Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion

Emmanuel Paradis

Molecular dating of phylogenies by likelihood methods: a comparison of models and a new information criterion

Emmanuel Paradis

Institut de Recherche pour le Développement, ISEM UMR 226/5554 – UM2/CNRS/IRD, Jl. Taman Kemang 32B, Jakarta 12730, Indonesia

Abstract

Dating the divergence in a phylogenetic tree is a fundamental step in evolutionary analysis. Some extensions and improvements of the penalised likelihood method originally presented by Sanderson are introduced. The improvements are the introduction of alternative models, including one with non-correlated rates of molecular substitution ("relaxed" model), a completely reworked fitting algorithm that considers the high-dimensionality of the optimisation problem, and the development of a new information criterion for model selection in the presence of a penalised term. It is also shown that the strict clock model is a special case of the present approach. An extensive simulation study was conducted to assess the statistical performance of these improvements. Overall, the different estimators studied here appeared as unbiased though their variance varied depending on the fitted and the simulated models and on the number of calibration points. The strict clock model gave good estimates of branch lengths even in the presence of heterogeneous substitution rates. The correlated model gave the best estimates of substitution rates whatever the model used to simulate the data. These results, which are certainly the first from an extensive simulation study of a molecular dating method, call for more comparison with alternative methods, as well as further work on the developments introduced here.

Keywords: large-scale estimation, molecular dating, penalised likelihood, rate smoothing, relaxed molecular clock

1. Introduction

The molecular divergence among a set of molecular sequences is the product of the time they separated and the rate at which substitutions accumulated. The crux of the problem of phylogenetic molecular dating lies in the fact these two components are generally confounded, and untangling them can be done only with some assumptions. One of these assumptions is that molecular substitutions accumulate at a constant rate—the molecular clock. When this assumption is applied in its most radical version (i.e., a single rate for all taxonomic groups and constant in time), the exercise of molecular dating is considerably simplified since it can be accomplished even without a phylogenetic tree, for instance with pairwise genetic distances.

It is a fact that the molecular clock cannot be held generally, though this assumption may be

of some value for recently diverged sequences (Brown and Yang, 2011). A number of studies have characterized variation in substitution rates among taxonomic groups and through time (e.g., Pereira and Baker, 2006). Thus, a rooted phylogenetic tree appears as appropriate to handle this problem because the internal nodes of such a tree might be interpreted as branching points through time. Sanderson (1997) was one of the first authors to propose a general solution by assuming that substitution rates on two contiguous branches are likely to be similar (in other words, they are auto-correlated). By contrast, previous authors handled the same problem by applying strong restrictions on how rates vary in a tree (see citations in Sanderson, 1997).

In a subsequent paper, Sanderson (2002) refined his approach by introducing a likelihood component while keeping the assumption of auto-correlated rates. At the same time, Bayesian methods were developed with similar modeling assumptions though with different fitting approaches (e.g., Thorne et al., 1998; Thorne and Kishino, 2002). Since then, the Bayesian approach to molecular dating has flourished in a diversity of refinements (e.g., Drummond et al., 2006; Yang and Rannala, 2006; Lepage et al., 2007; Lartillot et al., 2009; Guindon, 2010; dos Reis and Yang, 2011; Wilkinson et al., 2011). The Bayesian approach has not been the only one explored during the last decade. Britton et al. (2002) proposed the mean path length to estimate divergence times from a non-ultrametric tree with branch lengths in number of substitutions. Xia and Yang (2011) developed a method based on least squares to estimate a chronogram including the possibility to model auto-correlated rate variation. These two methods have the advantage of being easily and quickly run.

In the next section, I first review Sanderson's framework and show how it can be generalized to include other models of substitution rate variation. I then introduce a new information criterion that helps to decide which models of rate variation best describes the data. I also propose some improvements on model fitting by penalised likelihood. An extensive simulation was done to assess the statistical properties of the methods presented in this paper.

2. Methods

2.1. Sanderson's penalised likelihood framework

Sanderson's penalised likelihood function (denoted Ψ) is based on two components:

$$\Psi = \ln L - \Phi, \tag{1}$$

where L is the likelihood function of a model of the branch lengths, and Φ is a function constraining rate variation. Essentially, L is a parametric component and Φ is a nonparametric one. Sanderson (2002) originally used a Poisson model for the first component and a constraint that rates on contiguous branches are more likely to be similar. Specifically, the likelihood function is:

$$L = \prod_i \zeta_i^{x_i} \frac{e^{-\zeta_i}}{x_i!},$$

with x_i being the number of substitutions observed on branch i, and $\zeta_i = r_i t_i$ where r_i is the substitution rate and t_i is the time length of the same branch. The x_i's are the input data usually from a non-ultrametric tree estimated by maximum likelihood, Bayesian estimation, or a distance-based method. Typically, they will be expressed as mean number of substitution per site, so it is not required to know the sequence length. The product is made over all branches of the tree. The log-likelihood function is thus:

$$\ln L = \sum_i x_i \ln \zeta_i - \zeta_i - \ln x_i!. \tag{2}$$

The nonparametric part is:

$$\Phi = \sum_{k,j} (r_k - r_j)^2 + \mathrm{var}(r_{\mathrm{basal}}),$$

where k and j denote two contiguous branches (i.e., consecutive in time), and the subscript 'basal' denotes the basal branches of the tree. Φ is further multiplied by a "smoothing" parameter λ which controls the trade-off between both components. In the present paper, I will leave aside the problem of estimating the smoothing parameter and will set, somehow arbitrarily, $\lambda = 1$ in the rest of the study.

It appears clearly that such a penalised likelihood approach can be a general framework which does not need to be restricted to the above assumptions on rate variation. Kim and Sanderson (2008) discussed how this can be generalised in a broader context of phylogenetic inference in-

cluding linking likelihood, parsimony, and Bayesian methods. In the next section I show how non-correlated relaxed clock models can be implemented using this framework in the context of molecular dating.

2.2. Non-correlated relaxed clock model

It is possible to relax the assumption of auto-correlation among rates; however, to permit estimation it is necessary to introduce another constraint on rate variation. Bayesian methods typically use a lognormal prior on substitution rates: such a distribution can be used to constrain the overall variation of these rates over the tree. Other distributions may be used, such as Γ, uniform, or Beta.

Let F be the theoretical cumulative density function of the distribution of the substitution rates. The nonparametric component can now be defined as:

$$\Phi = \sum_i \left[\tilde{\mathcal{F}}(r_i) - F(r_i)\right]^2,$$

where $\tilde{\mathcal{F}}(r_i)$ is the empirical cumulative distribution function of the rates. This is then used in Equation (1) to estimate the parameters in the same way than for the auto-correlated model (see *Fitting Algorithm*). This penalty function can be seen as a special case of the general formulation from Kim and Sanderson (2008, their eq. 3).

How to choose a formulation for F? Typically, previous studies suggest that substitution rate variation follows a power-like distribution with most rates being small and few large. Several probabilistic functions may be used to model such distributions. The gamma law, $\Gamma_{\alpha,s}$, appears as a practical choice here since the shape parameter α can be estimated with the mean of the rates throughout the tree, and the scale parameter s can be fixed equal to one, resulting in an appropriate distribution. Thus the penalty term above can be calculated without the need to estimate additional parameters.

It must be noted that the expression 'relaxed clock' has been used with different meanings in the literature. To avoid confusion, in this paper I name "relaxed" model the model assuming that each branch of the tree has its own substitution rate with no assumption on correlation between contiguous branches.

2.3. Maximum likelihood approach with discrete rate variation

An issue in dating a phylogenetic tree with variable rates is how one decides which part of the tree evolves faster than others. Yang (2004) discussed the problem of assigning a rate value to a particular branch (see also a critique of earlier approaches by Sanderson, 1997). On the other hand, the penalised likelihood approach does not assign rates to branches, though it requires to estimate a rate value to each branch which may result in the need to estimate a large number of parameters. Yang (2004) developed a method to assign rates to branches based on the initial (non-clock) tree. The divergence times are then estimated by maximum likelihood.

Here I propose a method that is inspired from discrete rate category method used in phylogenetic inference (Yang, 1994). With this method, it is assumed that rates vary in a discrete way so that we can make categories of branches characterised by different rates. However, we do not know in which category each branch belongs to, but by estimating the frequency of each category, we can calculate the contribution of each branch to Equation (2) by summing the contribution of each rate weighted by its frequency. Let us denote as c the number of categories, so the number of parameters to be estimated, apart of the dates, will be $2c-1$. Typically, c will be small (between 1 and 10) because if it is large then the present model tend to the "relaxed" one with the additional cost of estimating frequencies. Obviously, $c=1$ implies a strict clock model.

By contrast to the two previous non-clock models, the present one is fitted by full maximum likelihood since there is no penalty term Φ.

2.4. A new information criterion

The recent literature has seen a number of contributions developing alternative models relaxing the molecular clock hypothesis. An important issue is how to select the correct model. Bayesian methods use criteria based on Bayes factors (e.g., Lepage et al., 2007; Ho and Lanfear, 2010). In the present framework, a likelihood-based criterion must be used. A widely used model selection criterion is the Akaike information criterion (AIC, Akaike, 1973). We recall that $\mathrm{AIC} = -2\ln L + 2k$ with k being the number of parameters estimated from the data (sometimes called the free parameters). Here $\ln L$ is given by Equation 2, but we must also include the contribution of the penalty term Φ so that the smaller Φ the smaller the information criterion. In other words, we should

favour models with predictions on rate variation that best conform to the data. A simple solution could be to compute $-2\ln L + 2k + \lambda\Phi$ but this would ignore the fact that Φ is actually made of several elements. A solution is to make a multivariate decomposition of Φ which will summarize how its elements vary. Among the possibilities to do this, a singular value decomposition comes as a good choice since it does not require a special form of Φ (eigenvalue decomposition requires a square matrix). So the new criterion, which I denote as ΦIC, is defined as:

$$\Phi\text{IC} = -2\ln L + 2k + \lambda\delta_{\{\Phi\}},$$

where δ_x is the singular value of x. The notation $\{\Phi\}$ is to point out that the decomposition is done on a vector made with the elements of Φ and not on Φ itself. Like for other information-based criteria (see review in Konishi and Kitagawa, 2008), the model with the smallest value of ΦIC is selected as the best model describing the data.

2.5. Fitting algorithm

Molecular dating of phylogenies usually implies estimating a large number of parameters. For instance, fitting the penalised models to a fully dichotomous tree with n tips and a single calibration point will need to estimate $3n-4$ parameters ($2n-2$ rates and $n-2$ dates) whereas fitting the discrete categories model will imply estimating $n+2c-3$ parameters. For instance, if $n = 100$ (a medium-sized tree in today's standards) and $c = 1$, the number of free parameters will be 296 and 99, respectively. Giving the difficulties of optimising a function with so many parameters, Bayesian methods have clearly known a wide success.

Sanderson (2002) reported several algorithmic difficulties when optimising his penalised likelihood function. He used a gradient-free method as well as quasi-Newton gradient-based methods. Here, I use the PORT optimization routines (Gay, 1990) as implemented in R version 2.15.2 (R Development Core Team, 2012). The advantage of PORT is that they are relatively robust to irregularities in the objective function (by contrast to some other non-linear optimisation methods such as Schnabel et al., 1985). Furthermore, they can handle infinite values, and can be used with or without gradients (i.e., first partial derivatives) of the objective function. If these gradients are not provided, the routine computes them numerically using the objective function. If this one has

few parameters (less than ten), both methods (with or without gradients) have similar computing times; however, in the case of many parameters the gain in running times can be at least one order of magnitude. The gradients of the penalised likelihood functions used in this paper and their derivations are provided in Appendix A.

Another advantage of PORT is that bounds for parameter estimation can be set. This is used for dates known within intervals which are thus considered as additional free parameters to be estimated within these intervals. For the substitution rates, the lower and upper bounds are set to ε and $100 - \varepsilon$ where ε is by default set to 10^{-8} (this can be changed by the user). A convergence diagnostic is output by PORT which makes possible to assess the quality of the solution. It is also possible to control the number of iterations used during optimisation which appears useful here because convergence may be very slow to achieve due to the high dimensionality of the problem; so, the user may assess whether increasing the number of iterations could improve convergence.

Another control on the optimization process is the size of the steps around the current solution to examine whether a better solution can be found. Setting the minimum and the maximum step size correctly improves substantially the optimisation. In the present problem, there are two kinds of parameters, the rates and the dates, which typically have distinct ranges of variation: 10^{-8}–10^{-2} for the former, and 1–100 for the latter. Therefore, I have implemented an algorithm of alternate optimizations which can be summarised as follows:

1. A first set of estimates is obtained by optimising the objective function over all parameters. The step size is set to 1 (the default of PORT).
2. Optimise the objective function over the rates keeping the dates constant at the values of the current solution. The step size is allowed to vary between 10^{-8} and 0.1.
3. Optimise the objective function over the dates keeping the rates constant at the values output at step 2. The step size is allowed to vary between 10^{-3} and 0.5.
4. If the objective function value is improved, the solution found from the combination of steps 2 and 3 is taken as the new solution. These are repeated until convergence—or until a fixed number of alternate optimisations has been performed.

Finally, the unknown dates are initially set using a random algorithm which, when repeated several times, allows one to assess the importance of these initial values on the final results.

The methods presented in this paper have been implemented in ape (Paradis et al., 2004). The computing time of this new implementation was assessed with random trees of size 20, 50, 100, or 200 tips, simulated using the three models described in the next section, and fitting to each of them four different models: the "relaxed" model, the correlated model, the clock model ($c = 1$), and a discrete model with $c = 10$. The number of calibration points were 1, 3, 5, or 10 as described below. For each combination of these components, the simulation was replicated ten times resulting in 1920 recorded running times which were measured with the R function `system.time` on a modern laptop (hardware: CPU duo-core 2.93 GHz; software: Ubuntu 12.04 64-bit, R 2.15.2, ape 3.0-6). A regression analysis was performed to characterise the most important factors affecting computing times. The best model can then be used to predict the computing times for larger data sets. Note that this simulation protocol is almost similar to the one described below with the difference that one additional model was fitted and fewer replications were used.

2.6. Simulation study

Ultrametric trees were simulated with random topologies generated with the `rtree` function in ape (Paradis, 2012). In order to not bias branch lengths in a particular direction, the branching times of these trees were drawn from a uniform distribution between 1 and 50. The number of tips were equal to 20, 50, 100, or 200. This was replicated 100 times so that 400 initial ultrametric trees were generated. From each of them, three trees were simulated with different models of variation in the substitution rates: (i) the rates were generated from a Γ distribution with shape $\alpha = 0.5$ and rate $\rho = 1$; (ii) the rates were simulated along the initial tree with a Brownian motion model with variance $\sigma^2 = 0.01$ and rescaled to have a similar range of variation than in the previous model; (iii) a clock model by adding to each branch length a small noise generated from a normal distribution with mean 0 and variance 0.5. In the first two models, the branch lengths of the initial trees were multiplied by the simulated rates. Thus, 1200 such trees were generated.

For each of the 400 initial trees, a set of ten calibration points was made by extracting the branching times of the root node (which was always included in the subsequent analyses) and nine randomly chosen nodes. The 1200 non-ultrametric trees were analysed by fitting three models: "relaxed", correlated, and clock (with $c = 1$) using 1, 3, 5, or 10 calibration points (which were considered as exactly known dates), so 4800 data sets were analysed (1200 trees $\times$ 4 sets

of calibration points). In the end, 14,400 chronograms were estimated (4800 data sets $\times$ 3 models). All trees, chronograms, and sets of calibration points were saved for further analyses. For each estimated chronogram, the mean of the difference between the real and the estimated branch lengths and the mean of the difference between the real and the estimated substitution rates were computed.

3. Results

The 1920 computing times varied between 0.017 and 325.777 seconds (mean = 20.71, median = 2.42, SD = 45.28). Table 1 presents a summary of computing times for the considered tree sizes. The longest computing time (5 min 26 sec) was obtained fitting a clock model to a tree with $n = 200$ simulated with a correlated model using five calibration points. Remarkably, fitting the same model to the same tree with 10 calibration points was much faster (< 2 sec). However, the computing times had a very asymmetric distribution with a mean much higher than the median. A relatively complex linear model was found to fit well:

$$\ln T \propto \ln n + Ncal + Model + Sim + n : Model$$

with T the computing time, $Ncal$ the number of calibration points, $Model$ a categorical variable specifying the fitted model, Sim a categorical variable specifying the simulated model, and $n : Model$ the first-order interaction term between these two variables. The effect of $\ln n$ was positive (estimated coefficient: 2.18, SE = 0.09) while the effect of $Ncal$ was negative (-0.061, SE = 0.007). Interestingly, n was a better predictor than the number of estimated parameters (k which is highly correlated with n). The explained variance was high (adjusted $R^2 = 0.79$), and the residual standard-deviation was $\hat{\sigma} = 1.059$ (as a comparison the standard-deviation of $\ln T$ was 2.32). An examination of the residuals of this model showed a relatively homogeneous distribution suggesting a satisfactory overall fit (see Supplementary Information). Giving values of n, $Ncal$, $Model$, and Sim it is possible to predict the mean computing time, $\mathbb{E}(T)$, taking the exponential of the above linear model. Because of the logarithmic transformation of T, a 95% prediction interval has to be computed with $\mathbb{E}(T)/\mathrm{e}^{1.96\hat{\sigma}}$ and $\mathbb{E}(T) \times \mathrm{e}^{1.96\hat{\sigma}}$ instead of the usual $\pm 1.96\hat{\sigma}$. An easy-to-use R program is provided in the Supplementary Information to perform these calculations. For instance,

for $n = 1000$ and $Ncal = 1$, $\mathbb{E}(T)$ = 7 min 38 sec for fitting a clock model on data following the same model (95% prediction interval: [57 sec, 1 h 1 min]). On the other hand, for the same values of n and $Ncal$, if the data come from a correlated model and one wishes to fit the same model, then $\mathbb{E}(T)$ = 34 h ([4 h 16 min, 271 h 6 min]).

Figure 1 shows one of the 400 sets of simulated trees. Among the 4800 simulated data sets, 174 could not be analysed completely because of lack of convergence of the model fitting procedure. Overall, the ΦIC selected the correct model in 61% of the cases; however, this percentage depended on the simulated model since when the simulated model was the "relaxed" one the correct model was selected in 82.7%, and the percentage reached 100% when the simulated model was the clock one (Table 2).

Overall, and somehow surprisingly, the results were not affected by the size of the tree, so they are presented below pooled over the different values of n. Figure 2 shows the distribution of the error on branch length estimation with respect to whether the correct model was used for fitting the chronogram and the number of calibration points. The eight boxes reveal no bias in these estimations even when the wrong model was fitted. On the other hand, using the correct model substantially decreased the dispersion of the estimates. Without surprise, increasing the number of calibration points led to better estimates even with the wrong model.

Figure 3 displays the same results broken down with respect to the nine combinations of simulated–fitted models. Overall, the clock model gave good estimates of branch lengths even if the trees were simulated with other models. When the data were simulated with a clock model, the fit with this model yielded almost unbiased branch length estimates with a very small variance (rightmost panel).

Like for the branch lengths, the mean error rate on the estimation of substitution rates was not affected by tree size. No bias was observed in these estimates; however, their dispersion was reduced when fitting the correct model and/or when the number of calibration points increased (Fig. 4). Figure 5 shows how this error varied with respect to the simulated and fitted models. In most cases, no bias was observed. The estimates from the "relaxed" model were slightly positively biased (i.e., overestimated) when the trees were simulated with another model (first box in second and third panels). Whatever the simulated model, a surprising result appears: the correlated model seems to estimate substitution rates better than the two other models, and it performed almost as

well as the clock model when the data were clock-like. The dispersion of these errors were much larger when fitting a "relaxed" model compared to the two others though this was much reduced with the data simulated from this model in spite of a number of outliers with positive errors.

4. Discussion

Molecular dating is a fundamental step in evolutionary analysis. It is used to assess the tempo of phenotypic evolution (Garland, 1992), the rates of lineage diversification (Nee et al., 1992), or to date the colonization of new areas or habitats (Warren et al., 2003). Currently, two classes of methods are widely used: Bayesian methods, which are by far the most widespread with the computer programs Multidivtime (Thorne and Kishino, 2002), BEAST (Drummond and Rambaut, 2007), and PAML (Yang, 2007), and the penalised likelihood method as implemented in the program r8s (Sanderson, 2003). Other methods have been proposed: mean path length (Britton et al., 2002), likelihood methods (e.g., Yang, 2004), or least squares using distances as input data (Xia and Yang, 2011).

In spite of this diversity of methods, there is apparently a confusion among practitioners about the relative merits of each method. Probably one reason for this is the scarcity of assessment of the statistical performance of these methods. Brown and Yang (2011) compared the performance of strict and relaxed clock methods by simulating phylogenies with $n = 5$, 10, or 20 species. To my knowledge, the present paper is the first extensive study of the statistical properties of a molecular dating method in realistic conditions of data analysis.

Computing time is clearly an issue in data analysis. In a recent paper, dos Reis and Yang (2011) stated "a typical Bayesian analysis for a phylogeny of $<$50 species might take several days." Such long calculations are at the expense of the time spent in the interpretation and understanding of the results. The computing times recorded in the present study were always less than six minutes, but this masks a great heterogeneity since 50% of these, regardless of the tree size, were less than or equal to 2.42 seconds. Furthermore, the linear regression analysis, characterising almost 80% of the variation among the recorded computing times, showed that they are affected by a number of factors including some interactions among them. The possibility to fit several models in reasonable times is clearly a strength for the present likelihood and penalised likelihood methods. Besides,

short computing times make easier the assessment of statistical properties with simulations as done here.

Several results emerge from the present simulation study. First, using the correct model (i.e., the one used to simulate the data) gives better estimates of branch lengths. Second, using more calibration points results also in better estimation. Third, and this is more surprising, using the strict clock model gives good estimates of branch lengths whatever the model used to simulate the data. A possible explanation for this result may come from the way substitution rates were simulated. These rates were generated with stable distributions, so that in the estimation process replacing those rates by their mean value could result in easier estimation, especially considering that the clock model has less parameters to estimate than the two penalised models.

The fact that the number of tips n in the tree did not affect estimation may seem surprising at first because bigger trees have more observations (the x_i's in Eq. 2). However, the number of parameters to estimate for all the models considered here is proportional to n, so increasing the size of the tree does not actually increase the quantity of information available for estimation.

The simulation study reported here considered that the input phylogenetic trees are correctly estimated. In reality this may not be always the case, but the goal here was to focus on some issues that have not been studied in-depth until now. Moreover, the simulation approach used here is similar to the one used in other studies (e.g., Brown and Yang, 2011). Another assumption of the present simulations is that the calibration points are known exactly and without error. This also is not likely to be true in real situations. This point has been extensively discussed in the literature. For instance, Yang and Rannala (2006) have shown that, in the case of Bayesian estimation, incorrectly assumed "hard bounds" (exactly known ages) will lead to wrong results whatever the quantity of sequence data.

A point that will need future attention is how the present models behave in the presence of more complex variation in substitution rates. Some studies have suggested that these rates may vary in a non-equilibrium fashion (e.g., Magallón, 2010), and others suggest a relationship with life-history traits (Lartillot and Poujol, 2011; Mayrose and Otto, 2011). Since the penalised models estimate a different substitution rate for each branch, they seem appropriate to handle more complex cases than considered here. This clearly requires further investigation.

The present study clealy calls for a wider assessment of the statistical properties of the methods

developed here in comparison with other methods. This will require a much larger simulation survey (e.g., Wertheim et al., 2010). An issue that also needs more attention is the characterisation of the relative merits of different methods. For instance, the mean path lengths method (Britton et al., 2002) provides a test of the molecular clock for each node of the tree. Whether this is a valuable tool for preliminary explorative analyses before running more complex model fitting, with likelihood or Bayesian models, is not yet clear.

A point of the present study that will require further study is how to select the best model in a penalised likelihood framework. The proposed information criterion, ΦIC, gave some equivocal results. The criterion correctly selected the "relaxed" model in the majority of cases which appears as a positive result. On the other hand, when the data were simulated with the correlated model, the ΦIC selected the clock model in most cases. This may be understood easily because fitting the correlated model tends to minimise the difference in substitution rates among contiguous branches, thus tending to a clock model. In other words, if the rate of autocorrelation is strong, the variation among branches will be low so the observed variation will tend to a clock model. On the other hand, if the autocorrelation is weak, then the rates will not be serially correlated among branches and they will tend to vary in a "relaxed" way. This may explain why the correlated model is likely to not be selected by the proposed criterion. A possible solution to this problem would be to use the ΦIC criterion only to compare the "relaxed" and the clock models. A related issue, which was left aside for the present study, is that the amount of smoothing is actually controlled by a parameter, denoted as λ by Sanderson (2002), which was fixed to one in all the results reported here. This clearly needs further study. Furthermore, Kim and Sanderson (2008) showed some difficulties in finding the best value of λ by cross-validation and called for the development of a better scheme to this end.

The present study aimed to investigate how Sanderson's penalised likelihood approach could be improved by extending it to other models of substitution rate variation. Some computational improvements have also been done, and a new information criterion for model selection has been developed. These improvements provide a framework for molecular dating of phylogenies which seems an attractive alternative to Bayesian methods, particularly considering the possibility to analyse large trees ($n = 200$) in short times (a few minutes). Four perspectives may be drawn from this work. First, the interplay between the clock and correlated models may lead to improved

estimation. Since the clock model provided the best estimates of branch lengths, and the correlated model provided the best estimates of substitution rates, both models could be combined. Second, the derivation of standard-errors of the estimates was left aside for the present study. As done elsewhere (Xia and Yang, 2011), bootstrap or other resampling methods can be used. Another approach might be to use profile likelihood-based methods to infer uncertainty of the parameter estimates. Because of the penalty term, this does not seem straightforward and is currently under study. Third, it will be necessary to study the issue of the smoothing parameter λ and its role in parameter estimation. Finally, the statistical properties and scope of the proposed information criterion ΦIC will need to be further investigated.

Acknowledgements

I am grateful to two anonymous reviewers and the Associate Editor for their constructive comments on a previous version of my manuscript. The simulations benefited from the ISEM computing cluster platform. This is publication IRD-ISEM 201x-xxx.

Appendix A. Derivation of gradients and Hessian

Gradients

The objective function is made of the difference of two functions, so the derivatives can be decomposed in two parts (with θ_i being any free parameter of the model):

$$\frac{\partial \Psi}{\partial \theta_i} = \frac{\partial \ln L}{\partial \theta_i} - \frac{\partial \Phi}{\partial \theta_i}.$$

So we consider first the log-likelihood part:

$$\ln L = \sum_i x_i \ln \zeta_i - \zeta_i - \ln x_i!,$$

where the index i is for all branches of the tree. Since $\zeta_i = r_i t_i$, this has to be written as:

$$\ln L = \sum_i x_i \ln r_i + x_i \ln t_i - r_i t_i - \ln x_i!.$$

We note that t_i, the length of branch i, is not a parameter and is given by $t_i = d_{a_i} - d_{b_i}$ where the d's are dates (measured from present back to the past) and a_i is the node ancestor of node b_i along branch i. The dates are parameters unless they are fixed calibration points.

We can now derive the first partial derivatives of $\ln L$ which is simple for the rates:

$$\frac{\partial \ln L}{\partial r_i} = \frac{x_i}{r_i} - t_i,$$

and slightly more complicated for the dates. We first rewrite the log-likelihood as:

$$\ln L = \sum_i x_i \ln r_i + x_i \ln(d_{a_i} - d_{b_i}) - r_i(d_{a_i} - d_{b_i}) - \ln x_i!,$$

which makes clearer how we obtain the derivates with respect to the unknown dates (this uses $\ln(u)' = u'/u$):

$$\frac{\partial \ln L}{\partial d_k} = \sum_i \left(\frac{x_i}{t_i} - r_i\right) \mathrm{A}(k,i) - \left(\frac{x_i}{t_i} - r_i\right) \mathrm{D}(k,i),$$

where the index k is for all unknown dates, $\mathrm{A}(k,i)$ is an indicator function taking the value 1 if k is a node ancestor (basal) of branch i or 0 otherwise, and similarly for $\mathrm{D}(k,i)$ if k is a node descendant (terminal). The advantage of this formulation, from a computational point of view, is that A and D can be built once as incidence matrices and the above equation is then calculated with two matrix products after calculating the terms within parentheses for all branches.

We now consider the penalty term Φ which differs depending on the model. Only the rates are concerned here since the dates are not involved in the calculation of Φ.

For the auto-correlated model, we rewrite Φ as:

$$\Phi = \sum_{k,j} r_k^2 - 2 r_k r_j + r_j^2 + \mathrm{var}(r_{\mathrm{basal}}),$$

reminding that (k, j) is for all pairs of contiguous branches. So for branch i we have (ignoring the variance term for the moment):

$$\frac{\partial \Phi}{\partial r_i} = \sum_{i,j} (2r_i - 2r_j) + \sum_{k,i} (-2r_k + 2r_i).$$

The first sum is for all pairs (i,j) where i is an ancestral branch connected to branch j, and the second sum has only one term for the pair (k,i) with k being the ancestral branch of i. So, this can be simplified as:

$$\frac{\partial\Phi}{\partial r_i} = 2\left(\eta_i r_i - \sum_j^{\eta_i} r_j\right),$$

where η_i is the number of branches connected to branch i ($\eta_i = 1$ for the terminal branches), and the sum is now over all branches connected to i (ancestral or not, but not sister-ones). Like discussed above, this can be computed in a straightforward way since the η_i's can be built once, and also an incidence matrix indicating whether two branches are connected.

For the basal branches (i.e., connected to the root), a term related to the variance of their rates must be added to the above, that is for branch i:

$$2\frac{r_i\left(1-\frac{1}{n_{\text{basal}}}\right) - \frac{1}{n_{\text{basal}}}\sum_{j\neq i}^{n_{\text{basal}}} r_j}{n_{\text{basal}}-1},$$

where i and j is for the basal branches and n_{basal} is their number. This simplifies to $r_i - r_j$ if $n_{\text{basal}} = 2$.

For the "relaxed" model, we use the fact that by definition the derivative of the empirical cumulative distribution function, which is a step function, is zero, and the derivative of the theoretical cumulative distribution function $F(x)$ is, also by definition, the density function $f(x)$. So, using $(u^2)' = 2u'u$:

$$\frac{\partial\Phi}{\partial r_i} = -2f(r_i)\left[\tilde{\mathcal{F}}(r_i) - F(r_i)\right]$$

Hessian

The matrix of second partial derivatives is obtained by deriving the first partial derivatives with respect to all parameters:

$$\frac{\partial^2\Psi}{\partial\theta_i\partial\theta_j}.$$

The Hessian is a square matrix with its diagonal elements being $\partial^2\Psi/\partial\theta_i^2$. As above, me way decompose for the two components of Ψ, so:

$$\frac{\partial^2 \ln L}{\partial r_i^2} = -\frac{x_i}{r_i^2} \qquad \frac{\partial^2 \ln L}{\partial r_i \partial r_j} = 0 \quad i \neq j.$$

Rewriting the above first derivative as $\partial \ln L/\partial r_i = x_i/r_i - (d_{a_i} - d_{b_i})$, we easily find:

$$\frac{\partial^2 \ln L}{\partial r_i \partial d_k} = -\mathrm{A}(k,i) + \mathrm{D}(k,i).$$

We now turn to the first partial derivatives with respect to d and derive them a second time with respect to r:

$$\frac{\partial^2 \ln L}{\partial d_k \partial r_i} = -\mathrm{A}(k,i) + \mathrm{D}(k,i).$$

The second partial derivatives with respect to d are more complicated to derive (this uses $(1/u)' = -u'/u^2$):

$$\frac{\partial^2 \ln L}{\partial d_k^2} = \sum_i \frac{x_i}{t_i^2}\mathrm{A}(k,i) + \frac{x_i}{t_i^2}\mathrm{D}(k,i),$$

$$\frac{\partial^2 \ln L}{\partial d_k \partial d_l} = \sum_i -\frac{x_i}{t_i^2}\mathrm{A}(k,i)\mathrm{D}(l,i) - \frac{x_i}{t_i^2}\mathrm{A}(l,i)\mathrm{D}(k,i) \quad k \neq l.$$

However, this last equation is simpler, since this is equivalent to consider only the branch i defined by nodes k and l and thus simplifies to $-x_i/t_i^2$. This is because, either $\mathrm{A}(k,i)\mathrm{D}(l,i)$ is equal to 1 only if k is the ancestor of l along i, or $\mathrm{A}(l,i)\mathrm{D}(k,i)$ is equal to 1 only if l is the ancestor of k along i.

We now consider Φ. For the auto-correlated model, we have:

$$\frac{\partial^2 \Phi}{\partial r_i^2} = 2\eta_i \quad \frac{\partial^2 \Phi}{\partial r_i \partial r_j} = -2,$$

where j are the branches connected to i (see above). Clearly, these depend only on the η_i's and so

do not need to be updated during optimisation.

For the "relaxed" model, the generic form may be written as:

$$\begin{aligned}\frac{\partial^2\Phi}{\partial r_i^2} &= -2\frac{\partial f}{\partial r_i}\tilde{\mathcal{F}}(r_i) + 2\frac{\partial f}{\partial r_i}F(r_i) + 2f(r_i)^2, \\ &= -2\frac{\partial f}{\partial r_i}\left[\tilde{\mathcal{F}}(r_i) - F(r_i)\right].\end{aligned}$$

Because of the difficulty in deriving the density function f, this is not considered further.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B. N., Csaki, F. (Eds.), Proceedings of the Second International Symposium on Information Theory. Akadémia Kiado, Budapest, pp. 267–281.

Britton, T., Oxelman, B., Vinnersten, A., Bremer, K., 2002. Phylogenetic dating with confidence intervals using mean path lengths. Mol. Phyl. Evol. 24 (1), 58–65.

Brown, R. P., Yang, Z., 2011. Rate variation and estimation of divergence times using strict and relaxed clocks. BMC Evol. Biol. 11, 271.

dos Reis, M., Yang, Z. H., 2011. Approximate likelihood calculation on a phylogeny for Bayesian estimation of divergence times. Mol. Biol. Evol. 28 (7), 2161–2172.

Drummond, A. J., Ho, S. Y. W., Phillips, M. J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biology 4 (5), 699–710.

Drummond, A. J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol. Biol. 7, 214.

Garland, Jr, T., 1992. Rate tests for phenotypic evolution using phylogenetically independent contrasts. Am. Nat. 140 (3), 509–519.

Gay, D. M., 1990. Usage summary for selected optimization routines. Tech. Rep. Computing Science Technical Report No. 153, AT&T Bell Laboratories, Murray Hill, NJ 07974, USA. URL `http://netlib.bell-labs.com/netlib/port/`

Guindon, S., 2010. Bayesian estimation of divergence times from large sequence alignments. Mol.

Biol. Evol. 27 (8), 1768–1781.

Ho, S. Y. W., Lanfear, R., 2010. Improved characterisation of among-lineage rate variation in cetacean mitogenomes using codon-partitioned relaxed clocks. Mitochondrial DNA 21 (3-4), 138–146.

Kim, J., Sanderson, M. J., 2008. Penalized likelihood phylogenetic inference: bridging the parsimony-likelihood gap. Syst. Biol. 57 (5), 665–674.

Konishi, S., Kitagawa, G., 2008. Information Criteria and Statistical Modeling. Springer, New York.

Lartillot, N., Lepage, T., Blanquart, S., 2009. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. Bioinformatics 25 (17), 2286–2288.

Lartillot, N., Poujol, R., 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28 (1), 729–744.

Lepage, T., Bryant, D., Philippe, H., Lartillot, N., 2007. A general comparison of relaxed molecular clock models. Mol. Biol. Evol. 24 (12), 2669–2680.

Magallón, S., 2010. Fossils to break long branches in molecular dating: a comparison of relaxed clocks applied to the origin of angiosperms. Syst. Biol. 59 (4), 384–399.

Mayrose, I., Otto, S. P., 2011. A likelihood method for detecting trait-dependent shifts in the rate of molecular evolution. Mol. Biol. Evol. 28 (1), 759–770.

Nee, S., Mooers, A. Ø., Harvey, P. H., 1992. Tempo and mode of evolution revealed from molecular phylogenies. Proc. Natl. Acad. Sci. USA 89, 8322–8326.

Paradis, E., 2012. Analysis of Phylogenetics and Evolution with R (Second Edition). Springer, New York.

Paradis, E., Claude, J., Strimmer, K., 2004. APE: analyses of phylogenetics and evolution in R language. Bioinformatics 20 (2), 289–290.

Pereira, S. L., Baker, A. J., 2006. A mitogenomic timescale for birds detects variable phylogenetic rates of molecular evolution and refutes the standard molecular clock. Mol. Biol. Evol. 23 (9), 1731–1740.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, iSBN 3-900051-07-0.
URL `http://www.R-project.org`

Sanderson, M. J., 1997. A nonparametric approach to estimating divergence times in the absence of rate constancy. Mol. Biol. Evol. 14 (12), 1218–1231.

Sanderson, M. J., 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19 (1), 101–109.

Sanderson, M. J., 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. Bioinformatics 19 (2), 301–302.

Schnabel, R. B., Koontz, J. E., Weiss, B. E., 1985. A modular system of algorithms for unconstrained minimization. ACM Trans. Math. Software 11 (4), 419–440.

Thorne, J. L., Kishino, H., 2002. Divergence time and evolutionary rate estimation with multilocus data. Syst. Biol. 51 (5), 689–702.

Thorne, J. L., Kishino, H., Painter, I. S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15 (12), 1647–1657.

Warren, B. H., Bermingham, E., Bowie, R. C. K., Prys-Jones, R. P., Thébaud, C., 2003. Molecular phylogeography reveals island colonization history and diversification of western Indian Ocean sunbirds (*Nectarinia*: Nectariniidea). Mol. Phyl. Evol. 29 (1), 67–85.

Wertheim, J. O., Sanderson, M. J., Worobey, M., Bjork, A., 2010. Relaxed molecular clocks, the bias–variance trade-off, and the quality of phylogenetic inference. Syst. Biol. 59 (1), 1–8.

Wilkinson, R. D., Steiper, M. E., Soligo, C., Martin, R. D., Yang, Z. H., Tavaré, S., 2011. Dating primate divergences through an integrated analysis of palaeontological and molecular data. Syst. Biol. 60 (1), 16–31.

Xia, X., Yang, Q., 2011. A distance-based least-square method for dating speciation events. Mol. Phyl. Evol. 59 (2), 342–353.

Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39 (3), 306–314.

Yang, Z., 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. Acta Zoologica Sinica 50 (4), 645–656.

Yang, Z., 2007. PAML 4: Phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24 (8), 1586–1591.

Yang, Z., Rannala, B., 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23 (1), 212–226.

Table 1: Mean, median, standard-deviation (SD), minimum, and maximum from a sample of 1920 computing times (in seconds). *n*: number of taxa.

n	Mean	Median	SD	Min	Max
20	0.67	0.14	2.04	0.02	16.04
50	3.71	1.10	8.06	0.10	69.84
100	15.65	5.41	22.73	0.35	170.13
200	62.84	26.49	71.65	1.39	325.78

Table 2: Number of times when the fitted model was selected with ΦIC with respect to the simulated model.

Simulated model	Selected model		
	"Relaxed"	Correlated	Clock
"Relaxed"	1276	3	263
Correlated	292	0	1250
Clock	0	0	1542

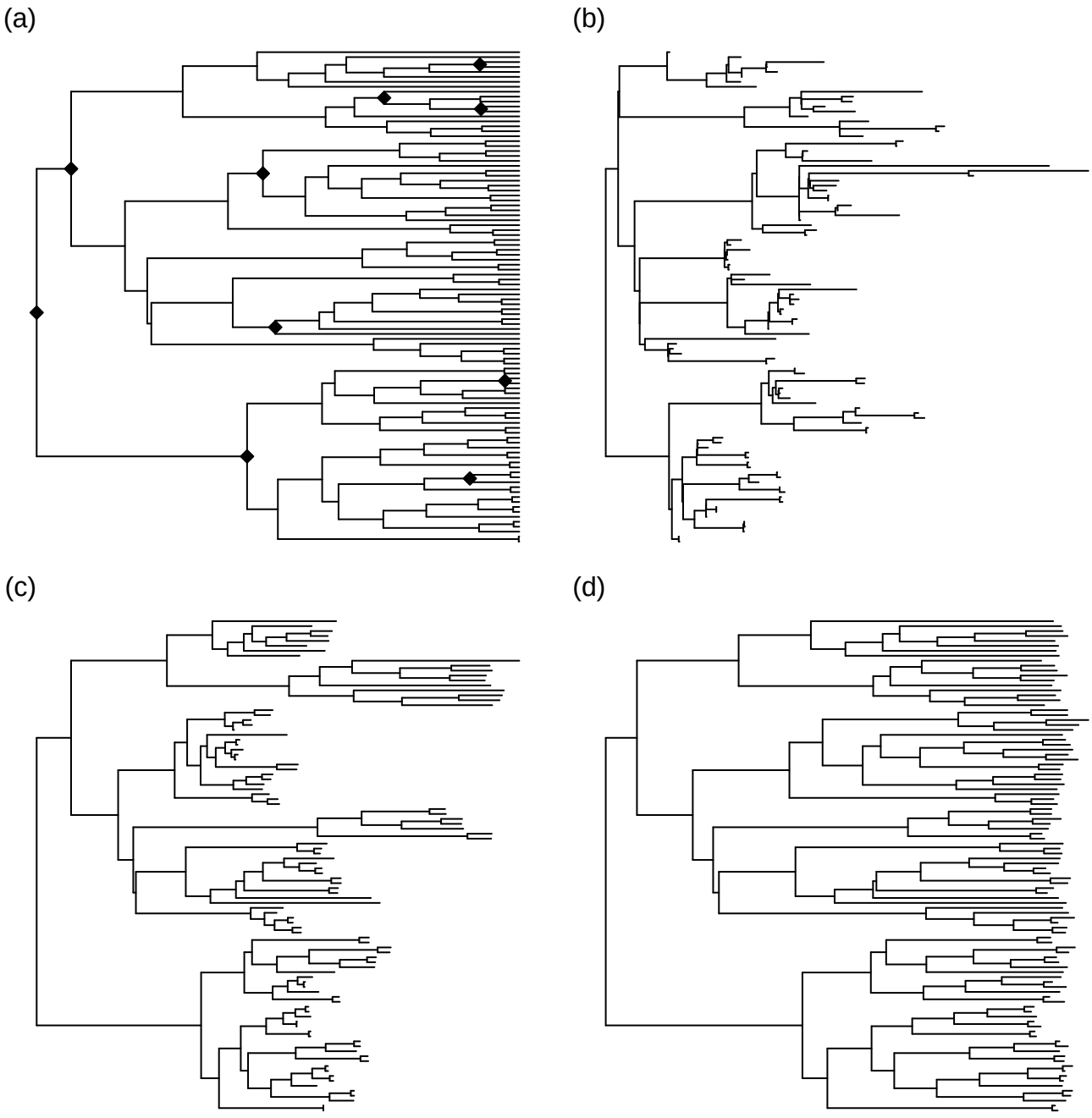
Fig. 1. (a) One of the 400 random ultrametic trees ($n = 100$). The diamonds indicate the selected calibration points. The other trees were generated from this one as explained in the text: (b) under the "relaxed" model with rates following a Γ distribution, (c) under the correlated model, and (d) under the clock model.

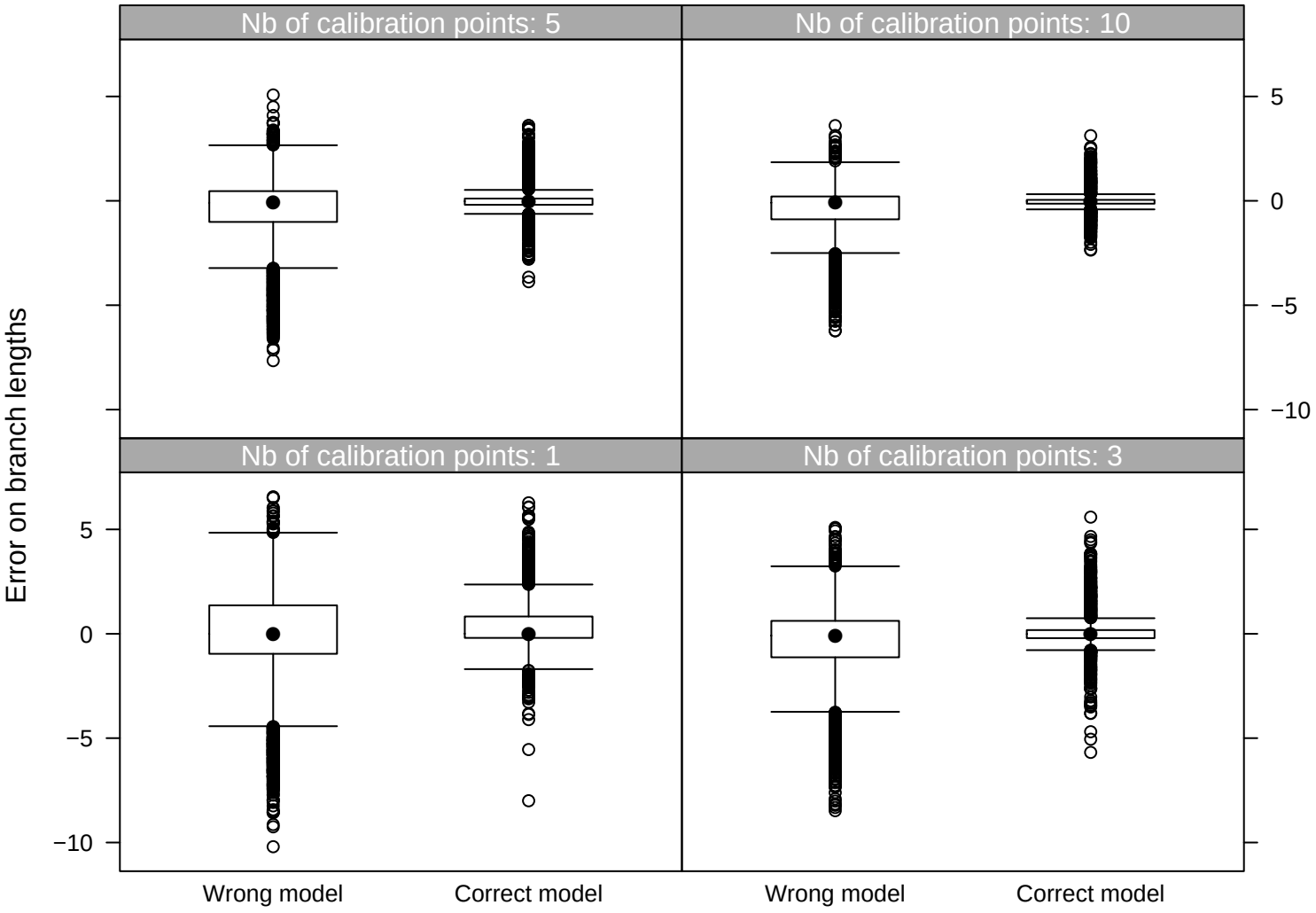
Fig. 2. Distribution of the mean error on branch length estimation with respect to whether the correct model was used for fitting the chronogram (x-axis) and the number of calibration points (panels). The mean error was calculated as the difference of the estimated and the true values of the parameters averaged over the whole tree. The box limits give the first and third quartiles and the whiskers extend to 1.5 times the box limits. The extreme values going beyond the whiskers are shown with open circles. The filled circle indicates the median.

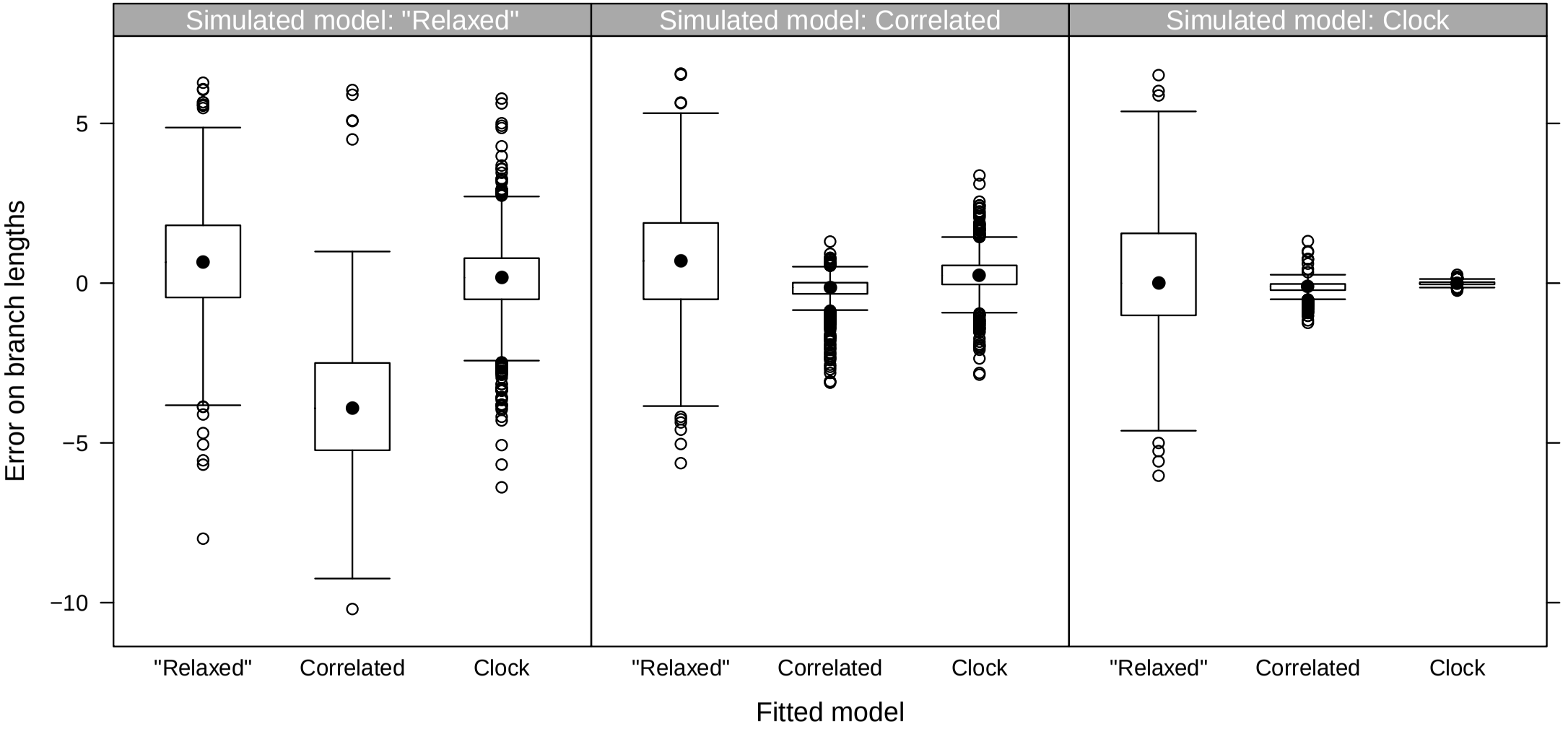
Fig. 3. Distribution of the mean error on branch length estimation with respect to the fitted model of the chronogram (x-axis) and the model used to simulate the data (panels). See Fig. 2 for details.

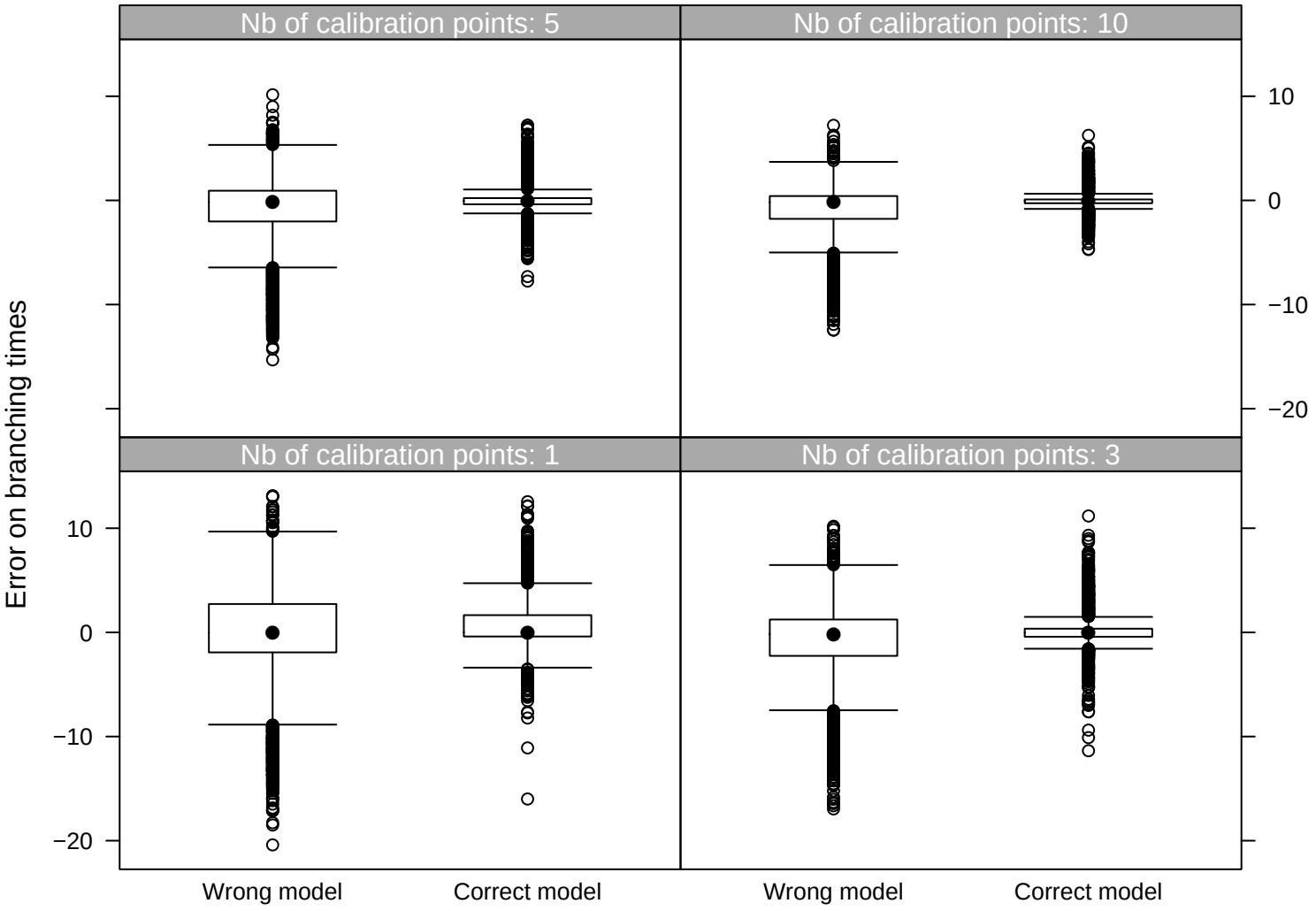
Fig. 4. Distribution of the mean error on substitution rate estimation with respect to whether the correct model was used for fitting the chronogram (x-axis) and the number of calibration points (panels). See Fig. 2 for details.

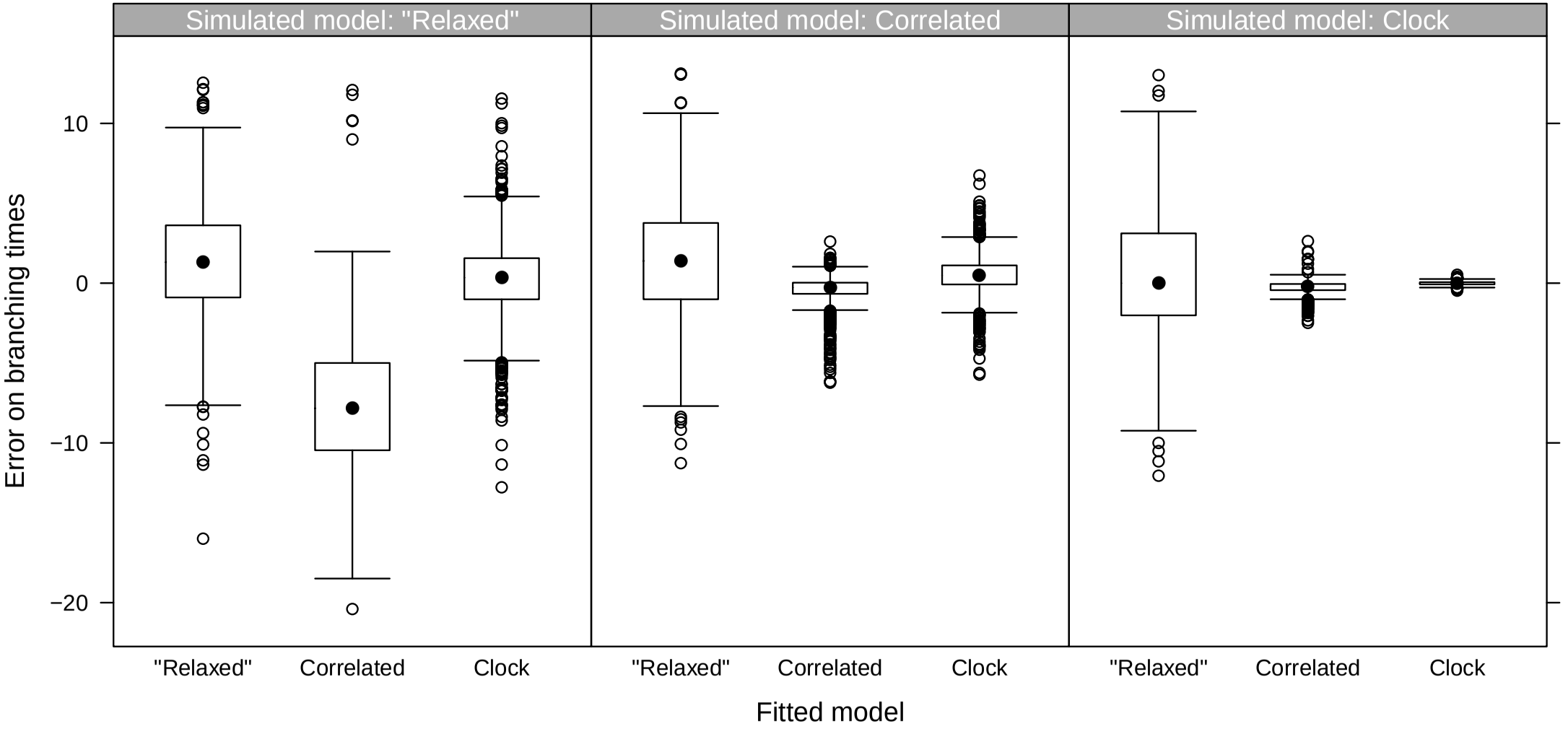
Fig. 5. Distribution of the mean error on substitution rate estimation with respect to the fitted model of the chronogram (x-axis) and the model used to simulate the data (panels). See Fig. 2 for details.


(a)
(b)
(c)
(d)


Nb of calibration points: 5
Nb of calibration points: 10
Nb of calibration points: 1
Nb of calibration points: 3
Error on branch lengths
5
0
−5
−10
Wrong model
Correct model
Wrong model
Correct model


Simulated model: "Relaxed"
Simulated model: Correlated
Simulated model: Clock
Error on branch lengths
5
0
−5
−10
"Relaxed"
Correlated
Clock
"Relaxed"
Correlated
Clock
"Relaxed"
Correlated
Clock
Fitted model


Nb of calibration points: 5
Nb of calibration points: 10
Nb of calibration points: 1
Nb of calibration points: 3
Error on branching times
10
0
−10
−20
Wrong model
Correct model
Wrong model
Correct model


Simulated model: "Relaxed"
Simulated model: Correlated
Simulated model: Clock
Error on branching times
10
0
−10
−20
"Relaxed"
Correlated
Clock
"Relaxed"
Correlated
Clock
"Relaxed"
Correlated
Clock
Fitted model