Online journal ISSN - 1676-5680
GMR Genetics and Molecular Research
EVOLUTION AND TECHNOLOGY

An empirical examination of the standard errors of maximum likelihood phylogenetic parameters under the molecular clock via bootstrapping

Carlos G. Schrago

Laboratório de Biodiversidade Molecular, Departamento de Genética, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil
Corresponding author: C.G. Schrago
E-mail: guerra@biologia.ufrj.br

Genet. Mol. Res. 5 (1): 233-241 (2006)
Received January 10, 2006
Accepted February 17, 2006
Published March 31, 2006

ABSTRACT. The molecular clock theory has greatly enlightened our understanding of macroevolutionary events. Maximum likelihood (ML) estimation of divergence times involves the adoption of fixed calibration points, and the confidence intervals associated with the estimates are generally very narrow. The credibility intervals are inferred assuming that the estimates are normally distributed, which may not be the case. Moreover, calculation of standard errors is usually carried out by the curvature method and is complicated by the difficulty in approximating second derivatives of the likelihood function. In this study, a standard primate phylogeny was used to examine the standard errors of ML estimates via the bootstrap method. Confidence intervals were also assessed from the posterior distribution of divergence times inferred via Bayesian Markov Chain Monte Carlo. For the primate topology under evaluation, no significant differences were found between the bootstrap and the curvature methods. Also, Bayesian confidence intervals were always wider than those obtained by ML.

Key words: Confidence intervals, Divergence time estimation

INTRODUCTION

The maximum likelihood (ML) method of phylogenetic reconstruction was formalized by Felsenstein (1981) and further developed by other authors (Kishino and Hasegawa, 1989; Kishino et al., 1990; Hasegawa et al., 1991). This method is known to perform very well under many circumstances in which other algorithms, such as parsimony and distance matrix-based ones (e.g., neighbor-joining), are error prone (Huelsenbeck, 1995). The good performance of ML has been attributed to the capacity for full incorporation of the phylogenetic information found in the data plus the inclusion of models of sequence evolution (Yang, 1994b).

Although the ML algorithm presents several advantages, it is computationally intensive, inhibiting its application when the number of sequences is large. Another delicate issue is the estimation of the standard errors (SEs) of ML estimates. This quantity is generally calculated by inverting the matrix of second derivatives of the likelihood function with respect to parameters (the curvature method, Edwards, 1972). In phylogenies, however, the number of parameters is often very large (e.g., branch lengths, transition/transversion rate ratio, α parameter of the gamma distribution, etc.), and the numerical approximation of second derivatives of the likelihood function may result in unreliable estimates (Yang, 1997). Moreover, ML estimates may not be approximated by the normal distribution, and thus the calculation of confidence intervals using the estimated SE and the normal density function would be biased.

In practice, however, this is not a problem. Researchers seldom use SEs of ML parameters in molecular phylogenetics, since the likelihood ratio is the preferred hypothesis testing procedure. But, in molecular clock studies, SEs are important to establish confidence intervals for divergence time estimates. Parametric estimates, such as divergence times, are biologically uninformative without associated errors. Hence, an evaluation of SEs and determination of the reliability of the normal approximation to the ML estimates are necessary. Such a study should be carried out by simulating sequences on topologies with known SEs. One could then test the curvature or any other method by verifying their accuracy in recovering correct SEs. Unfortunately, a simulation of this kind is difficult to perform in phylogenies.

Nevertheless, an independent measure of SE of the ML estimates can be obtained. This can be achieved by means of the bootstrap (Efron, 1979), which is a widely used statistical method to infer confidence intervals for parameters that have been successfully applied in molecular evolution and phylogenetics (Felsenstein, 1985; Nei and Kumar, 2001). Although one cannot test the systematic errors of a method by using another method, the bootstrap technique does not use the curvature of the likelihood surface to inform about the credibility of a parameter. Therefore, if the curvature and the bootstrap methods calculate identical SEs, it would indicate that the numerical approximation of second derivatives was accurate, since it is unlikely to wrongly infer the same value of SE twice.

In the present study, the bootstrap is used to infer the SEs of parameters estimated by ML of a standard primate phylogeny. A comparison was made between the SEs estimated by the curvature method and those obtained by the bootstrap technique. The confidence intervals inferred by the normal approximation to the ML estimates with those calculated using bootstrapping were also compared. Finally, 95% credibility intervals were also assessed by the recent Bayesian Markov Chain Monte Carlo (MCMC) method of Thorne and collaborators (Thorne et al., 1998; Kishino et al., 2001; Thorne and Kishino, 2002).

METHODS

The curvature method

Given an alignment X of s species and n sites, the likelihood for site h conditional on parameter θ is $\ell(x_h|\theta)$, and is estimated by the algorithm of Felsenstein (1981). The likelihood of the given alignment is then

$$L(\theta) = \prod_{h=1}^{n} \ell(x_h|\theta), \text{ or } \ln L(\theta) = \sum_{h-1}^{n} \ln \ell(x_h|\theta).$$

The ML estimate $\hat{\theta}$ is obtained by $\partial L/\partial\theta = 0$.

The usual method of communicating information about parameter θ is to obtain the second-order Taylor series approximation to the support curve (the $\ln L \times \theta$ plot) at $\hat{\theta}$:

$$L(\theta) = L(\hat{\theta}) + (\theta - \hat{\theta})\frac{\partial L}{\partial \theta} + 1/2(\theta - \hat{\theta})^2 \frac{\partial^2 L}{\partial \theta^2},$$

which becomes

$$L(\theta) = L(\hat{\theta}) + 1/2(\theta - \hat{\theta})^2 \frac{\partial^2 L}{\partial \theta^2}, \text{ since } \left.\frac{\partial L}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0.$$

Then, the second partial differential coefficients of the above equation are used to calculate the observed formation of the curve near $\hat{\theta}$ (Edwards, 1972). This value is given by

$$w^2 = -\left[\frac{\partial^2 L}{\partial \theta^2}\right]^{-1}.$$

Since the observed formation offers a measure of the radius of the curvature near $\hat{\theta}$, its square root (w), also called the span of the support curve, is a measure of the standard error of the ML estimate $\hat{\theta}$ ($\hat{\theta} \pm w$). If we assume the normality of the ML estimate, the $(1 - \alpha) \times 100\%$ confidence interval for $\hat{\theta}$ is $\hat{\theta} \pm z_{\alpha/2}w$. For example, the 95% confidence interval is given by $\hat{\theta} \pm 1.96w$.

The bootstrap

Consider alignment X with n sites, from which ML estimation $\hat{\theta}$ is inferred by applying the algorithm of Felsenstein (1981) on the assumed phylogenetic tree. Then, a bootstrap replicate can be generated by resampling (with replacement) n sites from X (Felsenstein, 1985). The bootstrap standard error of $\hat{\theta}$ is estimated by

$$SE(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\hat{\theta}_i - \bar{\theta}_B)}.$$

where B is the number of replicates used, $\hat{\theta}_i$ is the ML estimation for replicate i, and $\bar{\theta}_B$ is the average of all B ML estimations of θ. The standard $(1 - \alpha) \times 100\%$ bootstrap confidence limit for $\hat{\theta}$ is $\hat{\theta} \pm z_{\alpha/2}SE(\hat{\theta})$. However, several other measures of bootstrap confidence limits have been introduced (Manly, 1997). In the present study, the percentile method of Hall (1992) was also used. This confidence interval is given by $\Pr(\theta - \varepsilon_H < \theta < \theta - \varepsilon_L) = 1 - \alpha$, where ε_H and ε_L (the error limits) are obtained from the $(1 - \alpha) \times 100\%$ confidence interval of the distribution of $\varepsilon_B = \hat{\theta}_B - \hat{\theta}$, the difference between the estimate for a bootstrap replicate and the sample estimate $\hat{\theta}$. The distribution of the error ε_B should be an approximation of the error in $\hat{\theta}$, i.e., $\varepsilon = \hat{\theta} - \theta$.

Empirical study

Sequences of the mitochondrial ND5 gene were retrieved from selected primate genomes available in GeneBank. Eight species were used in this study: *Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus*, *Hylobates lar*, *Macaca sylvanus*, *Papio hamadryas*, *Cebus albifrons*, and the outgroup *Lemur catta* (Figure 1). The phylogenetic relationship among these species is not debated by any molecular or morphological work; hence, the topology in Figure 1 was assumed. The final alignment was composed of 1,806 sites. The HKY85+Γ_5 model of sequence evolution was used; it corrects for unequal base frequencies, transition/transversion rate ratio and accommodates among site rate heterogeneity using the discrete approximation of the gamma distribution with five categories (Hasegawa et al., 1985; Yang, 1994a). The hominoid-cercopithecoid separation at 25 million years ago (Mya) was used as a calibration point (Schrago and Russo, 2003).

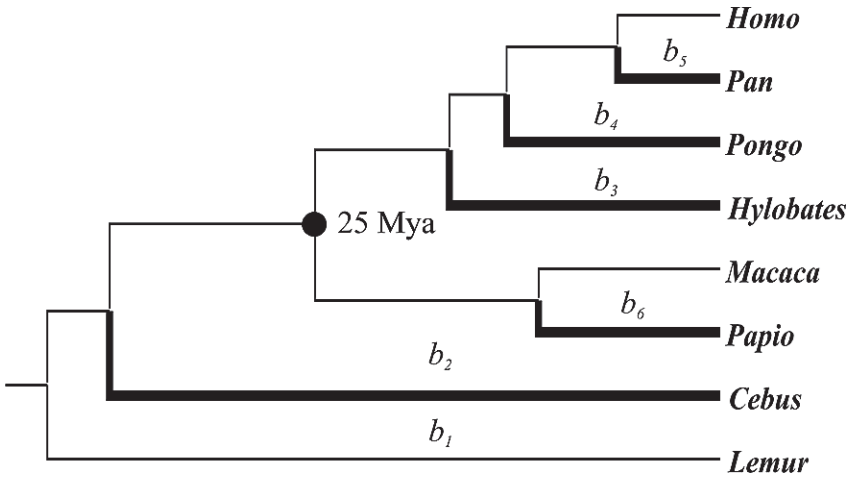

Figure 1. Standard primate phylogeny used in this study. The hominoid-cercopithecoid calibration point is depicted by the black circle. Branch lengths (times) are indicated following the classification used in the text. Mya = million years ago.

Since the molecular clock was assumed, the length of sister branches is constrained to be equal and, hence, a total of nine parameters were estimated by ML for the tree used: six branch lengths, i.e., times, the rate of molecular evolution μ (see Yang and Yoder, 2003), parameter κ (ts./tv. rate ratio) and parameter α of the gamma distribution. The likelihood function is thus $L(X \mid \Theta)$, with $\Theta = b_1,b_2,b_3,b_4,b_5,b_6, \mu, \kappa, \alpha)$. However, since the aim of the present study was to evaluate time parameters only; the results for parameters μ, κ, and α were ignored. The divergence time of the outgroup *L. catta* (b_1, Figure 1) was also ignored.

Maximum likelihood analyses were conducted in PAML 3.14 (Yang, 1997), and 1,000 bootstrap replicates of the original data set were also obtained in PAML. For each *B* replicate, the vector of parameters was estimated by ML ($\hat{\Theta}_B$). Then, bootstrap SEs and confidence intervals were calculated for each parameter using the procedure detailed in the previous section.

Further statistical analyses of the bootstrap samples were conducted in the R programming environment (www.r-project.org). Bayesian estimates of divergence times were obtained with the MULTIDISTRIBUTE program package (http://statgen.ncsu.edu/thorne/multidivtime.html). The Bayes method needs the adoption of maximum and minimum limits for the calibrations, instead of fixed points. Therefore, a narrow interval (24.5-25.5) was used to calibrate the hominoid-cercopithecoid divergence. This was done in order to make Bayes and ML estimates comparable, since ML uses fixed (25 Mya) calibration. The posterior distribution of divergence times was approximated by the MCMC algorithm. After a burn-in period of 50,000 generations, the Markov Chain was visited every 100 cycles, until 10,000 samples were taken.

RESULTS AND DISCUSSION

Bootstrap SEs were very close to those estimated by the curvature method (Table 1). The difference found between the two approaches was insignificant in practice. The highest difference value was found for parameter b_2 (0.05), the age of the New World primate split from Old World anthropoids, while the smallest differences were calculated for the smallest branches b_5 and b_6.

Table 1. Standard errors (SE) for maximum likelihood estimates (MLE) calculated by the curvature and the bootstrap.

Parameter	MLE	Bootstrap mean	Curvature SE	Bootstrap SE
b_2	37.704	37.754	2.288	2.338
b_3	16.745	16.799	1.003	1.048
b_4	13.185	13.199	0.928	0.974
b_5	6.301	6.333	0.599	0.584
b_6	11.216	11.197	0.894	0.882

The numerical calculation of inverted second derivatives with respect to parameters yielded essentially the same SEs as those obtained by the bootstrap techniques. It is important to notice that the iteration algorithm used to calculate derivatives is based on the difference ap-

proximation (Gill and Murray, 1981). In order to check the robustness of the estimates, the value of the variable that controls this approximation should be slightly changed (Yang, 1997). In this study, curvature SEs were identically estimated independent of the fine-tuning conditions of the algorithm (results not shown).

ML confidence intervals were very similar, independent of the method used (Table 2). Therefore, the normal approximation to the ML estimates worked as well as the bootstrap for the parameters in the tree that was considered. Actually, if sample size is very large, it is expected the ML estimate to be normally distributed around the true parameter value (Wackerly et al., 2001). Both bootstrap SEs rendered almost identical confidence intervals. The distributions of the bootstrap estimates were very close to the normal curve (Figure 2), and thus any measure of bootstrap confidence limits would likely be the same.

Table 2. Ninety-five percent confidence intervals for maximum likelihood estimates calculated with several methods.

Parameter	Curvature	Bootstrap (Standard)	Bootstrap (Hall's percentile)	Bayesian
b_2	33.220 - 42.188	33.122 - 42.216	32.891 - 41.857	31.352 - 44.824
b_3	14.779 - 18.711	14.691 - 18.768	14.620 - 18.669	13.238 - 19.411
b_4	11.366 - 15.004	11.276 - 15.065	11.206 - 15.017	9.478 - 15.548
b_5	5.127 - 7.475	5.156 - 7.428	5.075 - 7.296	4.005 - 7.890
b_6	9.464 - 12.968	9.487 - 12.918	9.384 - 12.894	8.995 - 15.163

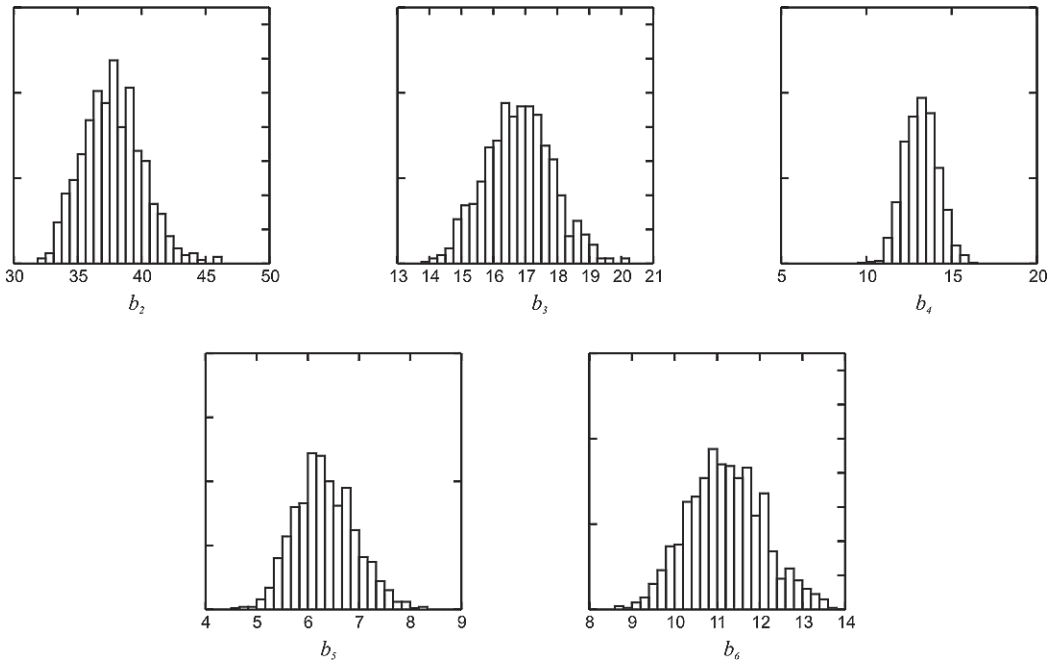

Figure 2. Histogram of 1,000 bootstrap replicates for each parameter.

The major difference found in the analyses was the difference between ML and the Bayesian method. Although Bayesian posterior probabilities and ML confidence intervals (calculated by the curvature or the bootstrap) are not analogous measures, in practice, they both inform the researcher about the credibility of the divergence times inferred. Bayesian estimates of 95% credibility intervals were wider than those obtained by ML for all parameters investigated (Table 2, Figure 3). This behavior was already reported by Yang and Yoder (2003), who suggested that narrow ML intervals could be caused by inappropriate assumption of the normality of the ML estimate or by not considering the uncertainties of calibrations, which are fixed.

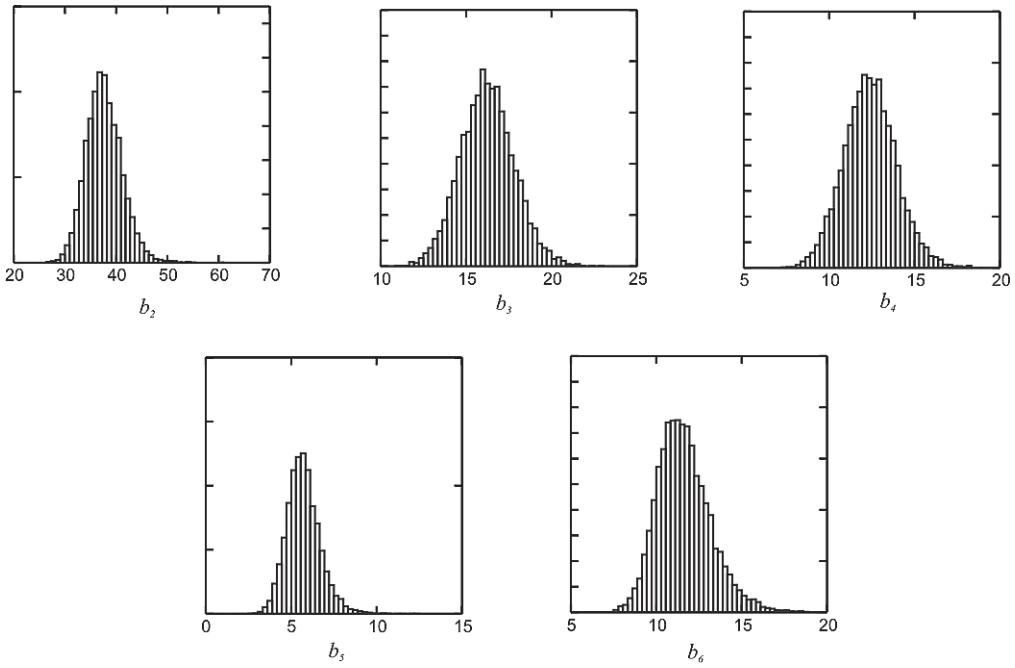

Figure 3. Posterior distribution for parameters inferred by the Bayesian MCMC method of Thorne and collaborators.

Here, when normality of the ML estimate was assumed, confidence intervals were similar to those from the bootstrap and the Bayesian calibration point was very narrow. Therefore, an explanation for the difference between ML and Bayesian confidence intervals is not simple. In fact, this explanation is theoretically meaningless. ML and Bayes are distinct approaches to statistical inference, and the researcher must evaluate if the use of prior distributions and the modeling of divergence times and rates of molecular evolution are realistic. It could also be argued that the differences between ML and Bayesian credibility intervals were due to failure in approximating the Bayesian posterior distribution by the MCMC algorithm. This could be caused by factors such as insufficient burn-in period. In this study this is not the case. MCMC analyses were conducted several times with different seed numbers, prior and heating periods. In every case, the posterior approximations were the same.

The analyses presented here showed that the SEs of ML estimates calculated by the curvature method and those independently inferred by the two bootstrap approaches were fundamentally unaltered and behaved regularly. Although not large, differences were found between the ML and Bayesian methods. In practice, such discrepancies are of little importance if the objective is to obtain time scales for splits between lineages.

Finally, it is reasonable to assume that the correspondence between the curvature SE and those from bootstrap found here should also be found in data sets including a similar number of taxa, with a sequence length of at least 1,806 bp (the size of the ND5 gene), using models of evolution with the number of parameters up to that of the HKY85+ Γ and adopting a minimum of one calibration point. However, a thorough investigation of this issue must be carried out with simulation studies that force variation on several parameters that potentially affect the SEs of ML estimates of divergence times, such as the number of sequences used, sequence length, model of sequence evolution, number of calibration points adopted, and tree shape.

ACKNOWLEDGMENTS

The author thanks CNPq (Brazilian Research Council) for the financial support received during his post-doctoral studies. This work has been significantly improved by comments from the anonymous referees.

REFERENCES

Edwards AWF (1972). Likelihood. Cambridge University Press, Cambridge, England.

Efron B (1979). Bootstrap methods: another look at the jackknife. *Ann. Stat.* 7: 1-26.

Felsenstein J (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach. *J. Mol. Evol.* 17: 368-376.

Felsenstein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39: 783-791.

Gill PE and Murray W (1981). Practical optimization. Academic Press Limited (AP), London, England.

Hall P (1992). The bootstrap and edgeworth expansion. Springer-Verlag New York Inc., New York, NY, USA.

Hasegawa M, Kishino H and Yano T (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160-174.

Hasegawa M, Kishino H and Saitou N (1991). On the maximum likelihood method in molecular phylogenetics. *J. Mol. Evol.* 32: 443-445.

Huelsenbeck JP (1995). The performance of phylogenetic methods in simulation. *Syst. Biol.* 44: 17-48.

Kishino H and Hasegawa M (1989). Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in *Hominoidea*. *J. Mol. Evol.* 29: 170-179.

Kishino H, Miyata T and Hasegawa M (1990). Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* 31: 151-160.

Kishino H, Thorne JL and Bruno WJ (2001). Performance of a divergence time estimation method under a probabilistic model of rate evolution. *Mol. Biol. Evol.* 18: 352-361.

Manly BFJ (1997). Randomization, bootstrap and Monte Carlo methods in biology. Chapman and Hall Ltd., New York, NY, USA.

Nei M and Kumar S (2001). Molecular evolution and phylogenetics. Oxford University Press, Eynshaw, Oxon, England.

Schrago CG and Russo CA (2003). Timing the origin of New World monkeys. *Mol. Biol. Evol.* 20: 1620-1625.

Thorne JL and Kishino H (2002). Divergence time and evolutionary rate estimation with multilocus data. *Syst. Biol.* 51: 689-702.

Thorne JL, Kishino H and Painter IS (1998). Estimating the rate of evolution of the rate of molecular

evolution. *Mol*. *Biol*. *Evol*. 15: 1647-1657.
Wackerly DD, Mendenhall W and Scheaffer RL (2001). Mathematical statistics with applications. 6th edn. Duxbury Press, Boston, MA, USA.
Yang Z (1994a). Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J*. *Mol*. *Evol*. 39: 306-314.
Yang Z (1994b). Statistical properties of the maximum likelihood method of phylogenetic estimation and comparison with distance matrix methods. *Syst*. *Biol*. 43: 329-342.
Yang Z (1997). PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput*. *Appl*. *Biosci*. 13: 555-556.
Yang Z and Yoder AD (2003). Comparison of likelihood and Bayesian methods for estimating divergence times using multiple gene loci and calibration points, with application to a radiation of cute-looking mouse lemur species. *Syst*. *Biol*. 52: 705-716.