There are several factors that may hamper the inference of reliable phylogenies, including insufficient data and oversimplified models. At best, these model deficiencies may lead to random noise that does not result in erroneous tree topologies. However, correlated parameters unaccounted for in the tree inference model can result in systematic biases towards highly supported erroneous topologies, branch lengths, and other parameter values.

Typically, probabilistic models proposed for the molecular evolution at a site are based on a discrete Markov process where each state represents a different character value (a nucleotide or amino acid) and transitions between states represent a site change fixation. Early methods assumed an identical and independent evolution process for all sites. Uzzell and Corbin (1971) accounted for the evolutionary heterogeneity among sites using rate-across-site (RAS) models. In these RAS models, site-specific relative rates are tied under a parametric sampling distribution, usually a gamma $\Gamma(\alpha, \beta)$ distribution. Despite their usefulness to approximate empirical data, RAS models are simplistic, as they do not take into account changes in the functional, environmental, and selective constraints acting on amino acid sites of a protein over time.

Covarion-like models were among the first to account for variations in evolutionary constraints among lineages. Fitch and Markowitz (1970) coined the term covarion, but covarion-like models were not given much attention until recently (e.g., Tuffley and Steel, 1998; Lopez et al., 1999; Galtier, 2001; Penny et al., 2001; Huelsenbeck, 2002; Kolaczkowski and Thornton, 2004; Philippe et al., 2005; Ane et al., 2005). Under these models, two separate Markov processes describe molecular evolution. The first is the typical transition between character states (amino acids for protein, nucleotides for DNA), whereas the second allows sites to switch between on-off states. Sites in the off state will remain constant, whereas those in the on state accept changes following 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. It also justifies the presence of extensive heterotachy, or changing rates, for a particular site across lineages that cannot be well explained by RAS models (Lockhart et al., 1998; Lopez et al., 2002; Philippe et al., 2005; Ane et al., 2005).

Extensions to this initial covarion model may be proposed to convey more accurately the complexity present in real data with a limited set of parameters. For example, Galtier (2001) considered a discrete gamma $\Gamma(\alpha)$ model that adds a fixed instantaneous rate of change, rate-of-rates ($v$), between rate categories for each lineage on every site. This extended covarion model does not allow, however, for rate-of-rates change between lineages that may be responsible of producing phylogenetic biases.

Currently, most phylogenetic studies use RAS models on simplicity and software availability grounds. Simultaneously, some methods have been proposed to determine significant within-site changes in evolutionary rates across lineages (Gu, 1999, 2001; Knudsen and Miyamoto, 2001; Mooers and Holmes, 2000; Susko et al., 2002). Other authors have carried out work on topologically misleading effects with special focus on shared invariant sites (Lockhart et al., 1998; Susko et al., 2004), on highly divergent or even random sites (Brinkmann and Philippe, 1999; Pisani, 2004; Susko et al., 2005), or without restriction to any of those two extremes (Chang, 1996;
Huelsenbeck, 1998; Inagaki et al., 2004; Kolaczkowski and Thornton, 2004).

Deviations from the RAS assumptions may lead to systematic errors. Chang (1996) showed that taxa tend to cluster together in a phylogeny when they display similarities in the distribution of rates among sites. Susko et al. (2004) reported that distance correction to account for heterogeneity in distance-based methods might become a one-way trip into long-branch attraction artifacts. Their results pinpoint a tendency towards long-branch attraction or repulsion, even using maximum likelihood (ML) methods, when the model and parameter values induce systematic misestimation of genetic distances. A covarion-like evolving data set, as those described by Tuffley and Steel (1998) or Galtier (2001), does not necessarily lead to an incorrect phylogeny. In fact, true sister groups may well present a greater correlation in their relative site rates and hence tend to be clustered together in the phylogeny. Heterogeneity in the rate-of-rates parameter values among lineages may however lead to mistaken phylogenetic groupings, as we will suggest in this study.

Here, we first investigate possible sources of bias from assuming RAS models in applying the maximum likelihood criterion on data that evolved following a covarion-like model of evolution. More precisely, we focus on evolutionary scenarios where lineages differ in the assignation of rate category per site. Then we propose a tool to visualize and test the presence of such biases.

METHODS

Simulation Model

To reproduce the bias caused by changes in evolutionary constraints, we used a similar approach as that used in previous work (Chang, 1996; Kolaczkowski and Thornton, 2004; Spencer et al., 2005; Philippe et al., 2005). However, we introduced important differences regarding the phylogeny used in simulations: the tree comprised a quartet of subtrees rather than a quartet of taxa, we considered a set of coefficients that account for the conservation of the relative site rate categories at the deepest divergence events (internal nodes) of the tree; we also accounted for possible differences in the shape parameter of the gamma RAS distribution of substitutions among subtrees. In this study we did not consider the effect of bipartitions of the data matrix based on different sets of branch lengths (see Fig. 2) as done in other studies; however, heterotachy is present due to random independent changes in relative rate category per site or difference in rate distributions across lineages.

We focused on unrooted phylogenies presenting four well-resolved monophyletic groups (Fig. 1). Each group is composed of a number of taxa and evolves under a standard RAS model. The shape of each subtree is controlled by three parameters (Fig. 1c): the total height of the tree, $h$, the number of simultaneous divergence events along every lineage, $d$ (thus the total number of leaves is $2^d$), and the relative branch length ratio, $r$, that modulates the temporal spacing between divergence events. The motivation behind considering subtrees rather than the typical unrooted four-taxon phylogeny is twofold: first, to provide enough data to infer back the rate for a site on each subtree and therefore its relative rate on that lineage, and second, to record the influence of different taxa sampling contexts on the bias strength.

We modeled heterogeneity among sites using a discrete approximation to $\Gamma(a)$ rate distribution with the number of rate categories $C = 16$ on all branches. Initially we considered an RAS model with each site
belonging to the same rate category throughout the tree. However, we included the possibility of changing the rate categories per site instantaneously on the two most internal nodes or divergence events, which led to the four well-supported subtrees. On both nodes we defined three transition probability matrices $P_{xx}$, $P_{xy}$, and $P_{yy}$, one per each incident branch. Each matrix element $p_{ij}$ indicates the probability for a site with category $i$ on the node to change to category $j$ that will apply to the incident branch and beyond. As we constrained all branch rates to follow a discrete gamma distribution, all matrices must yield equally probable stationary category frequency vectors $F = \{C^{-1}\}$. We further constrained the family of valid matrices in order to be able to parameterize the change in the distribution of rates across the tree with a single tuning value, $\theta$, which we used for the site rate category conservation coefficient throughout the text (Fig. 1a, b). Its value indicates the proportion of sites whose rate category is unchanged at the divergence event, whereas $(1 - \theta)$ indicates the percentage of sites that is susceptible to change to another rate category. In the latter case, the new category will be selected at random from all $C$ possible categories including its current value. Thus the resulting transition matrix is:

$$
P_{xy} = \begin{bmatrix}
\theta''_{xy} & \theta'_{xy} & \cdots & \theta'_{xy} \\
\theta'_{xy} & \theta''_{xy} & \cdots & \theta'_{xy} \\
\vdots & \vdots & \ddots & \vdots \\
\theta'_{xy} & \cdots & \theta'_{xy} & \theta''_{xy}
\end{bmatrix}
$$

(1)

where:

$$
\theta''_{xy} = (1 - \theta_{xy}) C^{-1}; \quad \forall \theta_{xy} \in [0, 1] \\
\theta'_{xy} = \theta_{xy} + \theta'_{xy}
$$

(2)

Under this restriction, reversion and transitivity are trivial because $P_{xy}$ is symmetric:

$$
P_{xy} = P_{yx}, \quad \theta_{xy} = \theta_{yx} \\
P_{xz} = P_{xy} \cdot P_{yz}, \quad \theta_{xz} = \theta_{xy} \cdot \theta_{yz}
$$

(3)

where $z$ is reachable from $x$ through node $y$. We based our simulations on a JTT (Jones et al., 1992) Markov process. Nonetheless, qualitative conclusions are extensible to other evolutionary models.

**Simulation Cases**

Due to computational limitations we only investigated covarion bias effects on three different parametric scenarios: (a) presence of changes in rate categories that are equally frequent in non-sister subtrees, (b) presence of changes in rate categories in the inner branch connecting phylogenetic pairs of subtrees; and (c) existence of different site rate sampling distributions. For all the simulations conducted here, we produce amino acid data sets following the model explained above based on a JTT substitution matrix. Each parameter values combination was repeated 50 to 100 times to reduce stochasticity. Unless otherwise indicated, alignments were 1000 amino acids long. We utilized the program CODEML from the PAML package version 3.14 (Yang, 1997) for evaluating the phylogenetic support based on RELL bootstrap proportions (Kishino and Hasegawa, 1989). For each data set we compared the support obtained by the three unrooted possible trees that were constrained so that all four subtrees are monophyletic and their topologies were known. Parameter values were specific for each simulation batch. All diagrams plot mean RELL support index among repetitions. Notice that this is not the same as considering the percentage of cases where that tree topology is the one that maximizes the likelihood, which is typically the statistic reported in this kind of studies. We used the R statistical package to produce all diagrams and specifically the locfit module to fit surfaces for the three-dimensional plots, and geneplotter for the triangular plots in Figure 10.

**Geometrical Visualization of Bias**

Intuitively, sites that may introduce bias towards a particular hypothesis (topology) are those for which the two non-sister lineages have similar rates and different from the other two lineages (Fig. 2). Therefore, assuming that we have enough data to accurately estimate rates per site per lineage, we can classify sites according to the hypothesis they favor. We expect that a proportion of sites support a particular topology due to stochastic events alone. However, their overall effect must cancel out and result in an unbiased sample.

There are three possible topologies for a fully resolved four-subtree-based unrooted phylogeny. A simple way to represent graphically these hypotheses is using the equidistant vertices of a regular triangle (Fig. 3). Others have used triangle geometry previously in a phylogenetic context although with a different purpose (Strimmer and von Haeseler, 1997; Kim, 2000). Within such a layout, impartial sites (those not favoring a particular hypothesis) will fall close to the middle point, $p = q$, which is defined as the proportion of sites that prefers topology $i$. The geometry of the layout and the size of the space occupied by the three topologies are determined by $p$ and $q$. For each data set we compared the support obtained by the three unrooted possible trees that were constrained so that all four subtrees are monophyletic and their topologies were known. Parameter values were specific for each simulation batch. All diagrams plot mean RELL support index among repetitions. Notice that this is not the same as considering the percentage of cases where that tree topology is the one that maximizes the likelihood, which is typically the statistic reported in this kind of studies. We used the R statistical package to produce all diagrams and specifically the locfit module to fit surfaces for the three-dimensional plots, and geneplotter for the triangular plots in Figure 10.

**Simulation Cases**

Due to computational limitations we only investigated covarion bias effects on three different parametric scenarios: (a) presence of changes in rate categories that are equally frequent in non-sister subtrees, (b) presence of changes in rate categories in the inner branch connecting phylogenetic pairs of subtrees; and (c) existence of different site rate sampling distributions. For all the simulations conducted here, we produce amino acid data sets following the model explained above based on a JTT substitution matrix. Each parameter values combination was repeated 50 to 100 times to reduce stochasticity. Unless otherwise indicated, alignments were 1000 amino acids long. We utilized the program CODEML from the PAML package version 3.14 (Yang, 1997) for evaluating the phylogenetic support based on RELL bootstrap proportions (Kishino and Hasegawa, 1989). For each data set we compared the support obtained by the three unrooted possible trees that were constrained so that all four subtrees are monophyletic and their topologies were known. Parameter values were specific for each simulation batch. All diagrams plot mean RELL support index among repetitions. Notice that this is not the same as considering the percentage of cases where that tree topology is the one that maximizes the likelihood, which is typically the statistic reported in this kind of studies. We used the R statistical package to produce all diagrams and specifically the locfit module to fit surfaces for the three-dimensional plots, and geneplotter for the triangular plots in Figure 10.

**Geometrical Visualization of Bias**

Intuitively, sites that may introduce bias towards a particular hypothesis (topology) are those for which the two non-sister lineages have similar rates and different from the other two lineages (Fig. 2). Therefore, assuming that we have enough data to accurately estimate rates per site per lineage, we can classify sites according to the hypothesis they favor. We expect that a proportion of sites support a particular topology due to stochastic events alone. However, their overall effect must cancel out and result in an unbiased sample. There are three possible topologies for a fully resolved four-subtree-based unrooted phylogeny. A simple way to represent graphically these hypotheses is using the equidistant vertices of a regular triangle (Fig. 3). Others have used triangle geometry previously in a phylogenetic context although with a different purpose (Strimmer and von Haeseler, 1997; Kim, 2000). Within such a layout, impartial sites (those not favoring a particular hypothesis) will fall close to the middle point, $p = q$, which is defined as the proportion of sites that prefers topology $i$. The geometry of the layout and the size of the space occupied by the three topologies are determined by $p$ and $q$. For each data set we compared the support obtained by the three unrooted possible trees that were constrained so that all four subtrees are monophyletic and their topologies were known. Parameter values were specific for each simulation batch. All diagrams plot mean RELL support index among repetitions. Notice that this is not the same as considering the percentage of cases where that tree topology is the one that maximizes the likelihood, which is typically the statistic reported in this kind of studies. We used the R statistical package to produce all diagrams and specifically the locfit module to fit surfaces for the three-dimensional plots, and geneplotter for the triangular plots in Figure 10.
Based on previous observations, lineages with more similar evolutionary processes should tend to attract each other. In this context, we may say that a site "prefers" a hypothesis \([ (x, y), v, w ] \) if for that site, ranks on lineages \( x, y \) and \( v, w \) are closer within pairs and different between pairs. A way to reflect this in quantitative terms is that a site will prefer a particular hypothesis if \( |R_x(i) - R_y(i)| \) and \( |R_v(i) - R_w(i)| \) are small (close to zero) and the average \( AVG(|R_x(i) - R_y(i)|, |R_v(i) - R_w(i)|) \) is big (close to one).

Instead of determining individual support indices for each site and hypothesis, we calculated coefficients that give to each hypothesis a weight proportional to the similarity of ranking of their two subtree pairs. Under this approach we yield three weight values, with only two degrees of freedom.

Let us consider the three hypotheses:

\[
H_0 = [(a_1, a_2), b_1, b_2] \\
H_1 = [(a_1, b_1), a_2, b_2] \\
H_2 = [(a_1, b_2), a_2, b_1] 
\] (4)

For each site \( i \) we calculated the value \( F \) that measures the amount of relative rate change between two subtrees \( (x \) and \( y) \) as:

\[
F(x, y, i) = 1 - |R_x(i) - R_y(i)|
\] (5)

Small differences in rates (rate conservation or convergence) will yield values close to 1, whereas extreme differences should yield \( F \) value closer to 0. \( F \) values are directly comparable only if they are estimated from subtrees with equivalent shapes (topology and branch lengths) as distributions of rate estimators used to rank sites depend on these subtree shapes. Therefore, in order to normalize ranking differences between all subtrees, we performed parametric simulations (using 1000 replicates) for each rate category based on RAS model assumptions. In these simulations, we used the parameter estimators obtained by ML analysis performed in PAML. Then we substituted the \( F \) values at each site by the complement of their corresponding cumulative probability \( F^c \) assuming that its rate category is the one with maximum posterior probability. This will yield uniformly distributed values from 0 to 1.

\[
F'(x, y, i) = Prob(F^c_{xy} > F(x, y, i))
\] (6)

Then we measure the total amount of change on a site \( i \) for each hypothesis with a second function, \( G \), defined as the product of the indices for its two individual phylogenetic subtree pairs:

\[
G_{H_0}(i) = F'(a_1, a_2, i) \cdot F'(b_1, b_2, i) \\
G_{H_1}(i) = F'(a_1, b_1, i) \cdot F'(a_2, b_2, i) \\
G_{H_2}(i) = F'(a_1, b_2, i) \cdot F'(a_2, b_1, i)
\] (7)
The resulting hypothesis indices also vary from 1 (total conservation of the substitution rate categories per site within pairs) to 0 (extreme change of the substitution rate categories per site at least within one pair). Then, we defined the weight, \( \omega_i \), on site \( i \) associated to a particular hypothesis \( H_j \) as:

\[
\omega_{H_j}(i) = \frac{G_{H_j}(i)}{G_{H_j}(i) + G_{H_0}(i) + G_{H_2}(i)}
\]  

(8)

Finally, we used these weights for combining arithmetically the three triangle vertices (hypotheses) in order to plot a single point that represents the site within the triangle. The bidimensional location \( (p_i) \) of that point can be calculated as:

\[
p_i = \frac{1}{3} (\omega_{H_0}(i) \cdot h_0 + \omega_{H_2}(i) \cdot h_1 + \omega_{H_2}(i) \cdot h_2)
\]  

(9)

Here \( h_j \) stands for the two-dimensional coordinates of the vertex that represents hypothesis \( H_j \). Notice that we can implement many alternative \( G \) or \( F \) functions generating different weight coefficients and plots with different statistical properties. Nonetheless, any alternative must be designed as to allow impartial sites to concentrate around the center of the triangle, whereas partial sites must cluster towards the corners (one hypothesis wins) or the edges (two draw and one loses).

**Bias Test and Measurement**

We can evaluate the assumptions in the RAS models by testing the goodness-of-fit of the expected to the observed distribution densities of the points in the triangle. Even though this may spot nonstandard samples, it would not automatically indicate which hypothesis benefits from that deviation. Besides, some samples may still be balanced and yet have significantly different plot densities. Alternatively, we could take a look at the mass center of the plot, that is, the arithmetic mean of all site points.

In the ideal case of equivalent subtrees with the same shape (topology and branch lengths), and assuming a genuine RAS evolution model, the mass center must be placed on the triangle barycenter equidistant from each tree hypothesis (topology) regardless of the presence or lack of any phylogenetic signal. Asymmetric heavy tails must pull this mean towards the winning alternative hypothesis (or further from the losing one). In consequence, one could use this to infer whether the sample is biased with respect to one or more hypotheses. Provided that the distribution of the mean plot point position approaches a bivariate normal distribution, we can readily obtain a simple formula to calculate a confidence area for the sample mass center as developed in the Appendix:

\[
P(D > d) = \exp \left[ -d^2 (2\sigma_D^2)^{-1} \right]
\]

\[
\sigma_D = \frac{\sigma_x}{\sqrt{L}}
\]  

(10)

where \( D \) designates the Euclidean distance from the sample mean to the barycenter, \( \sigma_x \) is the standard deviation of the sample's projection on any axis of choice with origin in the barycenter, and \( L \) is the sample size (data matrix length).

Nonetheless, distinct branch lengths and different number of taxa may affect the distribution of dots on the plot and change the mass center. This is due to the fact that the \( G \) values for each hypothesis, as described in Methods, are not independent from each other because they combine shared subtree estimators. In practice, however, the test starts to become liberal just with a great amount of data (10,000 sites) and severe tree shape height asymmetry (e.g., when trees present subtrees over 8 times taller, or with 4 times greater number of taxa, than others). A general solution to this issue is to further simulate the expected site plot and mass center, given the ML estimators of the full phylogeny, and to use the distance to this alternative center. Even though Equation (10) is still a good approximation under such circumstances, sample mean comparison tests, such as the Wilcoxon test, are probably more appropriate.

Another aspect to bear in mind is that since we sort rank ties by randomization, each execution can potentially return a different \( P \)-value. In order to stabilize the outcome, the test must be repeated a number of times and an averaged \( P \)-value must be selected. However, under this procedure the test becomes somewhat conservative. In this work we will proceed without repeats, as results for single data sets are not critical in the simulations.

As an alternative to the null hypothesis presented above, data sets can show different degrees of relative site rate correlation among lineages that may cause systematic distortions in the resulting topology and branch estimators. Interpretation of the results of such a test is independent of the amount of actual phylogenetic signal (inner branch length). Therefore, results can only point to the possibility of having inferred an artifactual phylogeny by ML under a RAS model.

The simulation and test programs were implemented in Java programming language utilizing functionality from the Java standard library, PAL library (Drummond and Strimmer, 2001), and Jakarta commons. We delegated on rates4site software (Mayrose et al., 2004) to obtain empirical Bayesian site rate estimates per each subtree.

**RESULTS AND DISCUSSION**

We investigated the effect of inducing different levels of site rate correlation between subtrees using the simulation model based on four monophyletic groups of taxa \( a_1, a_2, b_1, \) and \( b_2 \) as described in Methods and Figure 1. We intuitively expected that lineages with a more similar evolutionary process would attract each other due to a greater concordance on the rates across sites, or alternatively that lineages with more divergent evolutionary processes would repel accordingly. Higher pairwise site rate category conservation coefficients \( \theta_s \) (product of individual \( \theta \) along connecting nodes) should cluster lineages with increasing confidence as more data becomes available.
(b) Figure 4. Plot of the mean RELL support value for the wrong topology \((a_1, b_1), a_2, b_2\). (a) Simulated process: the conservation coefficient on the inner branch was fixed to 1 and the same values were assigned to non-sister subtrees: \(\theta_1\) for \(a_1\) and \(b_1\), and \(\theta_2\) for \(a_2\) and \(b_2\), respectively. Tree heights are set to 1 substitution per site, whereas the innermost branch length is set to 0.04. There are a total of four taxa per each group \((d = 2)\) and each branch is twice as long as the next branch towards the tree leaves \((r = 2)\). Rates are drawn from a gamma distribution \(\Gamma(\alpha)\) for all subtrees. (B) A plot of the mean RELL support value \((z\text{-axis})\) and the conservation coefficients \(\theta_1\) \((x\text{-axis})\) and \(\theta_2\) \((y\text{-axis})\). For this particular set of parameters, at very low conservation distance, \(|\theta_1 - \theta_2|\), the mean RELL value approaches a constant low support value (expected under a RAS model), whereas it increases considerably when the difference is extreme \((= 1)\).

The Bad

In the first simulation, we fixed \(\theta_{ab}\) (Fig. 1b) on the inner branch connecting the \(a\) and \(b\) groups to 1 so that there is no intrinsic rate category shuffle along this branch. \(\theta_{a1} = \theta_{b1} = \theta_1\) and \(\theta_{a2} = \theta_{b2} = \theta_2\) were assigned arbitrary values between 0 and 1 in steps of 0.1 (Fig. 4a). Initially, for each of these cases we fixed all subtrees total heights, \(h\), to a common average of one change per site, resulting in a symmetric and molecular clock-compatible phylogeny. The shape parameter value, \(\alpha\), was also set to 1 (i.e., a moderate to high degree of substitution rate heterogeneity among sites). Subtrees were composed of four taxa; the result of two rounds of duplication or speciation events on all lineages \((d = 2)\). Each branch was

![Figure 5](http://sysbio.oxfordjournals.org/)

**Figure 5.** Effect that changes in simulation parameters have on the RELL support for the three possible topologies. We investigate the marginal effect of five different parameters (one per each diagram column), fixing all other parameters to the values used in the surface plot in Figure 4: inner branch length \((0.01, 0.02, 0.04, 0.08, 0.16)\), shape parameter \((\alpha)\) \((0.05, 0.25, 0.5, 1, 1.5, 2)\), subtree common height \((h)\) \((0.25, 0.5, 0.75, 1, 1.5, 2)\), branch length factor \((r)\) \((0.25, 0.5, 1, 2, 4, 8)\), and number of taxa per subtree \((d)\) \((1, 2, 4, 8, 16)\). We tested these changes on a particular combination of conservation coefficients \(\theta_1 = 1\) and \(\theta_2 = 0.2\) situated at an average height on the plotted slope in Figure 4. These values are close to the inconsistency critical region. For each combination of values, we took the mean RELL support of 50 replicates for all three possible topologies. We performed the experiment for a number of different alignment lengths, although here we show results just for 1000, 3000, and 10,000 sites. There are no data for 10,000 sites and 16 taxa per subtree because it was computationally prohibitive.
Figure 6. Effect of subtree height asymmetry on the covarion bias strength. Here we investigate the effect of five distinct relative subtree height rearrangements. We simulated data sets of variable data lengths, although here we show results just for 1000 and 10,000 sites alignments. As in Figure 5, groups $a_1$ and $b_1$ share conservation coefficient $\theta_1 = 1$, whereas $a_2$ and $b_2$ share a different and much lower value $\theta_2 = 0.2$. In order to create variability in subtree heights, we study five different assortments where two or tree subtrees share the same height ($h$) and the remaining share the inverse ($h^{-1}$). $h = 1$ means the original symmetric tree that becomes more asymmetric as $h$ approaches 0 or infinite.

twice as short as the one that leads to the parent node ($r = 2$). The inner branch was set to 0.04 changes per site. This last value was chosen so that our simulated data sets are forced towards the inconsistency parametric edge. Consequently, relatively small changes in parameter values can make the results wander from yielding the right topology to the wrong one with increasing probability as the size of the data set increases. Under these conditions, the bootstrap support for the right topology following an RAS model of evolution was 80% out of 100 replicates.

A possible biological interpretation of such an arrangement is that smaller $\theta$s indicate a greater dissimilarity between evolutionary constraints on different lineages. For instance, paralogs within a gene family with distinct functions would show a similar degree of disparity in within-site evolutionary rates. If $\theta_1 = \theta_2$, there is the same degree of covariance between all subtrees. Consequently, we expect to obtain a roughly random type of error with no systematic bias. However, if two lineages have a higher degree of rate category conservation; for example, when the two paralogous groups of sequences are functionally redundant, there will be a higher correlation of site rates and a stronger attraction under non-covarion models between these two lineages. The opposite can be also true for the two lineages with lower site rate category conservation. Thus we would expect to see wrong topology resolution as the difference between $\theta_1$ and $\theta_2$ increases. In fact, our simulations, based on a symmetrical phylogeny where all subtrees have the same height, show that such an increase leads from reasonable support for the true topology towards false support for the hypothesis in which lineages with greater site rate conservation appear together (Fig. 4b).

Further, we took a particular combination of $\theta$ values ($\theta_1 = 1, \theta_2 = 0.2$) and explored the effect of changing values for key characteristics of the input phylogeny at different data sets sizes (Fig. 5). As expected, the increase of the amount of phylogenetic signal (inner branch length) neutralizes systematic error effects. On the other hand, the increase of subtree height does reinforce systematic error because of the loss of phylogenetic signal in the innermost branch. Covarion bias relies on the existence of heterogeneity of rates among sites because no site rate categories would be considered otherwise. Accordingly, a lower shape parameter value (more severe heterogeneity) favors resolution for the wrong topology, whereas a higher value does just the opposite. Besides, proper taxon
sampling may reduce the effect of bias. In fact, adding more taxa reduces the bias strength. Short deep and long-tip branches seem to be less problematic than long deep and short-tip branches are. Thus, we can conclude that the characteristics of the evolutionary process at the subtrees have a remarkable effect on phylogeny recovery because long deep branches seem to have a greater impact, something that is in agreement with previous studies (Kim, 1996, 1998; Bremer et al., 1999).

Finally, within-site rate heterogeneity among lineages can also alter the outcome significantly. Among all alternative combinations studied, it is only the ones that elongate or shorten lineages together with the same θ values that have a pronounced effect in comparison with the results obtained with the symmetric settings (Fig. 6). In fact, bias increases when lineages with high θ values are accelerated, whereas it decreases when lineages with low θ values accumulate a greater number of changes. We also studied the effect that different numbers of taxa among subtrees has on producing such phylogeny biases. Results here are qualitatively parallel to the ones obtained by tree height asymmetry, with accelerated lineages and more conserved lineages having equivalent effects to poorly sampled and well-sampled lineages, respectively. This, together with the results presented in Figure 5, suggests that long deep branches (caused by rate acceleration, poor sampling or high tree branch length ratio values) may have a major role in catalyzing phylogenetic bias.

The Good

For the second simulation scenario, we allowed θ_{ab} to vary from 0 to 1 in steps of 0.1, while we fixed the θ for each external branch to the same value θ_{12} taking the same set of values as formerly (Fig. 7a). The tree shape is the same as for “the bad” case, although here we set up the internal branch to 0.01 to hamper the recovery of the correct topology (48% bootstrap out of 100 repeats). Intuitively, simplistic assumptions may actually assist in obtaining the right topology, as closely related taxa should more often share evolutionary traits not accounted for by the model. Accordingly, decreasing θ_{ab} (no rate category conservation across the inner branch) and increasing subtree θ_{s} (conservation between sister groups) yields high support for the true topology (Fig. 7b) despite the limited phylogenetic signal available.

Although this may seem to be preferable for obvious reasons, we may well want to have an objective measure of the evidence towards each hypothesis. This same reasoning was used to debunk the arguments concerning the superiority of parsimony over likelihood in the so-called Farris zone (Siddall et al., 1998; Swoford et al., 2001), which is part of the long-standing literature battle regarding long-branch attraction effects and the performance of maximum likelihood and parsimony (e.g., Sullivan and Swoford, 2001; Kolaczkowski and Thornton, 2004; Gadagkar and Kumar, 2005; Spencer et al., 2005; Gaucher and Miyamoto, 2005; Brinkmann et al., 2005).

The Ugly

In the third simulated scenario, we explored the effect of distinct shape parameter values under different degrees of site rate category conservation. Two pairs of phylogenetic groups evolved under a gamma RAS model, with the first pair of subtrees (a1, b1) using a shape parameter, a1, and the other two subtrees using a different shape parameter, a2. For the inner branch, we used one of the pair values, a_{ab} = a_1 (Fig. 8b) and an interpolated version of both values, a_{ab} = exp(ln(a_1) + ln(a_2))/2 (Fig. 8c-f). This means that even if a site belongs to the same rate category throughout the phylogeny, its relative rate will change among lineages depending on the difference in site rate distributions. Once more, we intuitively expected to reproduce systematic error that would join lineages with convergent rate among site distributions a_1 ≠ a_2. Topology and branch lengths parameters are set up as in the good case including weak phylogenetic signal, so that we can see support oscillations in both directions (pro and against the wrong rate distribution convergent topology). We also simulated different shape parameters under different conservation levels between
FIGURE 8. Plot of the mean RELL support value for the wrong topology \((a_1, b_1, a_2, b_2)\) with different rate-across-site sampling distributions. (a) The simulated process: initially, rate category conservation is full along the inner branch \((\theta_{ab} = 1)\) and equals \(\theta_{d} \) for all branches leading to subtrees, so that the pairwise conservation coefficient is equal for all pairs of groups. RAS distribution shape parameters change across subtrees, \(a_1\) for \(a_1\) and \(b_1\), and \(a_2\) for \(a_2\) and \(b_2\). We assigned two different values to the shape parameter along the inner branch: we took one of the group values \(a_{ab} = \theta_{ab}\) (b) or alternatively used an interpolated value \(a_{ab} = \exp(\text{ln}(\theta_{a}) + \text{ln}(\theta_{b}))/2\) (c, d, e, and f). We simulated 100 repeats at shape parameter values \(a_1\) and \(a_2\) \(\in \{\exp(x)\mid x \in \{-5, -4, \ldots, 5\}\}\). (b) The effect of different combinations of shape parameters in the mean RELL support value where all \(\theta_s\) are set to 1. (d, e, and f) Changes in RELL support surface in (c) as we decreased the coefficient \(\theta_{d}\) to 0.75, 0.5, and 0.25, respectively. Bias fades away as more sites are given different random rate categories per subtree. Convergent RAS distributions have a weakening effect in this case.

1 and 0 in steps of 0.25 but maintained them equal among all subtrees (Fig. 8a) so that:

\[
\begin{align*}
\theta_{ab} &= 1 \\
\theta_1 &= \theta_2 = \theta_{12} \in [0, 1] \\
\theta_{xy} &= \theta_{12}^2, \forall x, y \in \{a_1, a_2, b_1, b_2\}
\end{align*}
\]

The outcome is quite surprising. When \(\theta_s\) are high (close to 1), for the most part, different shape parameter values drag together those subtree pairs with equal rate distribution shapes (Fig. 8b). Nonetheless, greater difference in shape parameters does not necessarily produce greater attraction between convergent subtrees. Moreover, there are zones where support falls below the level
achieved when there are no differences between rate distributions ($\alpha_1 = \alpha_2$). Consequently, repulsion takes place between lineages with the same shape parameter. In this particular simulation batch, this phenomenon occurred when medium/high heterogeneity (around $\alpha = 1$) and high heterogeneity (when $\alpha$ approaches zero) are combined. In fact, these sinuous support surface changes as the relative subtree heights vary to break the initial clock like overall structure (Fig. 9) or when the inner branch shape parameter is set up differently (Fig. 8b versus 8c).

This may be due to a delicate equilibrium between counterpoised attraction and repulsion effects between all lineages because a single shape parameter is used to maximize the global likelihood, despite that each combination of two subtrees would have a different marginally best-fitting pairwise shape parameter. Susko et al. (2004) concluded that the bias strength and the inconsistency zone would depend on the relative branch lengths.

Nonetheless, one may question the biochemical relevance of such a particular scenario; changes on the overall distribution of site rates under the RAS model or its tuning parameters must be accompanied by arbitrary changes in site rate categories, resulting in decreasing $\theta$s. We simulated the same shape parameter subspace at different overall rate category conservation coefficients (Fig. 8c–f). The outcome revealed that repulsion or attraction between subtrees caused by convergence in the sampling distribution of rates between lineages strongly depends on the conservation of site relative rates. This puts forward the conclusion that site rate category dependency between lineages is required for attraction or repulsion over possible influences of relative rate probability distribution differences or similarities. Therefore, different RAS distributions cannot be used alone as an argument to postulate possible systematic bias in a data set or to indicate what lineages should attract.

### Visualization and Test

Using the visualization method described in Methods, we plotted data sets simulated under different model conditions including the first and second parameter subspaces previously investigated. A regular RAS process ($\theta = 1$ for all branches) generates a balanced density surface (Fig. 10a) around the center of the triangle, although segments from the center to the vertices do accumulate a slightly greater amount of points. Covarion unbiased data sets ($\theta$ values are 0 for all the subtree-leading branches and $\theta = 1$ for the innermost branch) produce a visibly different point distribution with similar amount of points cumulating towards each hypothesis vertex (Fig. 10b). In contrast, covarion-biased data sets generated by correlating the conservation coefficients between pairs of subtrees yield unbalanced plots (Fig. 10c, d and f). Plot differences between biased and unbiased alignments are evident to the naked eye. The overall plot distribution may change if subtrees have clearly distinct shapes (number of taxa, branch length, and topology) as illustrated in Figures 10e and f. Nonetheless, even in these circumstances, the averaged mean point position remains close to the center of the triangle in genuine RAS data sets.

These features match the resulting real or artifactual tree support obtained by ML tree reconstruction. Accordingly, application of the proposed test on these simulated data sets allows the detection of deviations with increasing sensitivity as the bias strength grows (Fig. 11). Sensitivity here is measured as the percentage of positives with a significance level of 0.05 in 1000 amino acid long alignments.

Regarding “the ugly” case, the subtree pairwise rate difference ranking described above also detects parametric areas where systematic error occurs (Fig. 11d). Nonetheless, it seems not to be able to discern between attraction and repulsion, supporting the hypothesis that joins subtrees with similar site rate distributions. The fact that the parametric area of attraction and repulsion changes as different modeling elements are altered suggests that this will not be easily resolvable with an ad hoc approach. Here, the use of the right or fairly well approximated model of evolution is required. This issue will require further investigation; for the moment, we recommend to perform simulations when there is an important difference between shape distributions in different regions of the phylogeny.
FIGURE 10. Example of bias visualization plots based on the layout described in Figure 3. These are the product of 10000 amino acid long alignments after smoothing the resulting scattered site plot. Darker areas indicate a greater concentration of site points. The circle indicates the center of the triangle, whereas the x points to the mass center of the cloud. (a) Plot resulting from a RAS substitution process. In this figure we used the parameterization of “the bad” case but fixing all $\theta$s values to 1. (b) “Unbiased” covarion process: as in (a) but setting all conservation coefficients $\theta$ to zero so that rates are selected at random per each lineage. (c) A covarion evolving data set in which non-sister lineages attract: “the bad” case where $\theta_1 = 1$ and $\theta_2 = 0$. (d) Covarion process in which actual sister lineages attract: “the good” case where $\theta_{aa} = 1$ and $\theta_{ab} = 0$. (e) Represents an RAS process with a combination of different number of taxa and subtree heights. Subtrees $a_1$ and $b_1$ are 2 substitutions tall and have 8 taxa each, whereas the height of subtrees $a_2$ and $b_2$ is set to 0.25 substitutions and contain just 2 taxa each. (f) “The bad” case applied to (e) setting where $\theta_1 = 1$ and $\theta_2 = 0$. The distribution of dots clearly changes if subtrees are asymmetric. Biased data sets differ in that points migrate to the favored hypothesis vertex. Consequently, the sample mass center clearly breaks off from the center of the triangle (c, d, and f).
FIGURE 11. Sensitivity (solid line) of the proposed test at different values for model parameters that cause bias, plotted against the average RELL value of the favored hypothesis (dashed line). (a) The bad case, distantly related lineages attract as difference between conservation coefficient $\theta_1 - \theta_2$ increases from 0 to 1 (Fig. 4). $\theta_1$ takes values from 0.5 to 1 in steps of 0.05. $\theta_2$ is set to 1 - $\theta_1$. (b) On the good case when closely related lineages attract, the inner branch conservation coefficient $\theta_{ab}$ varies from 0 to 1 in steps of 0.1. Other coefficients on branches leading to the four subtrees are set to 1 (Fig. 7). (c) Again on the second simulation batch, but in this case as lineage conservation coefficients vary from 0 to 1 and the inner branch coefficient is set to 1. (d) In the ugly case, setting $a_{ab} = a_1 = 1$, all $\theta$s = 1 and giving $a_2$ values in $\{\exp(x): x \in (-2, -1.5, -1, ..., 3)\}$. Notice that in this particular case the bias is pro the wrong topology when $a_1 < a_2[\ln(a_2) > 0]$ and pro the right topology when $a_1 > a_2$. In all cases we repeated the experiment 100 times.

REMARKS

It is important to find solutions to the limitations of current phylogenetic methods and evolutionary models. Incorrect phylogenetic resolution may invalidate any result obtained by downstream analyses. A first step towards this aim is to characterize peculiarities of evolutionary processes that may cause bias and provide assessing tools to localize their occurrence in real data sets.

In this work, we illustrated possible misleading effects of some violations of RAS model assumptions regarding evolutionary rates among sites and lineages. Here we extend the observation done by previous work on this aspect of molecular evolution (Chang, 1996; Kolaczkowski and Thornton, 2004), with a more general and relaxed heterogeneous model of evolution. Bias is caused by different degrees of site rate drift among lineages rather than specially tailored conflictive rate convergence. This phenomenon may also be reproducible when assuming a process accounting for continuous changes in site rates categories (Tuffley and Steel, 1998; Galtier, 2001) that allows for different rate-of-rates among lineages. We investigated different overlapping effects of some model parameters on the intensity of the systematic error caused by covarion-like evolution. It is important to note the importance that taxa sampling may have in reducing the resulting systematic error. Also, we tentatively proposed a tool to visualize and test for the presence of bias on four well-supported subtree phylogenies based on simple geometric concepts. In other words, we test the suitability of using RAS models for resolving the deepest branches of the phylogenetic tree. Reliability is especially dubious at this depth, as any posterior fixed changes (rest of the tree height) constitute noise regarding the resolution of these relationships. The closest procedure to our approach has been to check, normally a posteriori, the changes on the support values after selective resampling of data (Lockhart et al., 1998; Inagaki et al., 2004; Pisani, 2004; Brinkmann et al., 2005).
Although here we have focused only on quartet analysis, it may be extendable to a greater number of monophyletic groups by means of quartet puzzling--like techniques (Strimmer and von Haeseler, 1996) or porting the geometrical framework described herein to a higher dimensional space. In order to test discrepancies in rates among lineages, we needed to have a reasonable number of taxa per each group. Therefore this method cannot be used accurately with poorly taxed groups. Unfortunately, this is not an uncommon situation especially in studies performed on poorly sequenced taxonomy groups. Perhaps this may be avoided using the other dimension of the data, increasing the number of sites compared each time using a sliding window.

In any case, it is important to bear in mind that there is more than a single source of systematic error; for example, compositional heterogeneity (Moore and Holmes, 2000; Foster, 2004). Moreover, bias may reinforce the right topology, as close phylogenetic relatives should intuitively tend to share traits that may not be accounted for by mainstream models. Therefore, thoughtful phylogeny curation is necessary in order to accurately assess the cause and effect of such model singularities.

We may devise some correcting protocols and alternative methods inspired by the simulation model and test proposed. Certainly, in molecular data sets, the accuracy of ML and Bayesian analyses may increase by including more sets of parameters in existing models such as the conservation coefficient per branch or subtree and by accounting for within-site rate changes. In this context, one could use likelihood-ratio tests or alternative model sorting criteria to determine whether the enhanced model explains the data significantly better and whether the simplified model is susceptible to bias. However, that may unbearably increase the computational cost and risk of overfitting. Model-based distance methods such as neighbor-joining could benefit from the fact that each pair distance need not to be estimated using the same global model. A plausible alternative is then to generate a distance matrix, where many intersubtree or intertaxa models approximate better the evolutionary path between pairs of monophyletic groups. Another possibility is to perform a priori site-weighted bootstrapping or removal in order to improve data fitness for the model of choice. Although this last option may not seem to be the most attractive, as we may need to discard a sizable portion of the data, it is in fact constantly being done; e.g., leaving out the third codon position in coding DNA analyses.

In summary, diagnostic analysis approaches, like the one proposed in this work, may complement existing methods such as the maximum-likelihood and Bayesian ones and provide a visual statistical framework to take into account hidden parameters (covariation evolution, compositional bias, an so forth) that, when ignored, may hamper accurate phylogenetic inference.


APPENDIX

Test Equation Demonstration

Under an RAS model of evolution and considering subtree with similar tree shapes, the bias plot proposed here must look the same (present the same density of points) from any hypothesis (tree topology) perspective. Let us define the pair (X_0, Y_0) as the mean point location of the plot when we set the origin of coordinates on the triangle center and when the y-axis is set to pass through the hypothesis H vertex based on the layout depicted in Figure 3. Then based on the symmetry assumption, variance should be equal for all three coordinate systems:

$$
\text{VAR}(X_0) = \text{VAR}(X_H) = \text{VAR}(X_R)
$$

$$
\text{VAR}(Y_0) = \text{VAR}(Y_H) = \text{VAR}(Y_R)
$$

It is easy to transform a point (X, Y) based on a coordinate system to another that is a simple rotation of the former with angle \(\lambda\):

$$
\begin{bmatrix}
X' \\
Y'
\end{bmatrix} =
\begin{bmatrix}
\cos(\lambda) & \sin(\lambda) \\
-\sin(\lambda) & \cos(\lambda)
\end{bmatrix}
\begin{bmatrix}
X \\
Y
\end{bmatrix}
$$

Knowing that:

$$
\text{VAR}[aX + bY] = a^2\text{VAR}(X) + b^2\text{VAR}(Y) + a b \text{COV}(X, Y)
$$

And using the transformation Equation (2), the variance for hypotheses H_0 or H_1 (\(\lambda = \pi/3\) and \(-2\pi/3\), respectively) can be expressed as a function of the variance for hypothesis H as follows:

$$
\text{VAR}(X_H) = \cos^2(\lambda)\text{VAR}(X_0) + \sin^2(\lambda)\text{VAR}(Y_0) + 2\cos(\lambda)\sin(\lambda)\text{COV}(X_0, Y_0)
$$

$$
\text{VAR}(Y_H) = \cos^2(\lambda)\text{VAR}(Y_0) + \sin^2(\lambda)\text{VAR}(X_0) - 2\cos(\lambda)\sin(\lambda)\text{COV}(X_0, Y_0)
$$

Let us summarize Equation (4) using the following notation:

$$
V(\lambda, X, Y) = \cos^2(\lambda) VAR(X) + \sin^2(\lambda) VAR(Y)
$$

$$
C(\lambda, X, Y) = 2\cos(\lambda)\sin(\lambda) COV(X, Y)
$$

As \(\sin(\lambda) = -\sin(-\lambda)\) and \(\cos(\lambda) = \cos(-\lambda)\), it follows that:

$$
V(\lambda, X, Y) = V(-\lambda, X, Y)
$$

$$
C(\lambda, X, Y) = -C(-\lambda, X, Y)
$$

Consequently, the covariance term, C, must be null for all three hypotheses:

$$
\text{VAR}(X_H) = V(\lambda, X_H, Y_H) + C(\lambda, X_H, Y_H) = 0
$$

$$
\text{VAR}(Y_H) = V(-\lambda, X_H, Y_H) - C(\lambda, X_H, Y_H) = 0
$$

$$
V(\lambda, X_H, Y_H) + C(\lambda, X_H, Y_H) = V(X, Y, X_H, Y_H)
$$

$$
- C(\lambda, X_H, Y_H) = 0
$$

First submitted 2 November 2005; revisions returned 6 January 2006; final acceptance 25 August 2006

Associate Editor: Tim Collins
This last equality is satisfied if and only if λ is multiple of π/2 (not the case) or when the covariance is zero. Taking this into account we can easily prove that the variance of $Y_{H1}$ must be the same as that of $X_{H1}$ to satisfy the symmetry condition:

$$\text{VAR}(X_{H1}) = \cos^2(\lambda)\text{VAR}(X_{H1}) + \sin^2(\lambda)\text{VAR}(Y_{H1}) + 0$$

This equation only holds if:

$$\text{VAR}(Y_{H1}) = 1; \text{VAR}(Y_{H1}) = \text{VAR}(X_{H1})$$

because $\cos^2(\lambda) + \sin^2(\lambda) = 1$.

Here we consider the fact that for any angle λ both coordinates have equal variances and their covariances are equal to zero. Assuming that the site position obtained constitutes an independent sample, the central limit theorem (CLT) ensures that the distribution of the means approaches a Gaussian and, because it has the same variance from every angle x-axis or y-axis, it must approximate a circular 2-D bell centered on the triangle barycenter where the standard error is $\frac{\pi}{2N}$. Nonetheless, because site point coordinates are calculated based on parameter estimators obtained using the same site sample, it is not certain that the independence condition for CLT is actually met with a limited amount of data. Consequently, we need to confirm any conclusion below through simulations.

We can determine the formula for the cumulative probability of the distance to the center taking as starting point the bivariate normal distribution or the chi-square. Here we develop the former. As the covariance is null, the mean is zero and the deviation is the same from every axis:

$$N(x, y) = \frac{1}{2\pi\sigma^2} \exp \left( -\frac{x^2 + y^2}{2\sigma^2} \right)$$

For further simplification, we integrate the probability volume along the axis $y = 0$, within certain distance $x$ using the circle perimeter equation as follows:

$$P(D \leq d) = \int_{0}^{d} N(x, 0) L(x) dx$$

$$L(x) = 2\pi x, \text{ perimeter}$$

$$P(D \leq d) = \frac{1}{2\pi} \int_{0}^{d} x \exp \left( -\frac{x^2}{2\sigma^2} \right) dx$$

$$= \frac{1}{2\sigma^2} \int_{0}^{\sqrt{2}} z \exp \left( -\frac{z^2}{2\sigma^2} \right) \sqrt{2} dz; \quad z = x^2$$

$$= \frac{1}{2\sigma^2} \left[ -2\sigma^2 \exp \left( -\frac{z^2}{2\sigma^2} \right) \right]_{0}^{\sqrt{2}}$$

$$P(D \leq d) = 1 - \exp \left( \frac{-d^2}{2\sigma^2} \right);$$

$$P(D > d) = \exp \left( \frac{-d^2}{2\sigma^2} \right)$$

We confirmed these results performing simulations and testing the goodness-of-fit of the $P$-value and mean point distributions using Kolmogorov-Smirnov test against a uniform and bivariate normal distribution, respectively. The simulations show that this constitutes a good approximation even with asymmetric data sets (different subtree heights and number of taxa) up to 10,000 sites in length. Here the distance must be taken from the mean point centre estimated through parametric bootstrapping. However, using general samples, mean or median comparison tests like Wilcoxon test are more recommendable in these cases.