_j, x_j) is the probability of transition from (r_i, x_i) to (r_j, x_j) in time t. Second, computing site likelihoods by summing over states and rate classes at internal nodes, and rate classes at leaves, using the pruning algorithm (Felsenstein 1981). For example, for a 3-taxon tree $(1{:}t_1, 2{:}t_2, 3{:}t_3)$, the site likelihood for a site pattern AAC is calculated as follows:

$$\begin{aligned}
P(AAC|\mathbf{t},\mathbf{Q}) = \sum_{x_0,r_0,r'_1,r'_2,r'_3} & P\{(R_0,X_0)=(r_0,x_0)\}\\
&\times P\{(r_0,x_0)\rightarrow(r'_1,x_1)|t_1\}\\
&\times P\{(r_0,x_0)\rightarrow(r'_2,x_2)|t_2\}\\
&\times P\{(r_0,x_0)\rightarrow(r'_3,x_3)|t_3\},
\end{aligned} \tag{6}$$

where $\mathbf{t}$ is a vector of edge lengths, R_0 and X_0 are the rate and amino acid at the internal node, respectively, and $x_1 = A$, $x_2 = A$, and $x_3 = C$. The unobservable rate classes r'_1, r'_2, and r'_3 at the leaves are summed independently over all possible values.

Finally, the likelihood of a sequence on a tree is the product of site likelihoods, assuming independence among sites:

$$P(Y|\tau) = \prod_j P(y_j|\tau, \mathbf{Q}), \tag{7}$$

where Y is the sequence data, τ is the given tree with edge lengths, and y_j is site likelihood for site j.

Calculating likelihoods for covarion models are expensive compared with noncovarion models because of the large number of states, even with a fast algorithm for diagonalizing the rate matrix (Galtier and Jean-Marie 2004). All of these covarion models can be expressed as special cases of a general Markov model with more states at internal nodes of the tree than at the leaves. Tree identifiability has been proved for some situations under the general Markov model (Allman and Rhodes 2006).

Implementation

The implementation of the Galtier model for protein sequences in a maximum likelihood framework (Felsenstein 1981) was based on the NHML package that implemented the covarion model for nucleotide sequences (Galtier and Gouy 1998; Galtier 2001). This together with the codes for the Huelsenbeck model and the general model form a package called PROCOV (available at http://www.mathstat.dal.ca/~hcwang/Procov) to optimize the parameters and evaluate the maximum likelihood of a given tree and protein alignment. Figure 2 illustrates the rate substitution models implemented in PROCOV and their relationships. PROCOV may be adapted to evaluate the tree topology under the covarion models. For simulation studies, we wrote a sequence simulator, adapted from seq-gen (Rambaut and Grassly 1997; Ané et al. 2005), to simulate amino acid data under a given tree using the 3 covarion models. The code (seq-gen-aminocov) is available at http://www.liv.ac.uk/~matts/.

Model Testing

Both Huelsenbeck and Galtier models have 2 more parameters than the RAS model. The general model has 4 more parameters than the latter. As the RAS model is nested within the 3 covarion models, and both the Huelsenbeck and Galtier models are nested within the general model, LRTs may be used to compare the covarion models with the RAS model and the general model with the other covarion models. The likelihood ratio statistic 2logΛ, which is twice the difference in log likelihoods between a model and its nested simpler model, is usually asymptotically χ^2 distributed with d.f. degrees of freedom. The appropriate d.f. is typically the difference in the number of free parameters between the 2 models in comparison. For instance, Galtier (2001) used the LRT with 2 d.f. to compare his covarion model and the RAS model. For comparing the Huelsenbeck and RAS models, however, Huelsenbeck (2002) noticed that because, under the simpler model, the parameters are on the boundary of the parameter space, the χ^2 approximation does not hold. For the same reason, the definition of d.f. described above is not appropriate for comparing the general model with the Huelsenbeck or Galtier models. In these cases, the appropriate distribution for the test statistic is a mixture of χ_0^2, χ_1^2, and χ_2^2 (Self and Liang 1987).

For the case of the general versus Galtier models, there is 1 parameter (s_{10}) on the boundary of the parameter space, which corresponds to "Case 6" in Self and Liang (1987). The limiting distribution is a mixture with equal weights of χ_1^2 and χ_2^2 distributions. The p value for a likelihood ratio statistic is calculated as

$$p_{\text{Galtier}} = P(\chi_1^2 > 2\Lambda)/2 + P(\chi_2^2 > 2\Lambda)/2. \qquad (8)$$

For the case of the general versus Huelsenbeck models, there are 2 parameters (s_{11} and π) on the boundary of the parameter space, which matches "Case 7" in Self and Liang (1987). The limiting distribution is a mixture of a point mass at χ_0^2, χ_1^2, and χ_2^2 distributions with weights $1/2 - p$, 1/2, and p, respectively, where

$$p = \frac{\cos^{-1}(I_{12}/\sqrt{I_{11}I_{12}})}{2\pi}, \qquad (9)$$

where I_{12} is the entry of the Fisher information matrix corresponding to s_{11} and π, and I_{11} and I_{22} are the entries corresponding to s_{11} alone and π alone, respectively. There is a positive probability that if the likelihoods for the Huelsenbeck and general models be the same, then the p value is simply 1/2 + p. When they are not, the p value can be calculated as:

$$p_{\text{Huelsenbeck}} = P(\chi_1^2 > 2\Lambda)/2 + pP(\chi_2^2 > 2\Lambda). \qquad (10)$$

Similarly, the case of the Huelsenbeck versus RAS models also matches "Case 7" (Huelsenbeck 2002), and the p value can be computed as equation (10), where I_{12} is the entry of the Fisher information matrix corresponding to s_{01} and s_{10}, and I_{11} and I_{22} are the entries corresponding to s_{01} alone and s_{10} alone.

For comparison between the general and RAS models, either $s_{01} = s_{10} = s_{11} = 0$ or $\pi = 0$ will give the RAS model. A closed form expression for the limiting distribution is not available. We, therefore, calculated a conservative p value as $P(\chi_4^2 > 2\Lambda)$; the real p value would be smaller.

For comparison between the Huelsenbeck and Galtier models, both have equal numbers of parameters and are not nested. Thus, information criteria like Akaike and Bayesian information criteria (AIC and BIC) favor the model with the larger likelihood.

Data Analysis

We examined 23 amino acid data sets with the 3 covarion models. Twenty-one of these were selected for analysis from online (Pfam Release 14.0) and in-house alignment databases. In order to have sufficient sequence length and taxonomic sampling, data sets were only retained if they had 30–100 taxa and >200 sites after alignment trimming. Alignment trimming was performed using the program GBlocks version 0.91b (Castresana 2000) with a maximum number of contiguous nonconserved positions of 16 and minimum block length of 5. The sequence alignments are available from one of us (A.J.R.) upon request.

The data sets used include 48 eukaryotic actin protein sequences, 36 acetyl-CoA carboxylase (Carboxyl_trans) sequences, 41 60-kDa chaperonin (CPN60) sequences, 65 CTP synthase sequences, 49 DNA topoisomerase IV subunit A (GyrA) sequences, 38 elongation factor 1a (EF-1a) sequences, 37 elongation factor 2 (EF-2) sequences, 36 intermediate filament protein (Filament)

Table 1
Performance of PROCOV on 3 Simulated Data Sets Based on a CPN60 Tree and the 3 Covarion Models

Data[a]	Model[b]	α	s_{01}	s_{10}	s_{11}	π	LnL
I		0.46	1.875	1.25	0	1	−15,241.17
	Huelsenbeck	0.52 (0.04)	1.71 (0.27)	1.12 (0.22)	0	1	−15,239.58
	Galtier	0.21 (0.03)	—	0	3.48 (0.78)	0.46 (0.04)	−15,253.0
	General	0.52 (0.05)	1.73 (0.27)	1.13 (0.22)	0 (0.03)	1 (0.02)	−15,239.58
II		0.46	—	0	1.5	0.6	−17,574.67
	Galtier	0.52 (0.05)	—	0	1.41 (0.66)	0.55 (0.15)	−17,573.71
	Huelsenbeck	1.50 (0.30)	0.73 (0.16)	0.30 (0.08)	0	1	−17,579.82
	General	0.6 (0.07)	0.004 (0.00)	0.0004 (0.00)	1.25 (0.49)	0.68 (0.17)	−17,571.96
III		0.46	1.5	2	2.5	0.6	−15,656.85
	General	0.57 (0.08)	0.94 (0.31)	1.87 (0.61)	2.0 (1.11)	0.50 (0.06)	−15,653.12
	Huelsenbeck	1.24 (0.31)	1.24 (0.15)	0.78 (0.12)	0	1	−15,665.57
	Galtier	0.28 (0.03)	—	0	1.89 (0.57)	0.66 (0.10)	−15,668.15

[a] Data set I, II, and III were simulated under the Huelsenbeck, Galtier, and general models, respectively. For each data set, the first line lists the true values for the fitted parameters and the log likelihood computed by PROCOV with the corresponding model that was used in the simulation; the second line lists parameters estimated by the model used in the simulation. The third and fourth lines are estimates by the models other than the one that was used for simulating the data. Values in brackets are standard errors.

[b] For the Huelsenbeck model, the s_{11} and π are defined as 0 and 1, respectively. For the Galtier model s_{01} is not relevant (can be any value greater than 0) and indicated by ‘‘—.’’ The s_{10} for the Galtier model is defined as 0. These 3 parameters were not optimized for the 2 models.

sequences, 40 glutamate synthase aminotransferase (Glu_synth_NTN) sequences, 34 70-kDa heat shock protein (HSP70) sequences, 54 90-kDa heat shock protein (HSP90) sequences, 51 ILVD_EDD dehydratase family sequences, 41 NADH dehydrogenase I chain F (NuoF) sequences, 40 minichromosome maintenance protein (MCM) sequences, 43 mitochondrial processing peptidase (MPP) sequences, 32 MreB/Mb1 sequences, 34 potyvirus coat protein sequences, 70 SecA sequences, 54 α-tubulin sequences, 46 β-tubulin sequences, and 36 fimbrial usher protein (Usher) sequences. Two multigene data sets corresponding to a published analysis of metazoa (Peterson and Butterfield 2005) and the chloroplast genomes of land plants (Leebens-Mack et al. 2005) were used to assess the impact of accounting for covarion-like evolution in multigene data sets. The Peterson and Butterfield data set (PB2005) consisted of 32 taxa with 8 concatenated proteins: mitochondrial cytochrome oxidase I, mitochondrial atpB, aldolase, methionine adenosyltransferase, triosephosphate isomerase, EF-1a, phosphofructokinase, and catalase. The chloroplast data set consisted of 24 taxa with 61 concatenated chloroplast-encoded proteins. Each data set consisted of a protein sequence alignment and an initial tree with edge lengths precomputed with PHYML with the JTT + Γ model (Guindon and Gascuel 2003).

For each data set, we ran PROCOV with 4 gamma rate categories, JTT substitution model (Jones et al. 1992), and the 3 covarion models. To compare with the RAS model, we also ran PROCOV for JTT + 4 gamma rate categories and set proportion of covarion sites (π) to 0 under the general model. The covariance matrices for the parameters were estimated from the inverse of the Fisher information matrix for all parameters other than edge lengths. They were also used to compute the Taylor series approximation to the likelihood surface around the estimated parameters.

For simulation studies, we used seq-gen-aminocov to simulate 3 data sets for the 3 covarion models based on a tree from a subset of the CPN60 data (17 taxa). For each data set, the simulated sequences were 1,000 amino acids long, using 4 gamma rates with shape parameter $\alpha = 0.46$. The covarion parameters were set according to the models. We then used PROCOV to estimate the parameters by fixing the topology and edge lengths at true values. The covariance matrices for the parameters were computed to obtain variances of the parameter estimates.

Results

Simulation Studies

We used seq-gen-aminocov to simulate a data set (data set I in table 1) based on the CPN60 tree under the Huelsenbeck model and fitted parameters (17 taxa, 1,000 sites, JTT substitution model, and 4 gamma rates with shape parameter $\alpha = 0.46$, $s_{01} = 1.875$, and $s_{10} = 1.25$). The log likelihood (LnL) of the tree for the true parameters is −15,241.17. Fixing edge lengths at their true values, we ran PROCOV on this data set under the Huelsenbeck model to estimate α, s_{01}, and s_{10} and their variances (table 1). We got $\alpha = 0.52$, $s_{01} = 1.71$, and $s_{10} = 1.12$. The estimated maximum LnL = −15,239.58, which is slightly better than the likelihood under the true values of the parameters. Figure 3*A* is a contour plot of the confidence regions computed by interpolating on a grid of values for s_{01} and s_{10} and based on the covariance matrix obtained with PROCOV after the optimization process is finished. The true values of s_{01} and s_{10} are located within the 50% confidence intervals (CIs) of the estimated parameters. The figure also shows that s_{01} and s_{10} are positively correlated. The correlation coefficient for the 2 switching rates calculated from the covariance matrix is 0.73. Furthermore, we used MrBayes (Huelsenbeck and Ronquist 2001) for this simulated data set to estimate s_{01} and s_{10} by fixing α, the tree topology, and edge lengths, using JTT and 4 gamma rates. The mean posterior s_{01} is 2.06 (posterior standard deviation 0.26), the mean posterior s_{10} is 1.30 (0.25), and mean posterior LnL = −15,241.50. The true values of the parameters are also within the 50% CIs of the estimates.

To test the performance of PROCOV on the Galtier model, we simulated a data set (data set II in table 1) based on the CPN60 tree under this model ($\alpha = 0.46$, $s_{11} = 1.5$, and $\pi = 0.6$). The LnL for the true parameter values is −17,574.67. Fixing edge lengths at their true values, we ran PROCOV under the Galtier model and obtained the

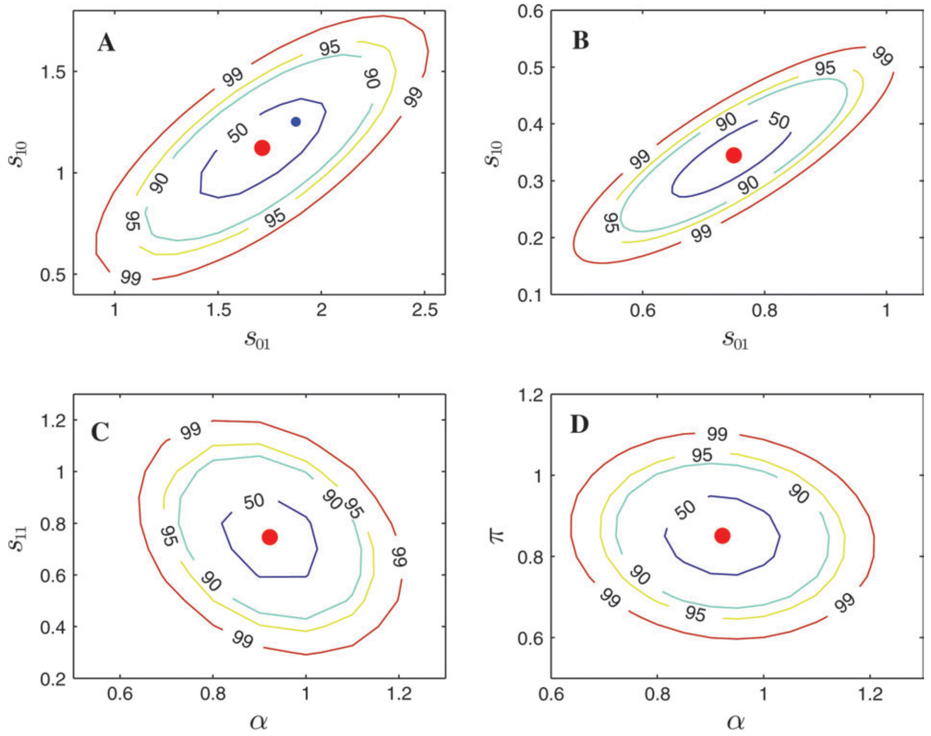

FIG. 3.—Contour plots for likelihood surfaces, computed by Taylor series approximation to the likelihood surface around the estimated parameters. The contours from inner to outer represents 50%, 90%, 95%, and 99% of the CIs of the estimated parameters. The points on the same contour line have the same likelihood, and points in the inner contour have higher likelihood than in the outer contour. The dot at the center represents the estimated values for the corresponding parameters. (*A*) Likelihood surface with respect to s_{01} and s_{10} for a simulated data set under the Huelsenbeck model (CPN60, 17 taxa, 1,000 sites, JTT + 4 gamma rates, $\alpha = 0.46$, $s_{01} = 1.875$, $s_{10} = 1.25$) and computed under the Huelsenbeck model. The smaller dot represents the true values of s_{01} and s_{10} used for the simulation. (*B*) Likelihood surface with respect to s_{01} and s_{10} for the HSP70 data set computed under the Huelsenbeck model. (*C*) Likelihood surface with respect to α and s_{11} for the MPP data set computed under the Galtier model. (*D*) Likelihood surface with respect to α and π for the MPP data set computed under the Galtier model.

following estimates: $\alpha = 0.52$, $s_{11} = 1.41$, and $\pi = 0.55$. The true values of the parameters are within the 50% CIs of the estimates (table 1). The maximum LnL is −17,573.71, very close to the true value.

Finally, to evaluate the performance of PROCOV on the general model, we simulated a data set (data set III in table 1) based on the CPN60 tree under this model ($\alpha = 0.46$, $s_{01} = 1.5$, $s_{10} = 2.0$, $s_{11} = 2.5$, and $\pi = 0.6$). The LnL for the true parameter value is −15,656.85. Fixing edge lengths at their true values, we ran PROCOV under the general model and obtained the following estimates: $\alpha = 0.57$, $s_{01} = 0.94$, $s_{10} = 1.87$, $s_{11} = 2.0$, and $\pi = 0.5$. The true parameter values are within the 50% (for s_{10} and s_{11}) or 95% (for α, s_{01}, and π) confidence regions of the estimates. The CIs are rather large for the general model (table 1). Therefore, it might be hard to get accurate parameter estimates under the general model from single-gene data sets. The maximum LnL (−15,653.12) is still better than the true value.

To test whether the general model can converge to the Huelsenbeck or Galtier models when the data set is simulated under these models, we applied PROCOV under the general model to the data sets I and II. For the data set I (simulated under the Huelsenbeck model), the general model got the following estimates: $\alpha = 0.52$, $s_{01} = 1.73$, $s_{10} = 1.13$, $s_{11} = 0.0$, $\pi = 1.0$, and LnL = −15,239.58. This result shows the general model can perfectly converge to the Huelsenbeck model when the data is constructed under the Huelsenbeck model. For the data set II (simulated under the Galtier model), the general model got the following estimates: $\alpha = 0.60$, $s_{01} = 0.0038$, $s_{10} = 0.0004$, $s_{11} = 1.25$, $\pi = 0.68$, and LnL = −17,571.96. Here, π and s_{11} are close to the true values (0.6 and 1.5, respectively), but the estimated α is a little higher than the true value. For s_{01} and s_{10}, the general model was close to the Galtier model, that is, s_{01} can be any positive value and s_{10} should be 0 or very small. Therefore, the general model also recovered the right covarion parameters when the data set was simulated under the Galtier model.

In contrast, the Galtier model did not perform well for the data sets simulated under the Huelsenbeck or general models (table 1). For data set I, the maximum LnL from the Galtier model is 13.42 less than that under the right model (i.e., the Huelsenbeck model). For data set III, the maximum LnL from the Galtier model is 15.03 less than that under the right model (i.e., the general model). Similarly, the Huelsenbeck model did not perform well for the data sets simulated under the Galtier or general models (table 1). For data set II, the maximum LnL from the Huelsenbeck model is 6.11 less than that under the right model (i.e., the Galtier model). For data set III, the maximum LnL from the Huelsenbeck model is 12.45 less than that under the right model (i.e., the general model).

Table 2
Increase of Maximum Log Likelihoods in Covarion Models Compared with the RAS Model

Data Set	Taxa	Sites	LnL RAS[a]	ΛHuelsenbeck[b]	ΛGaltier[b]	ΛGeneral[b]
Actin	48	363	−6,877.04	21.24	23.07	34.46
Carboxyl_trans	36	212	−9,648.41	92.15	85.97	93.32
CPN60	41	466	−17,233.67	47.7	37.46	51.83
CTP synthetase	65	212	−13,644.81	51.77	45.21	66.27
EF-1a	38	361	−9,543.94	63.22	45.68	69.09
EF-2	37	669	−20,559.4	67.73	36.66	72.63
Filament	36	210	−10,244.22	69.13	70.08	76.22
Glu_synth_NTN	40	253	−11,954.52	30.69	24.77	30.53
GyrA	49	228	−12,872.32	106.69	108.32	119.65
HSP70	34	432	−16,201.13	135.8	119.61	136.26
HSP90	54	459	−15,135.61	85.91	44.53	92.5
ILVD_EDD	51	310	−18,655.35	131.8	143.7	149.48
MCM	40	220	−9,046.91	65.85	71.39	78.6
MPP	43	203	−10,962.9	51.18	56.83	58.85
MreB/Mbl	32	275	−10,769.36	30.15	17.27	35.22
NuoF	41	405	−10,266.97	76.65	63.09	85.1
Potyvirus coat	34	212	−8,000.31	62.19	44.27	63.6
SecA/DEAD	70	203	−13,263.09	121.07	90.6	127.15
α Tubulin	54	375	−7,669.76	20.33	29.68	34.73
β Tubulin	46	382	−7,110.85	35.49	27.6	42.4
Usher	36	317	−17,936.73	41.33	37.16	45.38
PB2005	32	2,051	−44,774.16	165.94	48.34	176.82
Chloroplast	24	15,546	−175,920.8	420.78	224.75	499.23

[a] LnL RAS is the maximum log likelihood obtained from the RAS model.

[b] ΛHuelsenbeck, ΛGaltier, and Λgeneral are the log-likelihood difference between the Huelsenbeck and RAS models, between the Galtier and RAS models, and between the general and RAS models, respectively. The likelihood ratio statistic 2ΛHuelsenbeck, 2ΛGaltier, and 2Λgeneral are all very significant for each data set, see main text.

These simulation studies indicate that PROCOV can obtain good parameter estimates from data simulated under the Huelsenbeck or Galtier models. The general model has the advantage that can recover either model; however, because it has more parameters, we expect the standard errors to be larger.

Although the simulation studies were primarily designed to examine the performance of the 3 covarion models, we also wanted to evaluate the influence of the different covarion processes on the RAS measured by α estimated under the RAS model. For the 3 data sets simulated under the Huelsenbeck, Galtier, and general models (α was fixed at 0.46 for the simulations), the estimated α under the RAS model is 0.46 ± 0.024, 0.77 ± 0.044, and 0.57 ± 0.032, respectively. These suggest the RAS model would underestimate the rate variation among sites if the data set is constructed under Galtier-style covarion process, whereas it is less affected by the general process and virtually not affected by the Huelsenbeck process. The latter case is probably due to the fact that the rate multiplier from the Huelsenbeck process is equal for all sites, and thus the covarion process does not give much additional overall rate variation and hence α is not reduced. For the Galtier process, however, there is no overall rate multiplier and thus not as much overall variation in the data as in the data simulated under the Huelsenbeck process, and therefore a larger α was estimated.

In the above simulation studies, the edge lengths were fixed, as we wanted to specifically investigate the identifiability of the covarion parameters under the different models. We further did simulations where edge lengths were also optimized in addition to the covarion parameters. The estimated covarion parameters were not quantitatively different from that estimated by fixing edge length, and the likelihoods were better, as expected, than those obtained by using the true parameters in each case (data not shown). The original tree length is 3.92. The estimated tree lengths under the Huelsenbeck, Galtier, and general models were 3.76, 3.81, and 3.87, respectively.

Testing on 23 Protein Data Sets

Table 2 lists the maximum log likelihoods (LnL) estimated under the RAS model and the difference in log likelihoods between the 3 covarion models and the RAS model for 23 empirical data sets. The range of the increase in LnL is from 20.33 to 420.78 in the Huelsenbeck model. The likelihood ratio statistic is twice as big. Using equation (10), the p value is less than 10^{-9} for all cases. This is very significant, even considering the Bonferroni correction for the multiple tests with an overall $\alpha = 0.01$ being 0.01/23 (0.0004). The range of the increase in LnL is from 17.27 to 224.75 in the Galtier model. Simply using the LRT with 2 d.f., the p value for the test statistics is less than 0.0005 in all cases, which is also very significant. The range of the increase in LnL is from 30.53 to 499.23 in the general model. This is also very significant for a LRT with 4 d.f. Not surprisingly, for all 3 covarion models, the biggest increases in LnL over the RAS model is in the chloroplast data set that concatenates 61 protein sequences. Another multigene data set, PB2005, also shows second biggest increases in LnL for the Huelsenbeck and general models. For the Galtier model, the second biggest LnL increase is in the ILVD_EDD data set. It is also the third largest LnL increase for the general model and the fourth largest increase for the Huelsenbeck model. HSP70 shows the third largest

Table 3
Difference of Maximum Log Likelihood among 3 Covarion Models

Data Set	$\Lambda1$[a]	$\Lambda2$[b]	$\Lambda3$[b]
Actin	−1.83	13.22***	11.39***
Carboxyl_trans	6.18	1.17	7.35**
CPN60	10.24	4.13*	14.37***
CTP synthetase	6.56	14.50***	21.06***
EF-1a	17.54	5.87**	23.41***
EF-2	31.07	4.9*	35.97***
Filament	−0.95	7.09**	6.14**
Glu_synth_NTN	5.92	−0.16[c]	5.76**
GyrA	−1.63	12.96***	11.33***
HSP70	16.19	0.46	16.65***
HSP90	41.38	6.59**	47.97***
ILVD_EDD	−11.9	17.68***	5.78**
MCM	−5.54	12.75***	7.21**
MPP	−5.65	7.67**	2.02
MreB/Mbl	12.88	5.07*	17.95***
NuoF	13.56	8.45***	22.01***
Potyvirus coat	17.92	1.41	19.33***
SecA/DEAD	30.47	6.08**	36.55***
α Tubulin	−9.35	14.4***	5.05*
β Tubulin	7.89	6.91**	14.8***
Usher	4.17	4.05*	8.22**
PB2005	117.6	10.88***	128.48***
Chloroplast	196.04	78.44***	274.48***

[a] $\Lambda1$ is the log-likelihood difference between the Huelsenbeck and Galtier models; if it is positive, then the Huelsenbeck model is favored; the Galtier model is favored otherwise.

[b] $\Lambda2$ and $\Lambda3$ are the log-likelihood difference between the general and Huelsenbeck models and between the general and Galtier models, respectively. The likelihood ratio statistic 2Λ is a mixture of χ^2_0, χ^2_1, and χ^2_2 distribution. The Bonferroni correction for multiple tests with an overall $\alpha = 0.05$ is 0.00217. ***p value < 0.0001 (very highly significant); **p value < 0.00217 (very significant); *p value < 0.05 (significant).

[c] The slightly lower log likelihood of the general model than the Huelsenbeck model arose from the precision that PROCOV used to terminate the Newton–Raphson iterations in the algorithm. The precision was set to 10^{-4} in this case so that when the likelihood difference between any 2 successive Newton–Raphson cycles is less than 10^{-4}, then the program assumes that convergence is reached and stops further parameter optimization. However, if the precision is set to 10^{-8}, which requires a longer time to reach convergence, the general model obtains a higher likelihood than the Huelsenbeck model with a difference of 0.03.

LnL increase for the Huelsenbeck and Galtier models and the fourth largest increase for the general model.

Table 3 shows the differences in LnL among the 3 covarion models and the significance of the test statistics. Of the 23 data sets, 16 data sets have LnL greater for the Huelsenbeck model than for the Galtier model; the Galtier model has better likelihoods in the remaining 7 data sets. Because both models are not nested and have the same number of parameters in optimization, larger LnL means the model is favored according to the AIC or BIC criterion. Except for 1 data set (Glu_synth_NTN), the general model has higher likelihoods in the other 22 data sets compared with Huelsenbeck model, of which 19 are significant. The general model has higher likelihoods in all 23 data sets compared with the Galtier model. Except for 1 data set, the differences for the other 22 data sets are all significant.

Supplementary table S1, Supplementary Material online lists the parameter estimations for RAS and the 3 covarion models. Six data sets have estimated s_{11} for the general model equal to 0 (Carboxyl_trans, Glu_synth_NTN, HSP70, MreB/Mbl, and β-tubulin) or very small (0.05 in SecA/DEAD), which implies that a Huelsenbeck-style covarion process is favored. Indeed, the maximum LnL are greater for the Huelsenbeck model than for the Galtier model in these 6 data sets (see table 3). Of the 4 data sets that have nonsignificant likelihood difference between the Huelsenbeck and general models (Carboxyl_trans, Glu_synth_NTN, HSP70, and Potyvirus coat protein), 3 have s_{11} in the general model equal to 0, suggesting a Huelsenbeck-style model. For the fourth data set, the Potyvirus coat protein, the general model got the same s_{01} and s_{10} estimates as the Huelsenbeck model. The proportion of covarion sites (π), estimated at 0.95, is also very close to the Huelsenbeck π, which is defined as 1.0. For these data, the estimated general model was similar to a Huelsenbeck model, and therefore, no significant difference in the likelihoods was obtained between the 2 models. Across the 23 data sets for the Huelsenbeck model, the Pearson correlation coefficient (R) for s_{01} and s_{10} is 0.67; whereas the R for s_{01} and s_{10} with α are both very small. A contour plot for the likelihood surface for HSP70, which fits the Huelsenbeck model well, shows a positive correlation between s_{01} and s_{10} (fig. 3*B*). Huelsenbeck (2002) also found that s_{01} and s_{10} are positively correlated for each of the 11 genes he tested.

Only 1 data set (MPP) shows nonsignificant likelihood difference (2.02) between the Galtier and general models (table 3). The parameter estimates for the 2 models were similar and for the general model s_{01} was very small (0.09). Thus, for this data set, the general model behaved more like the Galtier model than the Huelsenbeck model. For the α-tubulin data set, the parameter estimates for the 3 models also suggest that a Galtier-style model is favored over the Huelsenbeck model. Indeed, the log-likelihood difference between the general and Galtier models (5.05) is second smallest among the differences for the 23 data sets, and the likelihood for the Galtier model is much better than that for the Huelsenbeck model.

Supplementary table S1, Supplementary Material online also shows that the estimated α values are smallest in all data sets for Galtier model and largest in 18 data sets for the Huelsenbeck model among the 4 models (RAS and 3 covarion models). Part of the reason probably has to do with the restriction that all sites undergo a covarion process under the Huelsenbeck model. The differences in residence times in ON states across sites provide a partial explanation for sites with unusually large or small numbers of amino acid differences without requiring highly variable rates. For the Galtier model, some proportions of sites are noncovarion and for these, the only explanation for unusually large or small numbers of amino acid differences is large or small site-specific rates. Consistent with this notion, we see a weak positive correlation between α and π across the 23 data sets for the Galtier model ($R = 0.24$, $p = 0.078$). There is also a weak negative correlation between α and s_{11} ($R = -0.26$). Figure 3*C* and *D* shows contour plots of likelihood surface with regard to α and s_{11} and to α and π, respectively, for MPP, which fits the Galtier model well. The figures show within this data set that there is also a negative correlation between α and s_{11} but no correlation between α and π. Galtier

Table 4
Original Tree Lengths and the Differences of Estimate Tree Lengths from the Original Tree Lengths

Data Set	Original	ΔRAS[a]	ΔHuelsenbeck[a]	ΔGaltier[a]	ΔGeneral[a]
Actin	4.33	−0.01	−0.33	−0.01	−0.2
Carboxyl_trans	21.27	−0.17	−0.34	−0.52	−0.44
CPN60	9.76	−0.02	0.06	0.35	−0.1
CTP synthetase	21.84	−0.08	−2.17	0.57	−0.45
EF-1a	8.26	−0.02	−1.14	−0.79	−1.14
EF-2	7.75	−0.03	0	0.12	0.08
Filament	15.09	−0.12	0.71	0.86	1.26
Glu_synth_NTN	15.69	−0.12	−0.67	0.42	−0.34
GyrA	19.79	−0.1	−0.27	1.84	1.39
HSP70	13.4	−0.08	−0.17	−0.33	−0.2
HSP90	7.8	−0.02	−0.4	−0.31	−0.78
ILVD_EDD	19.61	−0.09	−0.06	2.13	2.18
MCM	15.77	−0.11	−1.26	2.03	0.34
MPP	17.39	−0.05	1.37	1.6	1.55
MreB/Mbl	13.53	−0.03	−0.07	0.72	−0.74
NuoF	5.97	0	−0.45	0.02	−0.3
Potyvirus coat	15.9	0	−2.07	−2.41	−2.21
SecA/DEAD	25.65	0.02	−3	1.47	−2.08
α Tubulin	4.39	0	−0.03	0.17	0.09
β Tubulin	3.73	−0.02	0.04	0.01	0.06
Usher	18.87	−0.1	2.03	2.34	2.2
PB2005	6.17	0.08	−0.28	−0.06	−0.37
Chloroplast	1.82	0.57	0.53	0.56	0.53

[a] ΔRAS, ΔHuelsenbeck, ΔGaltier, and Δgeneral are the differences by the respective models from the original tree lengths.

(2001) noticed that α values are larger in the RAS model than in the covarion model for ribosomal RNA genes.

The general model estimated 5 parameters from each of the 23 data sets (supplementary table S1, Supplementary Material online). Among the data sets, there is a strong correlation between s_{01} and s_{10} ($R = 0.7$) and weak correlations between some other parameter pairs, such as s_{01} and α ($R = -0.46$), s_{11} and α ($R = -0.34$), etc. We also computed the correlations between the parameters within each of the 23 data sets from the covariance matrices obtained from the inverse of Fisher information matrices. Significant correlations for the 23 within data set correlations include s_{01} and s_{10} (mean $R = 0.61 \pm 0.048$), s_{11} and α ($R = -0.38 \pm 0.042$), and s_{11} and π ($R = -0.22 \pm 0.048$). The correlations between the other parameter pairs are not significant.

Table 4 lists the original tree lengths (the sum of edge lengths of all internal and terminal nodes) and the difference between the estimated tree lengths under the 4 models and the original tree lengths for the 23 data sets. All models have tree lengths greater than the original tree lengths in some data sets but shorter in the other sets. The RAS model has closest tree lengths to the original lengths. This is not surprising as the original trees were also evaluated under a RAS model with PHYML. The Huelsenbeck model tends to estimate shorter tree lengths, and the Galtier model estimated longer tree lengths than the original lengths. Galtier (2001) also noticed that tree lengths are longer in the covarion model than in the RAS model for the 16S and 23S ribosomal RNA genes. For the general model, the tree length estimates tend to be between the estimates by the Huelsenbeck and the Galtier models (in 15 out of the 23 data sets).

Table 5
Maximum Log Likelihoods for 4 Trees of Angiosperm Chloroplast Genomes Computed under the General Model and the RAS Model

Tree[a]	General	RAS
A	−175,421.41	−175,920.80
B	−175,417.99	−175,922.94
C	−175,441.63**	−175,944.71*
D	−175,515.31***	−176,028.59***

[a] The chloroplast genome trees were taken from Leebens-Mack et al. (2005), Fig. 1. Trees A, B, C, and D put *Amborella*, *Amborella* plus water lilies, water lilies, and monocots at the base of the angiosperms, respectively. The p values of the approximately unbiased tests for the 4 trees were computed with consel (Shimodaira and Hasegawa 2001). *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$.

Estimating a Tree Topology

One application of the general model would be to estimate tree topology. Although the current version of the PROCOV program cannot be directly used for tree topology search as extensive computations are required for the large number of amino acid states and switching rates, it can be used in comparing several competing tree topologies under the (general) covarion model. For instance, the place of *Amborella* within the radiation of angiosperms has evoked a debate about the basal node in angiosperm phylogeny (Goremykin et al. 2003, 2004; Soltis et al. 2004; Lockhart and Penny 2005; Martin et al. 2005). Using the chloroplast genome data (61 protein-coding genes from 24 plant taxa), Leebens-Mack et al. (2005) recently compared 4 hypothesized resolutions of the angiosperm phylogeny: 1) *Amborella* sister to all other angiosperms, 2) *Amborella* plus water lilies clade sister to all other angiosperms, 3) water lilies alone at the base of the tree, and 4) monocot at the base. Their study found weak support for *Amborella* and water lilies at the base of the angiosperms, that is, trees A and B are weakly supported in their RAS plus I (invariant sites) analyses of the amino acid and nucleotide alignments, respectively. Specifically, they found tree A is weakly supported by the amino acid sequence data and tree B is weakly preferred by the nucleotide data, whereas trees C and D are poorly supported by both. This result argued against some earlier studies (Goremykin et al. 2003, 2004) that put monocots at the basal node (i.e., tree D). As shown above, the covarion models applied to the chloroplast genome data based on tree A and the covarion models and especially the general model give better fits to the data than the RAS model (tables 2 and 3). It is interesting to see whether the general model can distinguish between the 4 tree topologies.

Table 5 shows that for the amino acid data, the general model prefers tree B marginally over tree A, whereas both trees C and D have significantly smaller log likelihoods. Furthermore, both the Huelsenbeck and Galtier models obtained qualitatively same results as the general model (data not shown). Thus, although the amino acid covarion models reject the same trees as the RAS model, the optimal topology is different (although the differences in the likelihood for these topologies are small and not significant in either case).

Discussion

We have developed a new covarion model that combines both Huelsenbeck and Galtier models, allowing evolutionary rates of sequence sites not only to switch from ON to OFF and OFF to ON, as in the Tuffley–Steel/Huelsenbeck model but also to switch among different ON states, as in the Galtier model. We have implemented these models in a maximum likelihood framework for amino acid sequence alignments in PROCOV. Simulation studies indicated that PROCOV can find the right parameter values (gamma shape parameter α and switching rates) for data sets simulated under the Huelsenbeck, Galtier, or general models, although the latter has bigger standard errors because of more parameters to be estimated. The behavior of the general model in PROCOV converges to the Huelsenbeck model or Galtier model, when the data set is simulated under either model.

Covarion processes are not directly observable. One can only infer them indirectly from the manner in which amino acids differ across sites throughout the tree. One might expect, therefore, that parameter estimates under a covarion process may not be reliable. The standard errors obtained in our simulations (table 1), although not small, were never so large as to render estimation meaningless. Estimation under a wrong model, however, did tend to give misleading parameter estimates, a point that is further illustrated by the large α estimates obtained under the Huelsenbeck models in the data analyses.

We tested the 3 covarion models on 23 empirical protein data sets and found significant likelihood increases in all data sets for the 3 models, compared with the RAS model. The increases in likelihoods by comparison with the RAS model were the largest regardless of the covarion models. This suggests that a substantial proportion of covarion-like rate variation can be explained by simple covarion models. Galtier (2001) found his model significantly increased the fit of the data for the 16S and 18S rRNA genes. Huelsenbeck (2002) showed his model gave a better fit of the data for 9 out of 11 genes. More recently, Ané et al. (2005) found covarion effects in 26 out of 57 plastid genes. Ignoring the nature of the data sets used by these authors and current studies, it seems that the method of Ané et al. is a little more conservative. Comparing the 3 covarion models implemented in PROCOV, the Huelsenbeck model gave better explanations than the Galtier model in 16 of the 23 data sets, whereas the Galtier model performed better in 7 data sets. This suggests that both models have advantages and disadvantages for different proteins, highlighting the usefulness of a general model. Indeed, the general model gave a significantly better fit to the data than either Galtier or Huelsenbeck model in the majority of the cases studied. The few data sets for which the general model did not perform significantly better than the other models (4 for the Huelsenbeck model and 1 for the Galtier model) either follow Huelsenbeck- or Galtier-style rate variation, so the general model simply converged to the simpler models, as indicated in the parameter estimations and also demonstrated in the simulation studies.

The covarion models considered here differ from models that have been previously shown to result in inconsistent topological estimation (Kolaczkowski and Thornton 2004; Susko et al. 2004) in that they are stationary processes throughout the tree. It is not clear that failing to account for covarion-like evolution will generally result in topological misestimation. The estimation of the basal node in the angiosperm phylogeny based on the chloroplast genome data that we considered here is only one of a few real data examples that we know of where incorporating covarion models of heterotachy results in different estimated topology. The others are a 3-gene analysis of opisthokont phylogeny (Ruiz-Trillo et al. 2004) and an analysis of plastid-derived genes in dinoflagellates (Schalchian-Tabrizi et al. 2006). In both of these cases, Bayesian analyses with a Huelsenbeck covarion model yielded different and more credible phylogenies than corresponding analyses using standard RAS models. The extent to which covarion models more generally impact on phylogenetic estimation therefore deserves further investigation.

Another application is in molecular dating. Even if tree topology estimated under the covarion model is same as a noncovarion model, the estimated edge lengths are different (see table 4), which can be used in computing divergence times among the lineages (Peterson and Butterfield 2005; Roger and Hug 2006). Finally, the covarion models implemented in PROCOV can be used to study functional shifts in protein families. It would be interesting to compare the distributions of site likelihoods between the RAS models and covarion models and between the Huelsenbeck and Galtier models. These differences may be combined with 3-dimensional structures of the proteins to study the coevolving amino acid residues, which could aid in understanding the molecular adaptation of the proteins.

Supplementary Material

Supplementary table S1 is available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

Thanks are due to Karen Li for assembling the majority of the protein data sets tested. We also thank Nicolas Galtier and Joe Felsenstein for help in coding PROCOV and Jim Leebens-Mack for sending us the chloroplast data. We appreciate the helpful comments of Stéphane Guindon and the referees. H.C.W. and M.S. were supported by postdoctoral funding from the Genome Atlantic/Genome Canada Large-scale Project "Towards a comprehensive understanding of prokaryotic evolution and diversity: from genomics to metagenomics." This research was supported by Discovery grants awarded to E.S. and A.J.R. by the Natural Sciences and Engineering Research Council of Canada. A.J.R. and E.S. are fellows of the Canadian Institute for Advanced Research Program in Evolutionary Biology. A.J.R. is supported by a fellowship from the Alfred P. Sloan foundation and the Peter Lougheed New Investigator Award from the Canadian Institutes of Health Research and the Peter Lougheed Medical Research Foundation.

Literature Cited

Allman ES, Rhodes JA. 2006. The identifiability of tree topology for phylogenetic models, including covarion and mixture models. J Comput Biol. 13:1101–1113.

Ané C, Burleigh J, McMahon M, Sanderson M. 2005. Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol. 22:914–924.

Blouin C, Boucher Y, Roger AJ. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. Nucleic Acids Res. 31: 790–797.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 17:540–552.

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

Fitch W, Markowitz E. 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. 4:479–593.

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

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

Galtier N, Jean-Marie A. 2004. Markov-modulated Markov chains and the covarion process of molecular evolution. J Comput Biol. 11:727–733.

Gaucher E, Gu X, Miyamoto M, Benner S. 2002. Predicting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem Sci. 27:315–321.

Gaucher E, Miyamoto M, Benner S. 2001. Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc Natl Acad Sci USA. 98:548–552.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2003. Analysis of the Amborella trichopoda chloroplast genome sequence suggests that Amborella is not a basal angiosperm. Mol Biol Evol. 20:1499–1505.

Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2004. The chloroplast genome of Nymphaea alba: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol Biol Evol. 21:1445–1454.

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

Huelsenbeck JP. 2002. Testing a covariotide model of DNA substitution. Mol Biol Evol. 19:698–707.

Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 17:754–755.

Inagaki Y, Blouin C, Susko E, Roger AJ. 2003. Assessing functional divergence in EF-1alpha and its paralogs in eukaryotes and archaebacteria. Nucleic Acids Res. 31:4227–4237.

Inagaki Y, Susko E, Fast NM, Roger AJ. 2004. Covarion shifts cause a long-branch attraction artifact that unites microsporidia and archaebacteria in EF1-alpha phylogenies. Mol Biol Evol. 21:1340–1349.

Jones D, Taylor W, Thornton J. 1992. The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci. 8:275–282.

Knudsen B, Miyamoto M. 2001. A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc Natl Acad Sci USA. 98:14512–14517.

Kolaczkowski B, Thornton J. 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature. 431:980–984.

Leebens-Mack J, Raubeson L, Cui L, Kuehl J, Fourcade M, Chumley T, Boore J, Jansen R, dePamphilis C. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Mol Biol Evol. 22:1948–1963.

Lockhart P, Huson D, Maier U, Fraunholz M, Van De Peer Y, Barbrook A, Howe C, Steel M. 2000. How molecules evolve in eubacteria. Mol Biol Evol. 17:835–838.

Lockhart P, Novis P, Milligan B, Riden J, Rambaut A, Larkum T. 2006. Heterotachy and tree building: a case study with plastids and eubacteria. Mol Biol Evol. 23:40–45.

Lockhart P, Penny D. 2005. The place of Amborella within the radiation of angiosperms. Trends Plant Sci. 10:201–202.

Lockhart P, Steel M. 2005. A tale of two processes. Syst Biol. 54:948–951.

Lockhart PJ, Steel MA, Barbrook AC, Huson DH, Charleston MA, Howe CJ. 1998. A covariotide model explains apparent phylogenetic structure of oxygenic photosynthetic lineages. Mol Biol Evol. 15:1183–1188.

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

Martin W, Deusch O, Stawski N, Grunheit N, Goremykin V. 2005. Chloroplast genome phylogenetics: why we need independent approaches to plant molecular evolution. Trends Plant Sci. 10:203–209.

Misof B, Anderson C, Buckley T, Erpenbeck D, Rickert A, Misof K. 2002. An empirical analysis of mt 16S rRNA covarion-like evolution in insects: site-specific rate variation is clustered and frequently detected. J Mol Evol. 55:460–469.

Miyamoto M, Fitch W. 1995. Testing the covarion hypothesis of molecular evolution. Mol Biol Evol. 12:503–513.

Naylor G, Gerstein M. 2000. Measuring shifts in function and evolutionary opportunity using variability profiles: a case study of the globins. J Mol Evol. 51:223–233.

Penny D, McComish B, Charleston M, Hendy M. 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol. 53:711–723.

Peterson K, Butterfield N. 2005. Origin of the Eumetazoa: testing ecological predictions of molecular clocks against the Proterozoic fossil record. Proc Natl Acad Sci USA. 102:9547–9552.

Philippe H, Casane D, Gribaldo S, Lopez P, Meunier J. 2003. Heterotachy and functional shift in protein evolution. IUBMB Life. 55:257–265.

Pupko T, Galtier N. 2002. A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc R Soc Lond Ser B Biol Sci. 269:1313–1316.

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

Roger AJ, Hug LA. 2006. The origin and diversification of eukaryotes: problems with molecular phylogenetics and molecular clock estimations. Philos Trans R Soc Lond B Biol Sci. 361:1039–1054.

Ruiz-Trillo I, Inagaki Y, Davis LA, Sperstad S, Landfald B, Roger AJ. 2004. Capsaspora owczarzaki is an independent opisthokont lineage. Curr Biol. 14:R946–R947.

Schalchian-Tabrizi K, Skanseng M, Ronquist F, Klaveness D, Bachvaroff TR, Delwiche CF, Botnen A, Tengs T, Jakobsen KS. 2006. Heterotachy processes in rhodophyte-derived second-hand plastid genes: implications for addressing the origin and evolution of dinoflagellate plastids. Mol Biol Evol. 23:1504–1515.

Self S, Liang K. 1987. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under non-standard conditions. J Am Stat Assoc. 82:605–610.

Shimodaira H, Hasegawa M. 2001. CONSEL: for assessing the confidence of phylogenetic tree selection. Bioinformatics. 17:1246–1247.

Simon C, Nigro L, Sullivan J, Holsinger K, Martin A, Grapputo A, Franke A, McIntosh C. 1996. Large differences in substitutional pattern and evolutionary rate of 12S ribosomal RNA genes. Mol Biol Evol. 13:923–932.

Soltis D, Albert V, Savolainen V, et al. (11 co-authors). 2004. Genome-scale data, angiosperm relationships, and "ending incongruence": a cautionary tale in phylogenetics. Trends Plant Sci. 9:477–483.

Spencer M, Susko E, Roger AJ. 2005. Likelihood, parsimony, and heterogeneous evolution. Mol Biol Evol. 22:1161–1164.

Susko E, Inagaki Y, Roger A. 2004. On inconsistency of the neighbor-joining, least squares, and minimum evolution estimation when substitution processes are incorrectly modeled. Mol Biol Evol. 21:1629–1642.

Tuffley C, Steel MA. 1998. Modelling the covarion hypothesis of nucleotide substitution. Math Biosci. 147:63–91.

Xu W. 2002. Covariotide models in phylogenetic analysis [M.Sc. thesis]. Dalhousie University.

Yang Z. 1994. Maximum-likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 39:306–311.

Peter Lockhart, Associate Editor

Accepted October 17, 2006