Statistics of Divergence Times

Bernhard Haubold and Thomas Wiehe
Max-Planck-Institut für Chemische Ökologie, Jena, Germany

Given the number of nucleotide substitutions between two species (K) and the substitution rate ν, the expectation of the corresponding divergence time is usually calculated as $K/(2\nu)$. This is strictly true only if ν is regarded as a constant because the ratio of two random variables, such as $K/(2\nu)$, has distributional properties different from those of the distribution of K. Therefore, both the mean and any confidence interval for divergence times are unknown in this situation. We model the distribution of K and ν using the Gamma distribution and calculate the mean and 95% confidence interval for the corresponding divergence time. These calculations are compared with results obtained by bootstrapping sequence data from the model plant *Arabidopsis thaliana* and its relatives. We show that for nonoverlapping pairs of phylogenetic distances, our method approaches the bootstrap results very closely. In contrast, regarding the mutation rate as a constant leads to strong underestimation of the confidence interval. An implementation of our method of computing divergence times is accessible through a web interface at http://www.soft.ice.mpg.de/cite.

Introduction

The question of how distantly two taxa are separated from their last common ancestor is probably as old as biology itself. With DNA data, the expected divergence time between taxa i and j, T_{ij}, can be computed in a straightforward way if the mutation rate is regarded as constant:

$$E(T_{ij}) = \frac{E(K_{ij})}{2 \times E(\nu)},$$

where K_{ij} is the number of substitutions per site between taxa i and j, and ν is the number of substitutions per site per year. If we further let $[K_1, K_2]$ be a $100(1 - \alpha)\%$ confidence interval for K_{ij}, the corresponding confidence interval for T_{ij} is

$$[T_1, T_2] = [K_1/(2\nu), K_2/(2\nu)].$$

However, ν is usually estimated from the number of substitutions between a pair of taxa that can be dated, e.g., by reference to fossil data, and hence is a random variable itself. This complicates the computation of both the mean and the confidence interval for divergence times. As we shall see, the difference between regarding ν as a random variable and regarding it as constant is much more marked for the confidence interval than for the mean.

In order to develop an intuition for the calculation of divergence times, we carried out a set of exploratory simulations. Let K and ν be normally distributed random variables with means 0.141 and 1.46×10^{-8} and standard deviations 0.024 and 0.025×10^{-9}. These are biologically meaningful values. We drew 10^5 random numbers from these distributions and calculated the expected 95% confidence interval for the corresponding divergence time as $[2.22 \times 10^6, 1.02 \times 10^7]$. Regarding the mutation rate as fixed led to the underestimation of this interval by approximately 30% (fig. 1).

Steel, Cooper, and Penny (1996) recognized this problem and suggested the following solution: let $[\nu_1, \nu_2]$ be a $100(1 - \alpha/2)\%$ confidence interval for ν. In this case, $[K_1/(2\nu_2), K_2/(2\nu_1)]$ is thought to be a $100(1 - \alpha)\%$ confidence interval of the divergence time (Steel, Cooper, and Penny 1996). This method did indeed lead to a wider confidence interval than that obtained for fixed ν, but this time, the interval was about 70% wider than expected (fig. 1). Furthermore, the mathematical justification for their proposed method is unclear. In the following, we shall present a solution to this problem and apply it to the divergence between the model plant *Arabidopsis thaliana* and its relatives among the crucifers.

Materials and Methods

Derivation of a Probability Density Function of the Divergence Time

Variation in the substitution rate among sites along a DNA sequence is often modeled by a Gamma probability distribution (Yang 1996). In these models, the number of substitutions is negative binomially distributed (Stuart and Ord 1994, p. 182). Equating the first and second moments of the negative binomial distribution with those of a Gamma distribution, the two parameters a (shape) and b (scale) of the Gamma distribution can be uniquely determined. For the biologically reasonable parameters tested, the Gamma distribution provides an excellent approximation of the negative binomial distribution. Furthermore, if the number of substitutions in a gene (H) is Gamma-distributed, then the number of substitutions per site $K = H/n$, where n denotes the number of sites, is also Gamma-distributed. Our method starts with this assumption. The density function of the Gamma distribution is

$$f(x) = \frac{b^{-a}\exp\left(-\frac{x}{b}\right)x^{a-1}}{\Gamma(a)}, \qquad x \geq 0.$$

We assume that we are provided with measurements for

Key words: divergence time, substitution rate, Gamma distribution, *Arabidopsis thaliana*.

Address for correspondence and reprints: Bernhard Haubold, Max-Planck-Institut für Chemische Ökologie, Carl-Zeiss-Promenade 10, D-07745 Jena, Germany. E-mail: haubold@ice.mpg.de.

Mol. Biol. Evol. 18(7):1157–1160. 2001

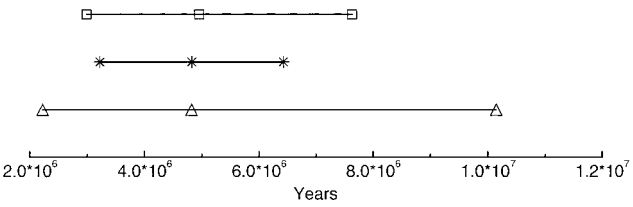

FIG. 1.—Ninety-five percent confidence intervals for the divergence time of a pair of taxa obtained by different methods. □ = true mean and confidence interval as determined by simulation; ★ = mean and confidence interval for fixed mutation rate; △ = mean and confidence interval according to Steel, Cooper, and Penny (1996).

(1) mean $\bar{K}$ and standard deviation σ of the per-base substitution rate between a pair of sequences, the target pair, and for (2) mean $\bar{\nu}$ and standard deviation τ of the mutation rate per base per year of a second sequence pair, the reference pair. As explained above, we assume that the random variables K and ν are Gamma-distributed. The corresponding parameters are $(\bar{K}/\sigma)^2$ and $\sigma^2/\bar{K}$ for the distribution of K, and $(\bar{\nu}/\tau)^2$ and $\tau^2/\bar{\nu}$ for the distribution of ν. Now, we need to determine the probability density of the ratio $Z = K/(2\nu)$. Note that 2ν is also Gamma-distributed with parameters $(\bar{\nu}/\tau)^2$ and $2\tau^2/\bar{\nu}$. Assuming that K and ν (and therefore also K and 2ν) are statistically independent, the density of the ratio is

$$f_Z(z) = \int_0^\infty x f_{2\nu}(x) f_K(xz)\, dx, \qquad (1)$$

where $f_{2\nu}$ and f_K denote the Gamma densities for K and 2ν, respectively. Abbreviating $\eta = \bar{\nu}/\tau$ and $\xi = \bar{K}/\sigma$ and substituting the respective Gamma density functions into equation (1), one finds

$$f_Z(z) = \frac{2^{-\eta^2}\left(\frac{\sigma}{\xi}\right)^{-\xi^2}\left(\frac{\tau}{\eta}\right)^{-\eta^2} z^{\xi^2-1}\Gamma(\xi^2+\eta^2)}{\left(\frac{\eta}{2\tau}+\frac{\xi z}{\sigma}\right)^{\xi^2+\eta^2}\Gamma(\xi^2)\Gamma(\eta^2)}.$$

This can be slightly simplified to

$$f_Z(z) = \frac{1}{B(\xi^2,\eta^2)\left(\frac{\text{scale}(K)}{\text{scale}(2\nu)}\right)^{\eta^2} z^{1-\xi^2}\left(\frac{\text{scale}(K)}{\text{scale}(2\nu)}+z\right)^{\xi^2+\eta^2}}, \qquad (2)$$

where $B(.,.)$ denotes the Euler Beta function. If the two scale parameters, $\text{scale}(K) = \sigma^2/\bar{K}$ and $\text{scale}(2\nu) = 2\tau^2/\bar{\nu}$, were identical, equation (2) would reduce to the Beta distribution of the second kind (Stuart and Ord 1994, p. 190). Numerical integration of equation (2) with appropriate integration bounds yields means and confidence intervals for the desired divergence times. A program implementing these computations is accessible via a web interface at http://soft.ice.mpg.de/cite.

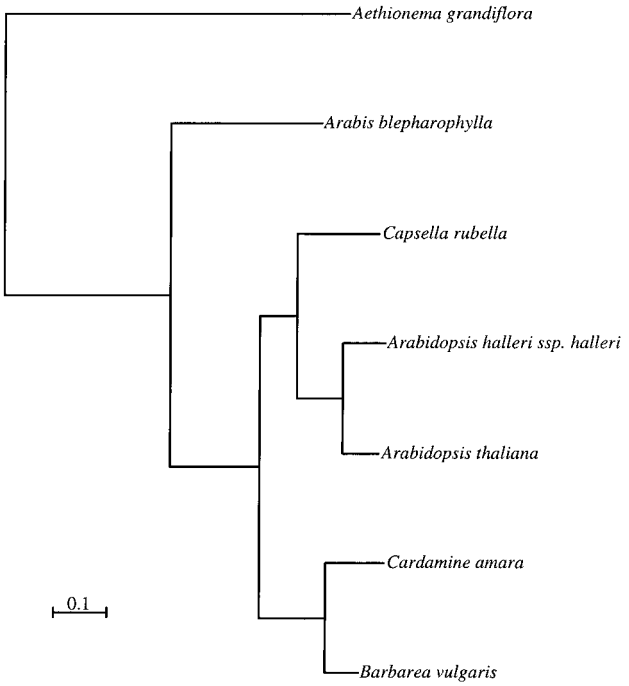

FIG. 2.—Neighbor-joining tree of *Arabidopsis thaliana* and some of its cruciferous relatives based on the number of synonymous substitutions at the chalcone synthase locus. Sequence data are taken from Koch, Haubold, and Mitchell-Olds (2000).

Bootstrap Simulation

In order to simulate the null distribution of divergence times, we generated pseudosamples using the bootstrap procedure (Efron 1979): an alignment of homologous protein-coding sequences was created, consisting of one reference pair, with a known (or assumed) divergence time, and a target pair. Pseudosamples were generated by sampling columns of codons with replacement and recalculating the synonymous mutation rate from the reference pair, the synonymous substitution rate from the target pair, and the corresponding divergence time. Substitution rates were calculated using the method of Li (1993) as implemented by Wolfe (1993). The average of the simulated divergence times was used as an estimator of the null distribution's mean. Further, the bootstrapped divergence times were sorted, and the desired $100(1-\alpha)\%$ confidence interval was obtained by removing the top and bottom $100 \times \alpha/2\%$ of their distribution.

Results

We applied bootstrap simulations and the numerical method outlined above to the complete coding sequence of the chalcone synthase locus (*Chs*) from *A. thaliana* and its relatives among the crucifers (Koch, Haubold, and Mitchell-Olds 2000). As a reference pair, we chose the crucifers *Cardamine amara* and *Barbarea vulgaris* (fig. 2). From pollen data, these are estimated to have diverged 6 MYA (Koch, Haubold, and Mitchell-Olds 2000). For the comparisons between *A. thaliana* and *Arabidopsis halleri*, *Capsella rubella*, and *Arabidopsis*

Table 1
Comparison of Divergence Times Calculated According to the Method Proposed in this Paper (*Gamma*) and According to the Traditional Method Based on a Fixed Mutation Rate (*Fixed*)

	Divergence Time (Myr)							
	Bootstrap		Gamma			Fixed		
Taxa	Mean	CI	Mean	CI	Error (%)	Mean	CI	Error (%)
A.t./A.h.	5.2	[3.2, 8.0]	5.2	[3.3, 8.0]	1.3	5.1	[3.4, 6.8]	29.9
A.t./C.r.	11.3	[7.6, 16.4]	11.3	[7.5, 16.4]	0.8	11.0	[8.3, 13.7]	38.8
A.t./A.b.	22.4	[15.4, 32.2]	22.3	[15.2, 31.9]	3.0	21.8	[17.2, 26.4]	45.2
C.a.B.v.	6.2	[4.2, 8.7]	6.2	[4.2, 8.8]	2.9	6.2	[4.3, 8.0]	18.9
A.t./B.v.	15.0	[10.4, 21.3]	15.0	[10.1, 21.6]	5.5	14.6	[11.2, 18.0]	37.6
A.b./B.v.	22.8	[15.9, 32.5]	28.8	[15.6, 32.6]	2.4	22.3	[17.5, 27.0]	42.8

Note.—For each method, the mean, 95% confidence interval (CI) and error are quoted. The error was calculated as the difference between the bootstrapped interval and the interval concerned, divided by the bootstrap interval, times 100. *A.t.* = *Arabidopsis thaliana*; *A.h.* = *Arabidopsis halleri* spp. *halleri*; *C.r.* = *Capsella rubella*; *A.b.* = *Arabis blepharophylla*; *C.a.* = *Cardamine amara*; *B.v.* = *Barbarea vulgaris*; c.f., figure 2.

blepharophylla, the results of our method deviated from the bootstrap results by $<3\%$, compared with $\geq 30\%$ for fixed mutation rate (table 1). In all of these examples, *C. amara* and *B. vulgaris* were used as the reference taxa. When we turned the calculation on its head and fixed the divergence of *A. thaliana* and *A. blepharophylla* at 22.4 Myr, the corresponding divergence time for *C. amara* and *B. vulgaris* was estimated as 6.2 Myr, which was again well approximated by our Gamma method (table 1).

In the examples presented so far, the calculations were always based on four taxa and the two nonoverlapping distances that could be formed between them (fig. 3). In order to investigate a triplet of sequences, we calculated the divergence between *A. thaliana* and *B. vulgaris* while using *B. vulgaris* and *C. amara* as reference sequences, as in the first three examples. This returned the worst agreement with the bootstrap results (5.5% deviation from bootstrap result; table 1), although our method was still much more reliable than the traditional method based on a fixed mutation rate (37.6% deviation from bootstrap result; table 1). When we considered pairs of distances with less overlap, the fit between bootstrap and analytical results improved again to 2.4% error (table 1; *A. blepharophylla*/*B. vulgaris*).

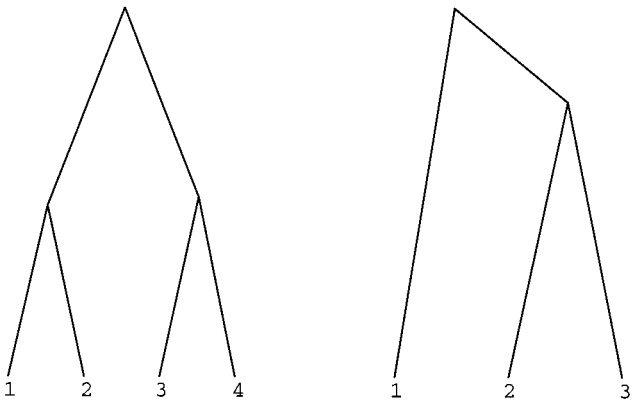

Fig. 3.—Random topologies for three and four taxa. In the left panel, distances between taxa 1/2 and taxa 3/4 do not overlap, while in the right panel distances between taxa 1/2 and 2/3 share part of the phylogenetic tree and are therefore not independent.

Discussion

The computation of divergence times is a standard part of phylogenetic analyses. Here, we concentrated on the apparently simple problem of computing divergence times given the number of substitutions (K) and the corresponding mutation rate (ν) for a particular pair of taxa. In the past, this calculation was often performed under the implicit assumption that the mutation rate could be regarded as constant. However, it is inconsistent to treat the mutation rate as a constant and the number of substitutions as a random variable. The justification for such an approach might be that for real-world examples it does not matter whether or not the mutation rate is treated as constant. Our bootstrap simulations show that the difference is rather large (fig. 1) and that the assumption that the mutation rate is constant leads to an underestimation of the confidence interval around the divergence time (table 1).

A hurdle to treating both the substitution and the mutation rate as random variables is the computational complications introduced by such an approach. Here, we sketched the derivation of equation (2), which leads to results that are close to those obtained by bootstrap simulation. The method works particularly well if it is based on pairs of nonoverlapping distances (fig. 3 and table 1). For pairs of overlapping distances, the bootstrap results may be less well approximated by formula (2), although the assumption of a constant mutation rate leads to a still greater error (table 1; *A. thaliana*/*B. vulgaris*). The reason for the greater error with strongly overlapping distances is that this violates the assumption that ν and K may be treated as independent random variables. This assumption is central to the derivation of formula (1) and therefore also to that of formula (2), on which we based our analytical calculations. If the overlap is reduced, the fit between simulation and analytical result also improves (table 1; *A. blepharophylla*/*B. vulgaris*). But even for comparisons with significant overlap, our method provides a reasonably accurate and computationally efficient alternative to bootstrap simulation for the calculation of confidence intervals around divergence times.

Acknowledgments

We thank two anonymous reviewers for comments on the manuscript. This work was supported by the Max Planck Society.

LITERATURE CITED

EFRON, B. 1979. Bootstrap methods: another look at the jackknife. Ann. Stat. **7**:1–26.

KOCH, M. A., B. HAUBOLD, and T. MITCHELL-OLDS. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae). Mol. Biol. Evol. **17**:1483–1498.

LI, W.-H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. Mol. Evol. **36**:96–99.

STEEL, M. A., A. C. COOPER, and D. PENNY. 1996. Confidence intervals for the divergence time of two clades. Syst. Biol. **45**:127–134.

STUART, A., and J. K. ORD. 1994. Kendall's advanced theory of statistics, Vol. 1. Distribution theory. 6th edition. Edward Arnold, London.

WOLFE, K. H. 1993. Software program li93. University of Dublin, ftp://acer.gen.tcd.ie/pub/khwolfe/li93.

YANG, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. TREE **11**:367–372.

FUMIO TAJIMA, reviewing editor

Accepted March 5, 2001