J Mol Evol (2008) 66:50–60
DOI 10.1007/s00239-007-9062-4

Topological Estimation Biases with Covarion Evolution

Huai-Chun Wang · Edward Susko · Matthew Spencer · Andrew J. Roger

Received: 1 August 2007 / Accepted: 19 November 2007 / Published online: 14 December 2007

Abstract Covarion processes allow changes in evolutionary rates at sites along the branches of a phylogenetic tree. Covarion-like evolution is increasingly recognized as an important mode of protein evolution. Several recent reports suggest that maximum likelihood estimation employing covarion models may support different optimal topologies than estimation using standard rates-across-sites (RAS) models. However, it remains to be demonstrated that ignoring covarion evolution will generally result in topological misestimation. In this study we performed analytical and theoretical studies of limiting distances under the covarion model and four-taxon tree simulations to investigate the extent to which the covarion process impacts on phylogenetic estimation. In particular, we assessed the limits of an RAS model-based maximum likelihood method to recover the phylogenies when the sequence data were simulated under the covarion processes. We find that, when ignored, covarion processes can induce systematic errors in phylogeny reconstruction. Surprisingly, when sequences are evolved under a covarion process but an RAS model is used for estimation, we find that a long branch repel bias occurs.

H.-C. Wang (✉) · E. Susko
Department of Mathematics and Statistics, Dalhousie University, B3H 4H7 Halifax, Nova Scotia, Canada
e-mail: hcwang@mathstat.dal.ca

H.-C. Wang · A. J. Roger
Department of Biochemistry and Molecular Biology, Dalhousie University, B3H 4H7 Halifax, Nova Scotia, Canada

M. Spencer
School of Biological Sciences, University of Liverpool, Liverpool, UK

Keywords Phylogeny estimation · Maximum likelihood · Simulation · Bias · Covarion · Heterotachy · Inconsistency · Protein evolution

Introduction

Phylogenetic inference with maximum likelihood (ML) methods relies on the correct specification of the molecular evolutionary process. Over the years, a variety of models of nucleotide and amino acid substitutions has been proposed to describe this process, including equal rates across sites and lineages; rate variation across sites (Uzzell and Corbin 1971); the proportion and/or distribution of (in)variable sites (Lockhart et al. 1998, 2000); rate variation across lineages and subtrees, such as covarion models (Fitch and Markowitz 1970) and heterotachy (Lopez et al. 2002); compositional heterogeneity (Lockhart et al. 1994; Galtier and Gouy 1995; Foster 2004); and site-heterogeneous amino acid replacement (Lartillot and Philippe 2004; Pagel and Meade 2004). Model misspecification is often cited as one of the primary causes of incorrect topology estimation by ML (Gaut and Lewis 1995; Lockhart et al. 1996, 2006; Bruno and Halpern 1999; Inagaki et al. 2004) and may also cause the method to become inconsistent (i.e., to converge to an incorrect tree with increasing certainty as more sequence data are used for estimation [Felsenstein 1978; Huelsenbeck 1998; Susko et al. 2004]).

Rate variation across sites (RAS) can be modeled with a discrete gamma distribution, which greatly improves the performance of ML methods (Yang 1994). However, the covarion model (and heterotachy in general) has recently attracted increasing attention, due to interest in reconstructing the deep structure of the tree of life. The covarion evolutionary process holds that selective pressures on an

amino acid or nucleotide site are dependent on the states of other sites. As these states change over time, the evolutionary rate at the site of interest also changes. The result is that along the branches of a phylogenetic tree, the rates at different sites may vary in different ways (Fitch and Markowitz 1970). In the covarion hypothesis, characters in a DNA or protein molecule are separable into variable and invariable classes, and the memberships of these two classes change over time, due to changes in functional and selective constraints (Fitch 1971). More specifically, the covarion model proposes the existence of three different categories of sites: the covarion pool of currently variable sites, the class of temporarily invariable sites that are potentially to become variable, and the class of permanently invariable sites (Fitch 1971; Miyamoto and Fitch 1995). The first evidence of covarion-like evolution was therefore based on detecting sites in homologous sequences that are invariable among taxa in one clade, but variable among taxa in another clade (Fitch and Markowitz 1970; Fitch 1971; Miyamoto and Fitch 1995; Gu, 1999; Gaucher et al. 2001; Pupko and Galtier 2002). This is also the basis for several statistical tests used to detect covarions (Lockhart et al. 1998, 2000; Ané et al. 2005).

Inspired by these qualitative ideas, Tuffley and Steel (1998) developed the first mathematical model of covarion evolution (the TS model; see also Penny et al. 2001). They hypothesized that the substitution process at a site can be turned *on* (variable) or *off* (invariable). When a site is *on* it evolves according to some substitution process and can be modeled with a reversible substitution rate matrix. The switching between *on* and *off* is modeled as an additional stationary Markov process. Huelsenbeck (2002) implemented a version of this model, with the addition of RAS, under the Bayesian framework for phylogenetic inference. Galtier (2001) developed a different covarion model with an arbitrary number of rate classes. In his model, the overall substitution rate multipliers are defined by a discrete gamma distribution, similar to models of rate variation across sites, except that sites may change rates. Rate switching is modeled by a Poisson process. However, the Galtier model does not allow rate switching to and from an invariable *off* state. We recently proposed a general covarion model that not only allows site rates to switch from *on* to *off* and from *off* to *on*, but also allows switching between different rates among the *on* states (Wang et al. 2007). Empirical studies on ribosomal RNA genes, protein coding genes, and protein sequences have demonstrated that the covarion models provided better fits to the majority of the data sets than the RAS models that do not allow rates at sites to change over time (Galtier 2001; Huelsenbeck 2002; Wang et al. 2007). Furthermore, several recent studies have shown that phylogenetic inference using the covarion models can support different optimal topologies from that using a model without covarions (e.g., the RAS model) (Ruiz-Trillo et al. 2004; Shalchian-Tabrizi et al. 2006; Wang et al. 2007). However, it is not clear that failing to account for covarion-like evolution will generally result in topological misestimation and the inconsistency of the phylogenetic methods.

In this study we performed analytical and theoretical studies of limiting distances under the TS model (Tuffley and Steel 1998) to investigate the extent to which the covarion process impacts on phylogenetic estimation. We then did four-taxon tree simulations to assess the ability of an RAS model-based ML method to recover the phylogenies when the sequence data were simulated under the TS model and the more complex general covarion model (Wang et al. 2007), respectively. We compared the effects of different sequence lengths and amino acid substitution matrices on the simulation results.

Analytical Results

We start by considering results in an idealized four-taxon setting: amino acid Jukes-Cantor (JC; 1969) distances with the neighbor-joining algorithm for tree estimation. While the setting is simplified to make analysis more tractable, we show that similar behavior arises with more complex substitution processes and the ML method. What we show here is that distances that are uncorrected for RAS variation will cause a long branch attraction (LBA) bias, while distances that are corrected for RAS will show a long branch repel (LBR) bias.

Our approach is similar to that described by Susko et al. (2004). With or without a gamma RAS adjustment, JC distances between a pair of taxa, i and j, are a continuous function of the proportion of sites with different amino acids in the sequences, $\hat{p}^{(i,j)}$; we denote the distance $d_{ij} = d(\hat{p}^{(i,j)})$. Since the proportion of sites with different amino acids converges to the probability of different amino acids for i and j at a site, $p^{(i,j)}$, we have that

$$d(\hat{p}^{(i,j)}) \rightarrow d(p^{(i,j)})$$

as the number of sites goes to infinity. It will be valuable to think of incorrectly specified distances in terms of their dependence on the true evolutionary distances. The probability of different amino acids at a site is dependent on the pair, i and j, only through the true evolutionary distance, t, between the pair: $p^{(i,j)} = w(t)$. Thus the limiting distance, $d(p^{(i,j)})$, is also a function $g(t)$: $= d(w(t))$ of the true evolutionary distance between the pair.

In the case of a four-taxon tree with taxa A, B, C, and D there are three topologies which can be described in terms of the neighbor of A: (A, B), (A, C), and (A, D). We assume throughout that the true topology is (A, B). With $b > a$ in Fig. 1, either the true tree will have long branches separate (Fig. 1A), which we will an LB-separate tree, or the true

tree will have long branches together (Fig. 1B), which we call an LB-together tree. For a four-taxon tree the neighbor joining algorithm can be shown (Saitou and Nei 1987) to determine the estimated topology according to the following rules:

1. *(A,B)* is preferred to *(A,D)* if

$$d_{AD} + d_{BC} - d_{CD} - d_{AB} > 0 \quad (1)$$

2. *(A,B)* is preferred to *(A,C)* if

$$d_{AC} + d_{BD} - d_{CD} - d_{AB} > 0 \quad (2)$$

The limiting behavior differs depending on whether the limiting distances, $g(t)$, are concave functions of the true distances, t, or not. We start by considering the case that $g(t)$ is concave and the generating tree has long branches separate. As discussed by Susko et al. (2004), with a large number of sites, (1) will for sure be satisfied so that the estimated topology will be (*A, B*) or (*A, C*). Let $b(a)$ be the solution of

$$g(2b + a) - 2g(a + b) + g(3a) = 0 \quad (3)$$

Then, as discussed by Susko et al. (2004), with a large number of sites, (2) will be satisfied, if and only if $b < b(a)$. In other words, for $b > b(a)$, the tree with long branches together, (*A, C*), will be estimated.

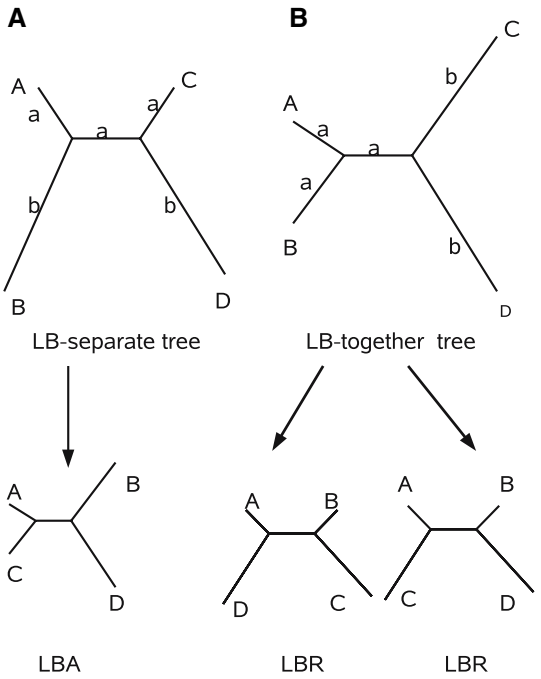

Fig. 1 The two types of four-taxon trees used for the analytical and simulation studies. **A** Tree of the LB-separate type, which can induce the LBA form of bias if the misestimated tree is an AC tree. It can also induce an AD tree (not shown) which does not represent an LBA. **B** Tree of the LB-together type, which may induce the AC and AD trees, both of which are of the LBR form of bias

Consider now a generating JC amino acid model combined with the covarion model described by Tuffley and Steel (1998). The TS model assumes a Markov process for rate switching along the edges of a phylogenetic tree. Rates along an edge switch from an *off* state to an *on* state and from *on* to *off*. The model has two parameters: s_{01} and s_{10}, the rate of transition from *off* to *on* and then the corresponding rate from *on* to *off*, respectively.

Explicit mathematical equations for the relationships between the limiting distance and the true distance (i.e., the function $g(t)$) under the covarion model are complicated, but they can be computed numerically. Figure 2A gives the estimated limiting distances plotted against the true distances for various choices of s_{01} and s_{10} when JC distances are used that make no adjustment for the TS model or even RAS. The concave shapes of the plots indicate that an LBA form of inconsistency will arise. The zones of inconsistency (defined by the function $b(a)$) are given in Fig. 2B. Values of b and a above and to the left of the lines correspond to regions where the topology with long branches together, the (*A, C*) topology, will be estimated. In some respects, the results are not surprising. Tuffley and Steel (1998) show that for a pair of taxa, the TS model is

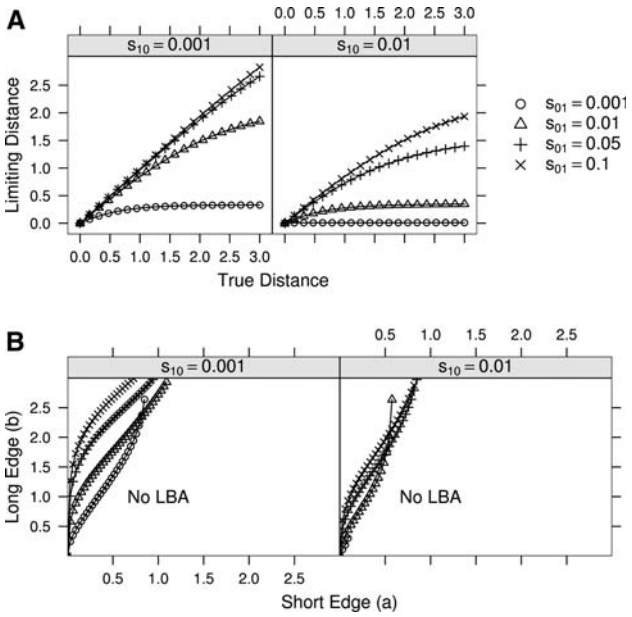

Fig. 2 Analytical results for a covarion generating process and equal rates distance estimation. **A** The relationship between the limiting estimated distances and the true distances. With the generating tree that has long branches separate, the TS + JC model was the generating process, but estimated distances are JC with no correction for RAS or a covarion process. **B** The zones of inconsistency. Edge lengths have been multiplied by $s_{01}/(s_{01} + s_{10})$ so that they are interpretable as expected numbers of substitutions. All values of a and b above and to the left of the boundary curves correspond to cases where the tree with long branches together will be estimated with long sequences

indistinguishable from an RAS model and it is well known that a failure to adjust for the RAS variation leads to LBA (Huelsenbeck 1995; Kuhner and Felsenstein 1995; Susko et al. 2004). For $s_{10} = 0.001$ and $s_{01} = 0.1$ the limiting estimated distances are almost-linear functions of the true evolutionary distances, giving an almost nonexistent zone of inconsistency. This too is not surprising when one considers that the proportion of time that the covarion process is *on*, $s_{01}/(s_{01} + s_{10})$, is almost 1; at any given site it is likely that there is little or no rate variation.

The case where the limiting distances are convex is different. As discussed by Susko et al. (2004), in this case with the generating tree that has long branches together, the trees with long branches separate, (*A, C*) and (*A, D*), will be estimated with probability approximately ½ each in the zone of inconsistency. This zone corresponds to values of $b > b(a)$ where $b(a)$ is the solution of

$$2g(b + 2a) - g(2a) - g(2b) = 0 \quad (4)$$

Figure 3A plots the limiting distances against true distances for a number of different choices of the α parameter in gamma-corrected RAS distances. The other parameters in the generating model were set to values estimated in real data (the HSP90 data considered by Wang et al. 2007): $s_{01} = 0.61$ and $s_{10} = 0.53$. With α small we see highly convex $g(t)$. In this case, with a generating LB-together tree, the trees with long branches separate will be estimated, resulting in an LBR form of inconsistency. The zones of inconsistency are plotted in Fig. 3B, which shows large zones when α is small. As α gets larger, the relationship between estimated and true distances gets closer to linear, with a corresponding small zone of inconsistency and, as expected given the results illustrated in Fig. 2, with no RAS variation ($\alpha \to \infty$), the relationship between estimated and true distance is concave, with no corresponding LBR zone of inconsistency, for the LB-together generating tree that has long branches together. In this case, because the shape of the curve is concave but close to linear, a small LBA zone of inconsistency will result.

The situation in usual practice is more complex. ML estimation is frequently used in practice rather than distance methods, empirical substitution models like the JTT model (Jones et al. 1992) are used and the value of α is estimated. Still, the results here suggest that in the common setting where an RAS adjustment is made, the likely consequence is an LBR form of bias. In the following sections we investigate this further through simulation.

Simulations

A sequence simulator program (Seq-gen-aminocov), modified from Seq-gen (Ané et al 2005; Rambaut and Grassly

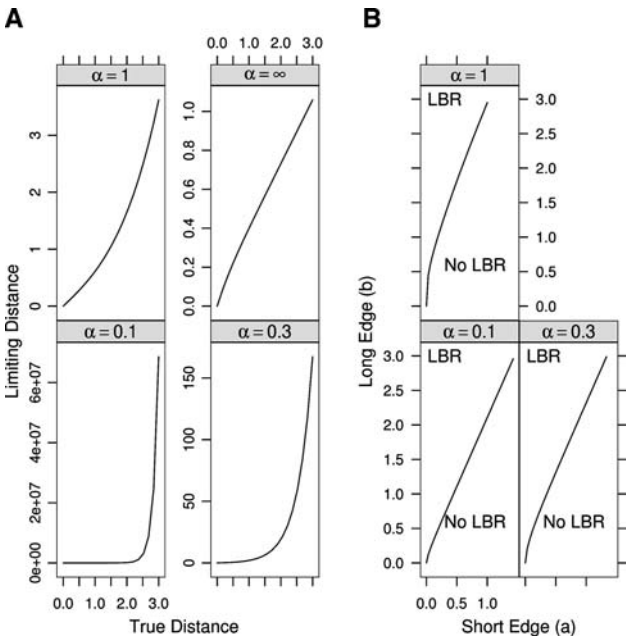

Fig. 3 Analytical results for a covarion generating process and RAS distance estimation. **A** The relationship between the limiting estimated distances and the true distances. With the generating tree that has long branches together, the JC + TS model ($s_{01} = 0.61$ and $s_{10} = 0.53$) is the generating model, but estimated distances are derived from the JC + Γ model for various choices of a fixed α shape parameter. B The zones of inconsistency. Edge lengths have been multiplied by $s_{01}/(s_{01} + s_{10})$ so that they are interpretable as expected numbers of substitutions. All values of *a* and *b* above and to the left of the boundary curves correspond to cases where a tree with long branches separate will be estimated with long sequences

1997), was developed for various covarion models. It first simulates switching between rate categories as a continuous-time Markov process. It then rescales the edge lengths so that the expected number of substitutions over the edge matches the required value. Finally, it simulates amino acid substitutions as another continuous-time Markov process, using the rescaled edge lengths. The rescaling is done as follows. For the TS and Huelsenbeck models, it divides the edge length by the time spent in the *on* state (i.e., $s_{01}/(s_{01} + s_{10})$). For the Galtier model, there is no *off* state. The rates switch among a set of equiprobable categories with mean of one substitution per unit time. The expected number of substitutions is thus unchanged. Therefore, no rescaling is necessary. For the general model, sites with a fixed rate need no rescaling. For covarion sites, the rescaling is done by dividing by the time spent in the *on* state, in the same way as in the TS and Huelsenbeck models. The source code of the program is available at http://www.liv.ac.uk/~matts/.

The two types (LB-separate and LB-together) of four-taxon trees (Fig. 1) were used to simulate protein sequence data. The edge lengths *a* and *b* varied from 0.05, 0.3, 0.6, 0.9, ..., 2.4, 2.7, 3.0. In total there are 121 trees corresponding to the various combinations of the *a* and *b*

settings. One hundred simulated data sets were generated for each setting. Seq-gen-aminocov was then used to simulate amino acid sequence data under the given models and trees. The simulated sequence lengths include 100, 459, 1000, 10,000, and 100,000 amino acids for different simulation experiments. The amino acid substitution models include uniform rates (i.e., the JC amino acid exchange matrix) and the JTT rate matrix. Three site-rate variation models were used in this study: the RAS model, the TS model, and the general covarion model. For the TS model we simulated the sequences with $s_{01} = 0.61$ and $s_{10} = 0.53$, which correspond to the equilibrium frequency of *on* sites (π) = 0.53 and switching speed per substitution (v) = 0.57 in the Seq-gen-aminocov parameterization. The general covarion model (Wang et al. 2007) combines the RAS model with both the TS and the Galtier models, allowing evolutionary rates of sequence sites not only to switch from *on* to *off* and from *off* to *on*, as in the TS and Huelsenbeck (2002) models, but also to switch among different *on* states, as in the Galtier (2001) model. The general model has three more parameters than the TS model, including the rate of switching from one nonzero rate to another nonzero rate (s_{11}), the proportion of covarion sites (π) ($1 - \pi$ is the proportion of sites evolving according to noncovarion RAS process), and the gamma shape parameter (α) for the RAS process, in addition to s_{01} and s_{10}.

Topology estimations were conducted with PAML version 3.12 (Yang 1997) under a discrete gamma RAS model and a uniform rate or with Tree-Puzzle version 5.2 (Schmidt et al. 2002) under the RAS model and JTT rates. Heatmaps, plotted with a script written in R (R Development Core Team 2007), were used to show the distribution of the estimated optimal topologies for different a and b settings.

The Effect of the TS Covarion Model on Phylogenetic Inference

Simulating Under the LB-Separate Tree

We simulated five data sets of different lengths (100, 459, 1000, 10,000, and 100,000 amino acids) under the TS model and amino acid JC model. The covarion parameters for the simulations were $s_{01} = 0.61$, $s_{10} = 0.53$. The estimations were conducted with PAML under the RAS model (with four gamma rate categories and allowing α to be optimized), the JC rate, and allowing edge lengths to be optimized. In this setting, the only "misspecification" of the model is the TS versus RAS process.

The results indicate that when the sequence lengths are very short (100 amino acids) quite a number of both AC and AD trees are estimated (470 and 405, respectively, of 12,100 trees). As the length increases, the numbers of both types of the misestimations decrease dramatically. For instance, at the length of 1000 amino acids the numbers of the misestimated AC and AD trees are only 12 and 41, respectively, and they become virtually zero when the lengths are 10,000 or 100,000 amino acids. It should be mentioned that for the generating tree being of the LB-separate form, the misestimated AC trees, but not the AD trees, represent an LBA bias when a is small and b is much greater than a (Fig. 1A). In order to see any estimation bias in these a and b settings that can potentially induce LBA artifacts, we computed average frequencies of the estimated AB, AC and AD trees among the cells in the region where $[b > a, a \leq 1.0]$ and their standard errors (Table 1). For the frequency data in the defined region, the Monte Carlo standard error of the proportion of the AB trees is obtained as

$$\frac{1}{c}\sqrt{\sum\nolimits_{a,b}\hat{P}_{AB}^{(a,b)}\left(1-\hat{P}_{AB}^{(a,b)}\right)}$$

where c is the number of a, b settings, the sum is over all a, b settings, and $\hat{P}_{AB}^{(a,b)}$ is the proportion of the AB trees in the cell. The standard errors of the proportions of the AC and AD trees were calculated in the same way.

Table 1 shows that even in this region where the LBA artifact is potentially plausible, the proportions of the AC trees are similar to or less than that of the AD trees and both types of the misestimations decrease dramatically as the simulated sequences get longer, and when the sequences are over 10,000 amino acids there is no misestimation. Therefore, it appears there is an absence of LBA bias for these simulation settings, i.e., data generated under the TS model and estimated under the RAS model.

Table 1 Proportions ± standard errors of the estimated AB, AC, and AD trees in the regions of $[b > a, a \leq 1.0]$ for simulations under TS + JC and estimation under RAS + JC for different lengths of the simulated sequences: simulating trees are of the LB-separate form

Sequence length	Simulation model	Estimation model	AB tree	AC tree	AD tree
100	TS + JC	RAS + JC	0.83 ± 0.006	0.09 ± 0.005	0.08 ± 0.004
459	TS + JC	RAS + JC	0.95 ± 0.003	0.01 ± 0.002	0.04 ± 0.003
1000	TS + JC	RAS + JC	0.98 ± 0.002	0.004 ± 0.001	0.01 ± 0.002
10,000	TS + JC	RAS + JC	1.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
100,000	TS + JC	RAS + JC	1.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Figure 4 shows three vertical panels of heatmaps that represent, respectively, the proportions of the misestimated AC and AD trees with regard to the edge lengths (a and b) for sequence lengths of 100, 459, and 1000 amino acids, respectively. The left two vertical maps are the distribution of the proportions of the AD and the AC trees, respectively, for the 100 amino acid data sets. Wrongly assigned AC and AD trees are obtained for all values of a and for $b > 0.05$, indicating that stochastic error is prevalent for analyzing the short sequences (especially in the cases where b is very large or a is very small). The middle two vertical maps are the distribution of the AD and AC trees for sequences of 459 amino acids. The misestimated optimal AC and AD trees are only present in $a = 0.05$ and $b > 0.6$. The right two vertical maps show the distribution for the AD and AC trees for the sequences of 1000 amino acids; the misestimated optimal AC and AD trees are only present in $a = 0.05$ and $b > 0.9$. For length = 10,000 and 100,000 amino acids, misestimated optimal AD and AC trees are present in only 1 simulation ($a = 0.05$ and $b = 3.0$) of a total of 12,100 simulations (heatmaps not shown; see Table 1).

The estimation of the tree topologies allowed the optimization of the gamma shape parameter (α). Figure 5 shows a heatmap of the estimated α values averaged for each cell of the estimated AB trees for sequence length of 1000 amino acids. This shows that α is small when both a (<1.0) and b (<1.5) are small. The reason is likely that if an edge is short, a site with a high rate at the start of the edge will probably still have a high rate at the end of the edge. Similarly, a low initial rate will also be maintained over a short edge. Consequently, over a short time, the variance in average rates across sites will be high, and a smaller α is expected. Heatmaps of α for the other sequence lengths (not shown) show a similar distribution of α, but the average values are smaller as sequence lengths increase.

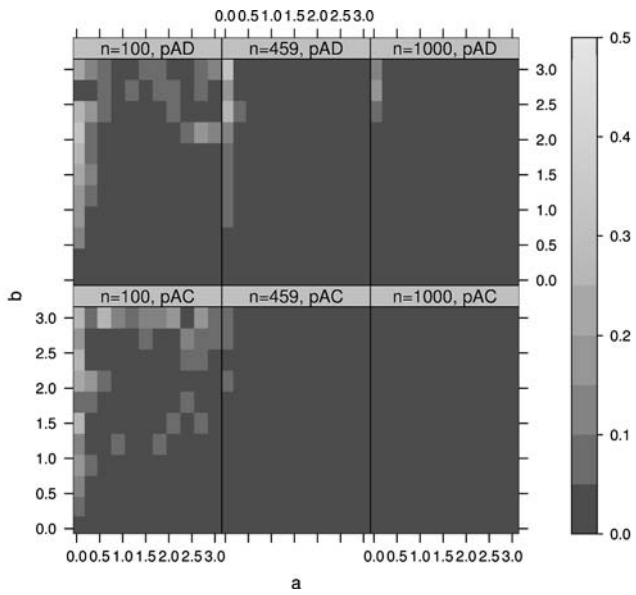

Fig. 4 Heatmaps of the proportions of misestimated AC and AD trees for different edge lengths a and b for simulations under LB-separate trees and for sequence lengths of 100, 459, and 1000 amino acids. Each vertical panel contains two heatmaps for the proportions of the AD and AC trees, respectively

Simulating Under the LB-Together Tree

For the generating tree being LB-together and simulations under the TS + JC model and estimated under the RAS + JC model, the numbers of misestimated AC and AD trees are much higher than for the simulations under the LB-separate tree. For instance, at the length of 100,000 amino acids the numbers of misestimated AC and AD trees for the current simulations are 514 and 492, respectively. Table 2 shows the proportions and standard errors of the estimated AB, AC, and AD trees for the current simulations (the simulating trees being of the LB-together form) among cells in the region where $[b > a, a \leq 1.0]$.

As shown in Fig. 1B, both misestimated AC and AD trees represent the LBR bias for simulations under the LB-together trees, which is supported by the comparable proportions of the AC and AD trees within the same sequence length settings (Table 2). Comparing Table 2 with Table 1 shows that the proportions of the AC and AD trees are much higher in Table 2. It also indicates that the proportions of the misestimated AC and AD trees, though showing a slight decrease with sequence length, are not

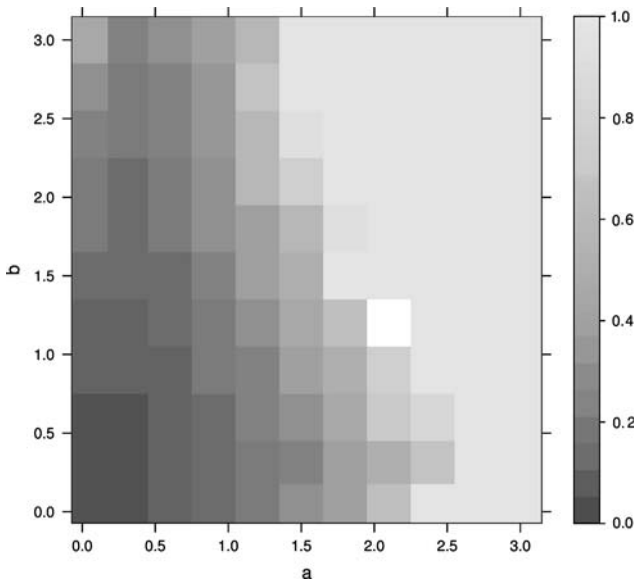

Fig. 5 Heatmap for the distribution of the estimated α shape parameter for the estimated optimal AB trees for simulations under LB-separate trees and a sequence length of 1000 amino acids

Table 2 Proportions ± standard errors of the estimated AB, AC, and AD trees in the regions of [$b > a$, a ≤ 1.0] for simulations under TS + JC and estimation under RAS + JC for different lengths of the simulated sequences: simulating trees are of the LB-together form

Sequence length	Simulation model	Estimation model	AB tree	AC tree	AD tree
100	TS + JC	RAS + JC	0.66 ± 0.007	0.18 ± 0.006	0.16 ± 0.006
459	TS + JC	RAS + JC	0.62 ± 0.007	0.20 ± 0.006	0.18 ± 0.006
1000	TS + JC	RAS + JC	0.66 ± 0.005	0.17 ± 0.006	0.17 ± 0.006
10,000	TS + JC	RAS + JC	0.69 ± 0.003	0.17 ± 0.005	0.14 ± 0.005
100,000	TS + JC	RAS + JC	0.71 ± 0.001	0.15 ± 0.005	0.14 ± 0.005

much different across the different sequence lengths in Table 2, in sharp contrast to the dramatic decrease in the proportions of the misestimations shown in Table 1. However, the proportion of the misestimations is an average over different edge length settings and its change with increasing sequence lengths can be very different for particular edge length settings, which is revealed in the following heatmaps.

Figure 6 shows heatmaps for the proportions of the misestimated AC + AD trees with regard to a and b for the simulations under different sequence lengths. The region where poor estimation occurs gradually contracts as the sequence length increases but the proportions of misestimations increase to 1 in these regions. For instance, starting with short sequences of 100 amino acids, wrongly assigned AC and AD trees are distributed mainly in the region composed of $a < 1.5$ and $b > a$. For length = 459 amino acids, the distribution of AC + AD trees is more restricted to the upper left corner. The region in which AC + AD trees are estimated continues to shrink as sequence length increases. When the sequence length is 100,000 amino acids, AC and AD trees are only estimated in the region where $a = 0.05$ or 0.3 and $b > 0.9$. However, within these regions, the frequency of misestimation increases with increasing sequence length, approaching 100% when the sequence length is 100,000.

For the simulated sequences of length = 1000 or 459 amino acids, the estimated average α parameters for the estimated optimal AB trees are relatively small (mean $\alpha < 2.0$) when $a < 1.0$, but b can be up to 3.0. For short sequences (sequence = 100 amino acids), $\alpha < 2.0$ only occurs when $a < 0.6$ and $b < 1.5$. Within this range of a and b settings, however, values of $\alpha > 5$ were still estimated.

The above simulations examine bias in phylogenetic inference for sequence data simulated under the TS model + the JC rate and estimated under RAS + the JC rate and demonstrated that the outcomes depend on the types of the simulating trees. If the data are simulated under the LB-separate trees, there is no LBA bias and the misestimations due to stochastic errors will be reduced with increasing sequence lengths and they essentially disappear when the sequence length reaches 10,000 amino acids. This suggests that data generated under a process similar to the TS covarion model can be handled with an RAS model for long sequences. However, if the data are simulated under the LB-together trees, both AC and AD trees when estimated represent LBR bias, and they significantly persist even when the sequence length reaches 100,000 amino acids, suggesting that the covarion process will likely cause an LBR bias of phylogenetic inference if the RAS method is used for estimation.

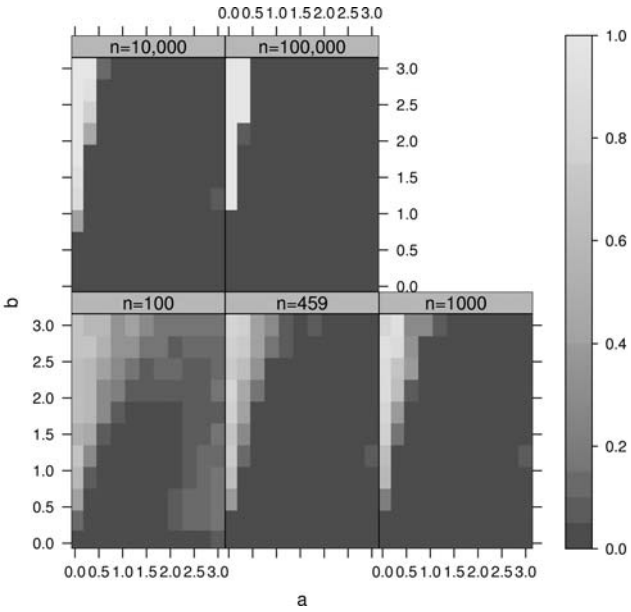

Fig. 6 Heatmaps of the proportions of misestimated AC and AD trees combined for different edge lengths a and b for simulations under LB-together trees and sequence lengths of 100, 459, 1000, 10,000 and 100,000 amino acids, respectively

Comparing the General Covarion and RAS Models

The foregoing analyses concerned estimation biases incurred by the simplest covarion model, the TS process. However, it is of interest to know whether these results generalize to more complex models, such as the general covarion process described by Wang et al. (2007). Four simulation experiments were conducted: Simulations I and II were performed over the LB-separate trees and

Table 3 Proportions ± standard errors of the estimated AB, AC, and AD trees in the regions of [$b \geq a$, $a \leq 1.0$] for the four simulation experiments indicated in the text: sequences are all 459 amino acids long in these simulations

Data set	Simulation model	Estimation model	AB tree	AC tree	AD tree
I[a]	RAS + JTT	RAS + JTT	0.96 ± 0.003	0.03 ± 0.002	0.02 ± 0.002
II[a]	General + JTT	RAS + JTT	0.95 ± 0.005	0.01 ± 0.002	0.04 ± 0.004
III[b]	RAS + JTT	RAS + JTT	0.76 ± 0.008	0.12 ± 0.007	0.12 ± 0.007
IV[b]	General + JTT	RAS + JTT	0.54 ± 0.008	0.22 ± 0.008	0.24 ± 0.008

[a] The simulating tree is of the LB-separate form

[b] The simulating tree is of the LB-together form

Simulations III and IV were performed over the LB-together trees. In all simulations sequence lengths were kept at 459 amino acids. The following four settings were used. (I) Data were simulated under the RAS model with $\alpha = 0.8$, JTT + 4 Γ rates. For each tree, edge lengths a and b vary from 0.05 to 3.0 and 100 data sets were simulated for each setting. (II) Data were simulated under the general covarion model. In addition to using the above parameters for the RAS model, covarion parameters also included a proportion of covarion sites (π) = 0.71, and switching rates of $s_{01} = 0.43$, $s_{10} = 0.57$, $s_{11} = 0.97$. These parameter settings were based on the optimized result for a HSP90 data set that was previously used for testing the covarion models (Wang et al. 2007). (III) The sequences were generated under the RAS model. The simulation conditions are the same as Simulation I except that the LB-together trees were used. (IV) The sequences were generated under the general covarion model. The simulation conditions are the same as Simulation II except that the LB-together trees were used.

Tree-Puzzle was used to estimate the topologies and compute the ML scores for the data sets, with the JTT + Γ model (four rates) and allowing α and edge length optimization. The simulation and estimation conditions are summarized in Table 3. Since the estimated wrong trees (i.e., the AC and AD trees) are restricted to the region where [$b > a$, $a \leq 1.0$], we computed average frequencies of AB, AC, and AD trees among cells in this region and their standard errors (Table 3). When the generating tree is of the LB-separate form and the RAS model is used to generate and estimate the data (setting I), the proportions of AC and AD trees recovered are both small and not significantly different from one another. By contrast, for data set II, where data are simulated under the general covarion model and the same type of the generating trees, the proportion of the estimated AD tree is significantly greater than that of the AC trees. As mentioned above, only the misestimated AC trees represent LBA bias for the generating trees being of the LB-separate form. For data sets III and IV, however, since the generating trees are of the LB-together form, both the misestimated AC and AD trees represent LBR. Table 3 and Fig. 7 show that both the proportions of AC and AD trees are significantly increased in data set IV (simulated under the general model) compared to data set III (simulated under the RAS model). In summary, the general covarion model and estimating trees with an RAS model significantly increased LBR bias compared with the RAS model simulations.

Furthermore, we also simulated sequence data under the TS + JTT models and estimated under the RAS + JTT models for both the LB-separate and the LB-together trees. The results (not shown) again demonstrate that sequence data simulated under the covarion model and estimated under the RAS model cause an LBR bias.

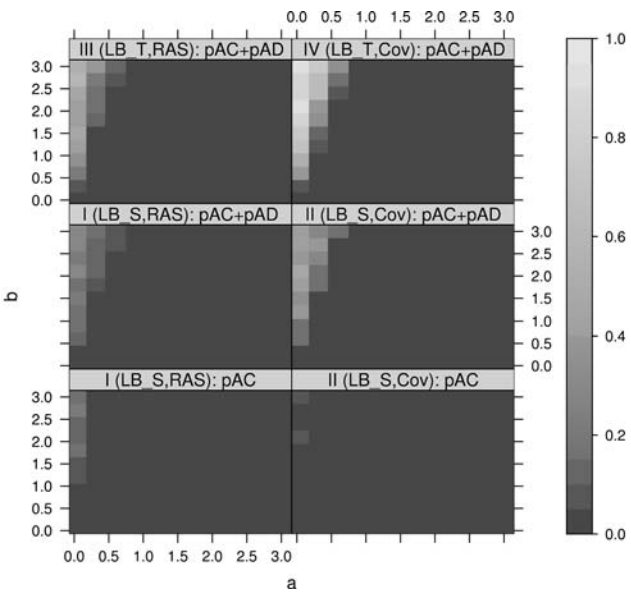

Fig. 7 Heatmaps of the proportions of misestimated AC and AD trees for different edge lengths a and b for simulations under the LB-separate trees (LB−S; the bottom panel for the proportion of the AC trees only; the middle panel for the proportions of the AC and AD trees combined) and under the LB-together trees (LB−T; the top panel for the proportions of the AC and AD trees combined). Data sets I, II, III, and IV are defined as in Table 3. For each horizontal panel, the left map shows the data that were simulated under the RAS + JTT models and the right map shows the data that were simulated under the general covarion + JTT models. The sequence lengths for all simulations were maintained at 459 amino acids. The estimations of the topologies were done using Tree-Puzzle with RAS + JTT models with four gamma rate categories

Does a Covarion Model Perform Better Than the RAS Model?

The above simulations evaluate the performance of the ML tree estimation with an RAS model when the data are simulated under covarion models, which show that an LBR bias persists when the data are simulated under the LB-together trees. It will be interesting to see whether this problem can be relieved if a covarion model is used for tree estimation. It is also of interest to know the effect of tree estimation with a covarion modeling for data simulated under an RAS process. Several software packages, including NHML, MrBayes, and Procov, are now available for implementing the covarion process in phylogenetic reconstruction methods (Galtier 2001; Huelsenbeck 2002; Wang et al. 2007). While a complete evaluation of the covarion methods on tree estimation is beyond the scope of this paper, we used Procov with the general covarion model to analyze the four data sets (459 amino acids long and 100 replicates for each data set) described in Table 3. The numbers of correctly estimated and misestimated trees (i.e., AB trees and AC + AD trees) are shown in Table 4. It also lists the corresponding numbers of the AB trees and AC + AD trees estimated with the RAS model and the *p*-values for χ^2 tests of the three 2 × 2 contingency tables. The results show that both the RAS and the general covarion models perform well for data simulated under the LB-separate trees. The slight increases in the numbers of the AC + AD trees in the covarion estimations can be explained by the large variances in parameter estimations under the general covarion model, which has four more parameters than the RAS model, especially when the sequences are relatively short (459 amino acids in these data). For data simulated under the LB-together trees, there is no significant difference between the uses of the RAS and covarion models for estimating the topology when the data are simulated under the RAS. However, there is a significant improvement using the covarion model, over the RAS model, for tree estimation when the data are simulated under the covarion model. Therefore, the use of the covarion model effectively reduces the LBR bias.

Discussion

Simulated four-taxon datasets have been widely used to produce controlled simulation of evolutionary processes and evaluate the success rate of different methods for recovering phylogenetic trees (Felsenstein 1978; Huelsenbeck 1995, 1998; Gaut and Lewis 1995; Chang 1996; Siddall 1998; Bruno and Halpern 1999; Swofford et al. 2001; Susko et al. 2004). These previous simulation studies sometimes focused on model misspecification in the RAS model and the relative performances of the maximum parsimony (MP) and ML methods, which have successfully identified model misspecification and particular edge length setting that may cause LBA (“Felsenstein zone’’) and LBR (“Farris zone’’) biases. More recently, the studies have shifted to simulate more heterogeneous evolutionary processes, the heterotachous rate variation across sites and lineages (Kolaczkowski and Thornton 2004; Ruano-Rubio and Fares 2007), and the resulting estimation bias for ML. The Kolaczkowski and Thornton study has generated renewed debate about whether the MP or ML methods should be preferred for data that evolved under a heterotachous process (Spencer et al. 2005; Gadagkar and Kumar 2005; Gaucher and Miyamoto 2005; Steel 2005; Philippe et al. 2005).

In this study we investigated the impact of data generated under the standard covarion process on phylogenetic estimation with widely used methods: the equal rates and RAS models. From both our analytical studies and the

Table 4 Numbers of estimated AB and AC + AD trees for the four simulated data sets indicated in Table 3

Data set	Simulation model	Estimation model	No. AB trees	No. AC + AD trees	*p*-value
I[a]	RAS + JTT	RAS + JTT	11,929	171	0.0011
I	RAS + JTT	General + JTT	11,864	236	
II[a]	General + JTT	RAS + JTT	11,989	111	0.002
II	General + JTT	General + JTT	11,928	172	
III[b]	RAS + JTT	RAS + JTT	11,572	528	0.35
III	RAS + JTT	General + JTT	11,542	558	
IV[b]	General + JTT	RAS + JTT	11,048	1052	<0.0001
IV	General + JTT	General + JTT	11,459	641	

Note. For each data set of 12,100 alignments, tree estimations were done with the RAS and general covarion models, respectively and a *p*-value was derived from a χ^2 test of the 2 × 2 contingency table

[a] The simulating tree is of the LB-separate form

[b] The simulating tree is of the LB-together form

simulations we see that, depending on the types of the simulating tree and the phylogenetic model used for estimation, the bias could be LBA or LBR. For sequence data simulated under the LB-separate trees and the TS model, the analytical results indicate that the neighbor-joining algorithm-based distance method for tree estimation will cause an LBA form of inconsistency if a uniform amino acid exchange rate model is used for estimation. The zone of inconsistency varies with the relative rates of s_{10} and s_{01} (Fig. 2B). The simulations show that using ML with RAS adjustment for tree estimation will cause some misestimations, but no apparent LBA bias. Moreover, increasing sequence length will effectively reduce the misestimations, indicating that the RAS method will be consistent under these settings. This supports the argument that part of the covarion process may be accounted for by an overall RAS-like heterogeneity, as every site relative rate would depend on the time interval it has spent in the *on* state: the longer the time, the higher the rate (Ruano-Rubio and Fares 2007). For sequence data simulated under the LB-together tree and estimated with an RAS-based ML method, the LBR bias will persist even when very long sequences are used, indicating inconsistency under these settings. The analytical results show that for the generating tree being of the LB-together form, the limiting distance is a convex function of the true distance for the generated sequences (Fig. 3A) and the zone of inconsistency of the ML estimation depends on the α parameter used for the RAS adjustment (Fig. 3B). Therefore, although the covarion process could cause LBA bias when using an equal-rates model to estimate, an LBR bias is much more of a concern and results in estimation under the RAS model to be inconsistent. It has been noticed elsewhere that covarion-type evolution is not always well explained by the RAS models (Lockhart et al. 1998; Ané et al. 2005).

Phylogeneticists are accustomed to being concerned about LBA bias when an estimated tree contains LB-together. We see here that LB-apart in an estimated tree can be of concern as well. This study has found no evidence of LBA bias but rather a substantial LBR bias when ML estimation under an RAS model is used but the generating model is a covarion model. Furthermore, we found that using a phylogenetic method that implements covarion models can effectively reduce the LBR bias. If the data under examination show evidence of covarion-like evolution, as is often the cases in the inference of deep phylogenies (Inagaki et al. 2004; Lockhart et al. 1998; Ané et al. 2005), it would be advisable to use a covarion model, in addition to the traditional RAS models, to infer phylogenies.

Acknowledgments We thank the two reviewers for useful comments. 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 Peter Lougheed New Investigator Award from the Canadian Institutes of Health Research and the E.W.R. Steacie fellowship from NSERC.

References

Ané C, Burleigh JG, McMahon MM, Sanderson MJ (2005) Covarion structure in plastid genome evolution: a new statistical test. Mol Biol Evol 22:914–924

Bruno WJ, Halpern AL (1999) Topological bias and inconsistency of maximum likelihood using wrong models. Mol Biol Evol 16:564–566

Chang JT (1996) Inconsistency of evolutionary tree topology reconstruction methods when substitution rates vary across characters. Math Biosci 134:189–215

Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:401–410

Fitch WM (1971) Rate of change of concomitantly variable codons. J Mol Evol 1:84–96

Fitch WM, 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

Foster PG (2004) Modeling compositional heterogeneity. Syst Biol 53:485–495

Gadagkar SR, Kumar S (2005) Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous. Mol Biol Evol 22:2139–2141

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

Galtier N, Gouy M (1995) Inferring phylogenies from DNA sequences of unequal base compositions. Proc Natl Acad Sci USA 92:11317–11321

Gaucher EA, Miyamoto MM (2005) A call for likelihood phylogenetics even when the process of sequence evolution is heterogeneous. Mol Phylogenet Evol 37:928–931

Gaucher EA, Miyamoto MM, Benner SA (2001) Function-structure analysis of proteins using covarion-based evolutionary approaches: Elongation factors. Proc Natl Acad Sci USA 98:548–552

Gaut BS, Lewis PO (1995) Success of maximum likelihood phylogeny inference in the four-taxon case. Mol Biol Evol 12:152–162

Gu X (1999) Statistical methods for testing functional divergence after gene duplication. Mol Biol Evol 16:1664–1674

Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44:17–48

Huelsenbeck JP (1998) Systematic bias in phylogenetic analysis: Is the Strepsiptera problem solved? Syst Biol 47:519–537

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

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 DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. Comput Appl Biosci 8:275–282

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp. 21–123

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

Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11:459–468

Lartillot N, Philippe H (2004) A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. Mol Biol Evol 21:1095–1109

Lockhart PJ, Steel MA, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA 93:1930–1934

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

Lockhart PJ, Huson D, Maier U, Fraunholz MJ, Van De Peer Y, Barbrook AC, Howe CJ, Steel MA (2000) How molecules evolve in eubacteria. Mol Biol Evol 17:835–838

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

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

Pagel M, Meade A (2004) A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. Syst Biol 53:571–581

Penny D, McComish BJ, Charleston MA, Hendy MD (2001) Mathematical elegance with biochemical realism: the covarion model of molecular evolution. J Mol Evol 53:711–723

Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5:50

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 B 269:1313–1316

R Development Core Team (2007) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org

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

Ruano-Rubio V, Fares MA (2007) Artifactual phylogenies caused by correlated distribution of substitution rates among sites and lineages: the good, the bad, and the ugly. Syst Biol 56:68–82

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing evolutionary trees. Mol Biol Evol 4:406–425

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics 18:502–504

Siddall ME (1998) Success of parsimony in the four-taxon case: long-branch repulsion by likelihood in the Farris zone. Cladistics 14:209–22

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

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

Steel M (2005) Should phylogenetic models be trying to "fit an elephant"? Trends Genet 21:307–309

Susko E, Inagaki Y, Roger AJ (2004) On inconsistency of the neighbour joining method and least squares estimation when distances are incorrectly specified. Mol Biol Evol 29:1629–1642

Swofford DL, Waddell PJ, Huelsenbeck JP, Foster PG, Lewis PO, Rogers JS (2001) Bias in phylogenetic estimation and its relevance to the choice between parsimony and likelihood methods. Syst Biol 50:525–39

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

Uzzell T, Corbin KW (1971) Fitting discrete probability distributions to evolutionary events. Science 172:1089–1096

Wang H-C, Spencer M, Susko E, Roger AJ (2007) Testing for covarion-like evolution in protein sequences. Mol Biol Evol 24:294–305

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

Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 15:555–556