A methodological bias toward overestimation of molecular evolutionary time scales

Francisco Rodríguez-Trelles*[†][‡], Rosa Tarrío*[§], and Francisco J. Ayala*

*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697-2525; [†]Instituto de Investigaciones Agrobiológicas de Galicia (CSIC), Avenida de Vigo s/n, Apartado 122, 15780 Santiago de Compostela, Spain; and [§]Misión Biológica de Galicia (CSIC), Apartado 28, 36080 Pontevedra, Spain

Contributed by Francisco J. Ayala, April 16, 2002

There is presently a conflict between fossil- and molecular-based evolutionary time scales. Molecular approaches for dating the branches of the tree of life frequently lead to substantially deeper times of divergence than those inferred by paleontologists. The discrepancy between molecular and fossil estimates persists despite the booming growth of sequence data sets, which increasingly feeds the interpretation that molecular estimates are older than stratigraphic dates because of deficiencies in the fossil record. Here we show that molecular time estimates suffer from a methodological handicap, namely that they are asymmetrically bounded random variables, constrained by a nonelastic boundary at the lower end, but not at the higher end of the distribution. This introduces a bias toward an overestimation of time since divergence, which becomes greater as the length of the molecular sequence and the rate of evolution decrease.

The hypothesis of the molecular clock holds that the number of amino acid (or nucleotide) replacements in any given protein (or DNA) sequence changes linearly with time (1, 2). If constant, rates of molecular evolution can be extrapolated for dating past evolutionary events. Rates used for extrapolation have to be first calibrated by reference to absolute dates drawn from the fossil record. A notable feature of the hypothesis of the molecular clock is multiplicity: every one of the thousands of proteins or genes of an organism is an independent clock, each ticking at a different rate, but all measuring the same events (3–5). Molecular clock projections have ostensibly pushed back fossil-based dates in many studies (6). Prominent examples are the time of origin of the metazoan phyla, which has been placed as twice as old as determined by paleontologists (but see ref. 7), dating to more than 1,000 Myr ago (8–10); or the split of the three multicellular kingdoms, timed at about 1,600 Myr ago (9–11), some 400 Myr earlier than predicted from the fossil record.

Two not mutually exclusive explanations have been adduced to account for molecular earlier than fossil dates: (*i*) incompleteness of the fossil record, such that paleontological data can provide only minimal divergence dates (6, 12, 13); and (*ii*) too few genes and proteins considered, which turns molecular dating methods inaccurate (6, 9). It has been proposed that discrepancies between fossil and molecular dates will fade away as new fossil findings continue to accumulate; but also, and more steeply, as the size of molecular data sets become increasingly larger, because averages across numerous estimates of the same date will converge toward more consistent estimates (6, 9, 10, 14). Yet although data sets have become much larger and methods of analysis considerably more sophisticated, the discrepancy between fossil and molecular dates has not disappeared (reviewed in ref. 6). We now show that common molecular estimates are upwardly biased because of a fundamental flaw in the molecular approach to dating.

Suppose three orthologous protein sequences related as in Fig. 1, which have passed some molecular clock criterion (usually a "relative rate" test), and that we seek to determine the date when lineages C and AB split (denoted as t_T, or target time in Fig. 1). Let us assume that the average number of amino acid replacements per site between A and B is $K_{AB} = 1$, and that C differs from either A or B by $K_{AC} = K_{BC} = 10$. Also, it is known from the fossil record that A and B split from a common ancestor 100 Myr ago (denoted as t_C, or calibration time). Hence, the absolute rate of molecular evolution between A and B would be $r_{AB} = K_{AB}/2t_C = 5 \times 10^{-9}$ replacements per site per year. If we assume that r_{AB} (hereafter denoted as r_R, or reference rate) is equal to the rate between C and AB (hereafter denoted as r_U, or unknown rate), then the unknown date would be placed at $t_T = [(K_{AC} + K_{BC})/4]/r_R =$ 1,000 Myr ago. After conducting analogous calculations separately for each of n independent, putatively rate-constant protein regions, conventional molecular dating approaches would set the time of the split between lineages C and AB as the arithmetic mean across the ensuing n t_T values (e.g., refs. 7–11 and 14).

Note, however, that (*i*) even if rate constancy holds, r_R and r_U represent different realizations of a stochastic process, subject to sampling variation such that they are not expected to be identical; indeed, the dispersion of the rate of molecular evolution has proved to be much larger than expected if the probability of change were constant (3, 4, 15, 16); and (*ii*) because of its definition as a quotient of (often nonindependent, gamma-distributed) rates, time-since-divergence is an asymmetrically bounded random variate: constrained to be non-negative (i.e., the lower boundary is nonelastic) but unbounded above zero (i.e., elastic boundary). Equivalent random deviations around target times scale divisively forward (i.e., to the present), but multiplicatively backward (i.e., to the past) on their target times. As a result of this reciprocal scaling of under- and overestimates, the frequency distribution of time-since-divergence estimates is squashed up near the origin with a long tail to the right, yielding arithmetic averages that are upwardly biased with respect to the true times. Suppose that in Fig. 1 100 and 1,000 Myr are, respectively, the true divergence times between A and B, and between either of them and C. Now consider two protein sequences with observed r_R two times r_U for one protein, and r_R half r_U for the other protein. The first protein would date the split between C and AB 500 Myr later than it happened (i.e., 500 Myr ago), whereas the second one would set the split 1,000 Myr earlier (i.e., 2,000 Myr ago). The arithmetic average across the two proteins is 1,250 Myr, which still overestimates the true time by 250 Myr. These numbers become increasingly disparate as the ratio r_R/r_U deviates from 1.

To evaluate the extent of the overestimation that results from equating target times to arithmetic means across

Abbreviation: Myr, million years.

[‡]To whom reprint requests should be addressed at: (% Francisco J. Ayala) Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525. E-mail: ftrelles@iiag.cesga.es.

www.pnas.org/cgi/doi/10.1073/pnas.122231299

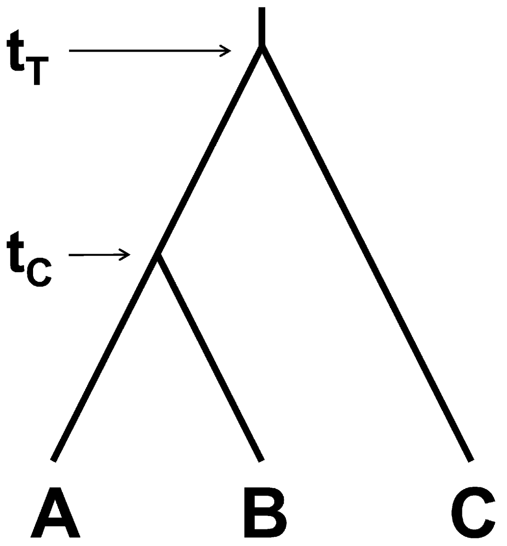

Fig. 1. Tree topology for lineages A, B, and C. t_C and t_T represent, respectively, calibration and target times.

multiple-gene data, we simulated the evolution of an ancestral amino acid sequence along the topology of Fig. 1 under different sets of conditions. For each condition set, the rate of replacement was fixed throughout the tree (i.e., $r_R = r_U$). Amino acid changes were generated conforming to the model of ref. 17, using the discrete gamma distribution with shape parameter α (the JTT+dG model; ref. 18) to accommodate among-site rate variation. Three different, biologically meaningful replacement rates were considered to represent slow (one replacement per site per 10^{10} years), intermediate (five replacements per site per 10^{10} years), and fast (ten replacements per site per 10^{10} years) evolving genes. Each replacement rate was combined with a specific value of α (0.5, 1.0, and 2.0, respectively), to take into account that slowly evolving proteins tend to have a high level of rate variation among sites, and *vice versa* (19). In all cases, t_C was set to 300 Myr, and for each rate class we considered three total tree lengths by setting alternatively t_T at 600 Myr, 1,100 Myr, and 3,000 Myr. We considered four sequence lengths (75, 150, 300, and 500 aa) that span most frequent alignment lengths (e.g., refs. 7–11 and 14). For each set of conditions we conducted 1,000 simulations. Each simulation produced three amino acid sequences related as in Fig. 1. With each sequence set we built a pair-wise distance matrix by using the same model (i.e., JTT+dG) and parameter values as in the original simulation. Then we extrapolated the inferred values of r_R to estimate corresponding t_T values. Simulations were performed with the EVOLVER program from the PAML package (20).

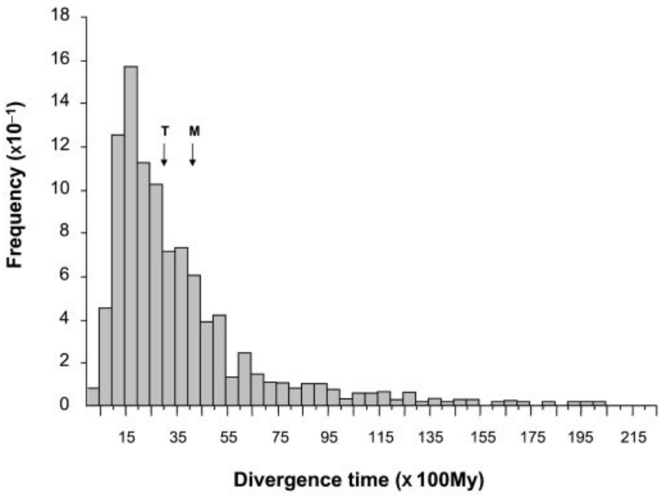

Fig. 2. Frequency distribution of 1,000 estimates of the divergence time between lineages C and AB in Fig. 1, set to have occurred 3,000 Myr ago, obtained using a short (75 residues long), slow evolving (one replacement per site per 10^{10} years) protein, and using the split between A and B, set to 300 Myr ago, as the calibration point. T and M represent target (i.e., 3,000 Myr) and estimated mean (i.e., 4,084 Myr; see Table 1) times, respectively.

The simulation results are shown in Table 1. As expected, owing to the distributional asymmetry of divergence times, even under a uniform rate model of evolution, arithmetic averages across molecular clock projections consistently overestimate the true date of divergence (i.e., Table 1 ratios enclosed in parentheses are all greater than one). The overestimation problem becomes aggravated as the rate of replacement decreases and/or the sequences become shorter. Both circumstances are expected to result in enhanced sampling variation of estimates, thus yielding increasingly right-skewed distributions. Fig. 2 illustrates the frequency distribution of 1,000 time estimates for the case of a short (75 residues long) slowly evolving (five replacements per site per 10^{10} years) protein used to date an episode 3,000 Myr old (first column in

Table 1. Mean divergence time estimates between lineages C and AB in Fig. 1, assuming that A and B split 300 Myr ago

				Protein length[§]			
r*	α	t_T[†]	Branch lengths[‡]	75	150	300	500
1	0.5	600	((A:3,B:3):3,C:6)	732 (1.22)	676 (1.13)	637 (1.06)	616 (1.03)
		1,200	((A:3,B:3):9,C:12)	1600 (1.33)	1372 (1.14)	1301 (1.08)	1243 (1.04)
		3,000	((A:3,B:3):27,C:30)	4084 (1.36)	3589 (1.20)	3210 (1.07)	3194 (1.06)
5	1.0	600	((A:15,B:15):15,C:30)	642 (1.07)	619 (1.03)	611 (1.02)	607 (1.01)
		1,200	((A:15,B:15):45,C:60)	1301 (1.08)	1237 (1.03)	1218 (1.02)	1212 (1.01)
		3,000	((A:15,B:15):135,C:150)	3308 (1.10)	3161 (1.05)	3053 (1.02)	3043 (1.01)
10	2.0	600	((A:30,B:30):30,C:60)	624 (1.04)	619 (1.03)	609 (1.02)	604 (1.01)
		1,200	((A:30,B:30):90,C:120)	1278 (1.07)	1238 (1.03)	1210 (1.01)	1219 (1.02)
		3,000	((A:30,B:30):270,C:300)	3505 (1.17)	3204 (1.07)	3122 (1.04)	3128 (1.04)

*Replacement rate (replacements per site per 10^{10} years).

[†] Target time.

[‡] The branch lengths for the topology in Fig. 1 (given in parenthetical notation) $\times 10^2$ are the expected absolute numbers of replacements per site.

[§] The ratios between estimated and target times are given in parentheses.

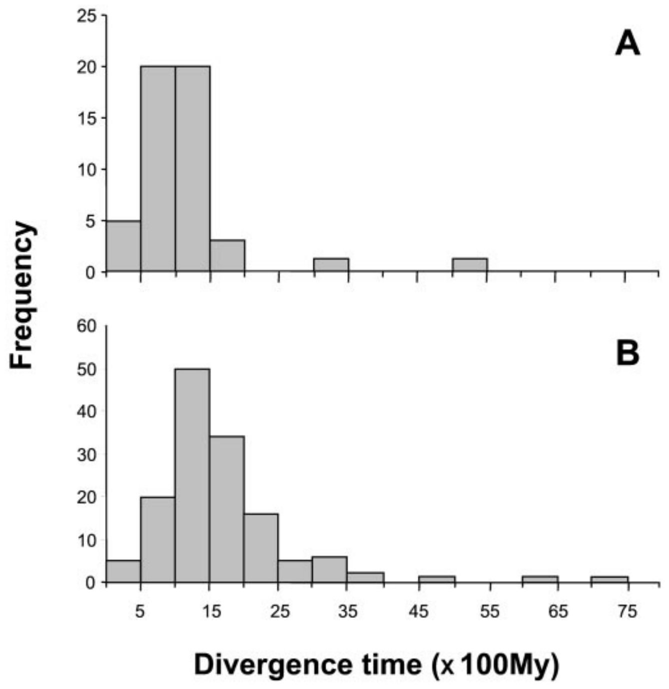

Fig. 3. Frequency distribution of divergence time estimates. (*A*) The split chordate-arthropod (50 estimates). (*B*) The split animal–fungi–plant (55 animal–fungi, 49 animal–plant, and 37 fungi–plant pooled together). Taken from table 1 of ref. 9.

third row of Table 1). The distribution is highly skewed to the right, giving an arithmetic average that places the event 4,084 Myr ago—i.e., more than 1,000 Myr earlier than actually happened. Table 1 also shows that overestimates grow as target times become increasingly remote. Apparently, this pattern results because, when the rate of replacement is low enough such that the sequences being handled become too short for accurately reflecting the expected number of variable sites, evolutionary rates become consistently underestimated. Underestimation is most acute for the reference rate, because it involves the shortest time span (i.e., there has been less time to accumulate replacements), and diminishes as the rate to be ascertained involves an increasingly remote divergence. Because of these systematic differences in sampling error between the calibration and extrapolation rates, the least related sequences will often appear to have diverged more, leading to inflated divergence times. Note that this methodological bias becomes enhanced as a consequence of the multiplicative scale of overestimates.

With real-world sequences, overestimates of divergence times are expected to be larger than suggested by our simulations, particularly because relative rate tests widely used to identify and exclude sequences that violate the rate-constancy assumption have but limited statistical power (7, 21–24). Relative rate tests neglect levels of rate variation between lineages where the rate of one lineage is as much as four times the rate of the other in most typical data sets (24). In addition, the power of relative rate tests decreases with the length of the sequences and the number of variable sites (23), which are precisely the conditions where sampling error differences between calibration and extrapolation rates become more pronounced.

Fig. 3 illustrates the distribution of divergence time estimates taken from a representative multiprotein study (see table 1 of ref. 9; see also figure 2 of ref. 10). As a calibration point, the study used 310 Myr for the date of the split between birds and mammals, considered to be reliably attested by the fossil record. On the basis of arithmetic averages, Wang *et al.* (9) placed the divergence between arthropods and chordates at 993 ± 46, and the three-way split of animals, fungi, and plants at 1576 ± 88 Myr ago (see also ref. 10)—i.e., some 400 Myr earlier than predicted from the fossil record in both cases. Yet it is apparent from Fig. 3 that the distribution of estimated divergence times is conspicuously asymmetric, and markedly right-skewed in the two examples, as expected from the reciprocal scaling of under and overestimates on the target time. This asymmetry was noted by Wang *et al.* (9), who attributed it to the presence of outliers.

Despite the booming amount of sequence information, molecular timing of evolutionary events has continued to yield conspicuously deeper dates than indicated by the stratigraphic data. Increasingly, the discrepancies between molecular and paleontological estimates are ascribed to deficiencies of the fossil record, while sequence-based time tables gain credit. Yet, we have identified a fundamental flaw of molecular dating methods, which leads to dates that are systematically biased toward substantial overestimation of evolutionary times. Moreover, as rate ratios, divergence times are highly sensitive to the vagaries of the molecular clock. It is thus not surprising that early molecular assessments inferred widely varying dates for the same event, some of them far earlier than those derived from the fossil record (reviewed in ref. 6). These studies typically focus on just one or a few, often slowly evolving (i.e., most easily alignable) proteins. Averages across multiple measures of the same divergence time are expected to converge to more consistent overestimates as molecular data sets become vastly larger in the future. If molecular-sequence-based time appraisals are to yield reliable estimates, centered around the target dates, they should take a new turn. Although enlarging the size of the data sets remains a critical issue, attention must be paid to careful choice of the sequences. Close approximation to the molecular clock premise should be a necessary condition. Given the limited power of available tests, however, acceptance of this premise seems safe only for long and fast evolving (yet alignable) sequences. Although proceeding with appropriate caution may not completely close the gap between clocks and rocks (for they still measure different events), it will likely contribute to its narrowing.

F.R.-T. and R.T. received support from contracts Ramón y Cajal and Doctor I3P, respectively, from the Ministerio de Ciencia y Tecnología (Spain). This work was supported by National Institutes of Health Grant GM42397 (to F.J.A.).

1. Zuckerkandl, E. & Pauling, L. (1965) in *Evolving Genes and Proteins*, eds. Bryson, V. & Vogel, H. J. (Academic, New York), pp. 97–166.
2. Kimura, M. (1968) *Nature (London)* **217,** 624–626.
3. Ayala, F. J. (1986) *J. Hered.* **77,** 226–235.
4. Gillespie, J. H. (1991) *The Causes of Molecular Evolution* (Oxford Univ. Press, New York).
5. Li, W.-H. (1997) *Molecular Evolution* (Sinauer, Sunderland, MA).
6. Wray, G. A. (2002) *Genome Biol.* **3,** 1–7.
7. Ayala, F. J., Rzhetsky, A. & Ayala, F. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 606–611.
8. Wray, G. A., Levinton, J. S. & Shapiro, L. H. (1996) *Science* **274,** 568–573.
9. Wang, Y.-C., Kumar, S. & Hedges, S. B. (1999) *Proc. R. Soc. London Ser. B* **266,** 163–171.
10. Heckman, D. S., Geiser, D. M., Eidell, B. R., Stauffer, R. L., Kardos, N. L. & Hedges, S. B. (2001) *Science* **293,** 1129–1133.
11. Feng, D.-F., Cho, G. & Doolittle, R. F. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 13028–13033.
12. Glaessner, M. F. (1984) *The Dawn of Animal Life* (Cambridge Univ. Press, Cambridge, U.K.).

8114 | www.pnas.org/cgi/doi/10.1073/pnas.122231299 Rodríguez-Trelles *et al.*

13. Conway Morris, S. (1993) *Nature (London)* **361,** 219–225.
14. Nei, M., Xu, P. & Glasko, M. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 2497–2502.
15. Ayala, F. J. (1999) *BioEssays* **21,** 71–75.
16. Rodríguez-Trelles, F., Tarrío, R. & Ayala, F. J. (2001) *Proc. Natl. Acad. Sci. USA* **98,** 11405–11410.
17. Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992) *Comput. Appl. Biosci.* **8,** 275–282.
18. Yang, Z., Nielsen, R. & Hasegawa, M. (1998) *Mol. Biol. Evol.* **15,** 1600–1611.
19. Zhang, J. & Gu, X. (1998) *Genetics* **149,** 1615–1625.
20. Yang, Z. (1997) *Comput. Appl. Biosci.* **13,** 555–556.
21. Dobzhansky, Th., Ayala, F. J., Stebbins, G. L. & Valentine, J. W. (1977) *Evolution* (W. H. Freeman, San Francisco).
22. Scherer, S. (1989) *Mol. Biol. Evol.* **6,** 436–441.
23. Robinson, M., Gouy, M., Gautier, C. & Mouchirod, D. (1998) *Mol. Biol. Evol.* **15,** 1091–1098.
24. Bromham, L., Penny, D., Rambaut, A. & Hendy, M. D. (2000) *J. Mol. Evol.* **50,** 296–301.