Testing the Cambrian explosion hypothesis by using a molecular dating technique

LINDELL BROMHAM*†, ANDREW RAMBAUT*, RICHARD FORTEY‡, ALAN COOPER§, AND DAVID PENNY¶

*Department of Zoology, University of Oxford, Oxford OX1 3PS, United Kingdom; ‡Department of Palaeontology, Natural History Museum, Cromwell Road, London SW7 5BD, United Kingdom; §Department of Bioanthropology, University of Oxford, OX2 6QS, United Kingdom; and ¶Theoretical Biology, Massey University, Palmerston North, New Zealand

Edited by James W. Valentine, University of California, Berkeley, CA, and approved August 9, 1998 (received for review June 3, 1998)

ABSTRACT Molecular studies have the potential to shed light on the origin of the animal phyla by providing independent estimates of the divergence times, but have been criticized for failing to account adequately for variation in rate of evolution. A method of dating divergence times from molecular data addresses the criticisms of earlier studies and provides more realistic, but wider, confidence intervals. The data are not compatible with the Cambrian explosion hypothesis as an explanation for the origin of metazoan phyla, and provide additional support for an extended period of Precambrian metazoan diversification.

The early history of the Metazoa, whether originating as part of a Cambrian "explosion" or with an extended, Precambrian "phylogenetic fuse," remains controversial (1–3). The Cambrian explosion hypothesis—that the phyla and even classes of the animal kingdom originated in a rapid evolutionary radiation at the base of the Cambrian at 545 million yr ago (Ma) or 10–15 Ma before this date—rests on the sudden appearance of a diverse range of animals in the fossil record. Although recent discoveries of Ediacaran metazoans have extended the record of sponges and bilateral animals to 570 Ma (4, 5), the biological affinities of many Ediacaran organisms remain controversial, and the earliest palaeontological evidence of metazoan life is no more than 600 Ma (6). However, the absence of earlier metazoan fossils could be caused by systematic biases in preservation that left the Precambrian history of recognized phyla unrecorded (7). Molecular data have the potential to provide an independent test of gaps in the fossil record by providing a continuous historical record of all extant taxa, which can be used to infer the history of lineages that have low fossilization potential, or the presence of lineages over periods for which the fossil record is poor.

"Molecular clock" studies, which use genetic distance to infer time since separation of lineages, can be criticized on the grounds that variation in the rate of molecular evolution, could affect the accuracy of the date estimates. We address these criticisms by using a molecular phylogenetic technique that deals explicitly with rate heterogeneity and allows the Cambrian explosion hypothesis to be tested within a statistical framework. The uncertainty in phylogenetic reconstruction and the stochastic nature of the molecular clock preclude precise dating (8), but when the imprecision is expressed as appropriate confidence intervals, molecular data are ideally suited to testing specific hypotheses. Here, we test the Cambrian explosion hypothesis by asking whether molecular date estimates are compatible with the initial radiation of animal phyla around the time of the Precambrian-Cambrian boundary.

We use a maximum likelihood implementation of the quartet method (9), which incorporates a test of rate constancy. For a given sequence alignment, pairs of taxa for which independently derived dates of origin are available are combined into quartets (Fig. 1). An array of quartets was formed from all possible combinations of monophyletic dated pairs, then each of the quartets was subjected to a rate constancy test, based on the likelihood ratio test (10). The maximum likelihood position of the deep node for each of the quartets provides an estimate of the date of divergence between the pairs.

We address the problem of rate heterogeneity in two ways. First, having a fossil date for each pair allows us to estimate a rate for each side of the quartet (pair plus internal lineage leading to that pair, Fig. 1), which allows for rate variation between pairs. Second, a quartet is rejected if the likelihood for a free-rate model (each branch allowed to have a different rate) is significantly better than that for the rate-constrained model (each pair has its own rate, Fig. 1). Only those quartets that passed the rate constancy test were used to generate date estimates. The combination of a rate constancy test with high resolution of long DNA sequences (two data sets totaling over 7,000 bp), use of a wide range of taxa and fossil dates (rather than reliance on a single calibration rate), and a maximum likelihood analysis makes this method more reliable than previous molecular dating techniques (11).

METHODS

We used two data sets: (*i*) protein-coding mitochondrial DNA, taken from whole mitochondrial genome sequences (5,676 bp) and (*ii*) nuclear-encoded 18S rRNA (1,710 bp). Sequences were aligned by eye and any saturated regions that could not be confidently aligned were removed. The list of sequences and alignments are available on: http://evolve.zoo.ox.ac.uk./Alignments/Cambrian.html. These sequences were then combined into dated pairs using fossil dates taken from the literature∥. We chose fossil dates that were between 540 and 200 Ma, to give date estimates from a reasonably well characterized part of the fossil record and yet old enough to allow extrapolation to the deep node. The use of many dates permits testing for rate constancy between taxa and avoids reliance on a single calibration rate to date all lineages. Quartets were then formed from all possible combinations of
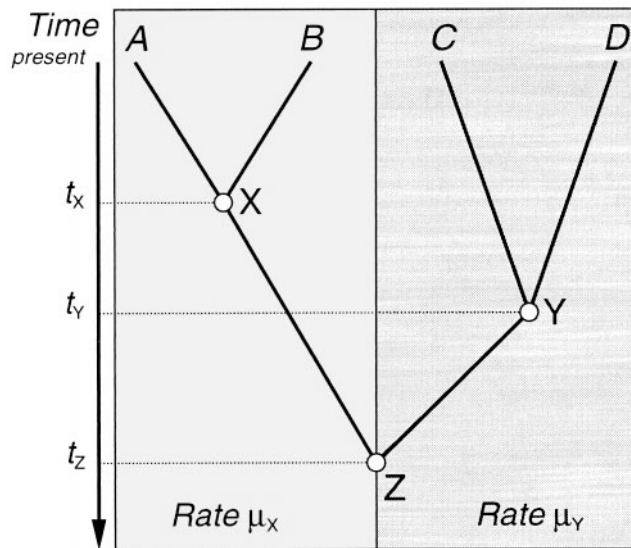
FIG. 1. Quartets are formed from two pairs of sequences (*A* and *B*; *C* and *D*), each with an independently derived date of origin (t_X and t_Y, respectively) connected by an internal node Z, the date of which is to be estimated (11). Maximum likelihood estimates of the two rates μ_X and μ_Y and the position of the deep node relative to the dated nodes allow estimation of the date of divergence between the pairs (t_Z).

dated pairs, as long as the monophyly of pairs was maintained (Fig. 1).

The maximum likelihood tree for all taxa was determined for each sequence alignment under the HKY85 model (12) with gamma-distributed rates across sites (13). Values for transition/transversion ratio and gamma shape parameter estimated for this tree were then used to calculate the maximum likelihood solution for each quartet. For the mitochondrial DNA, we obtained a transition/transversion ratio of 1.21 and gamma shape parameter of 0.38. For the 18S rRNA, these parameters were 1.74 and 0.38, respectively.

The likelihood of a quartet can be calculated for a given set of branch lengths. Under the rate-constrained model, branch lengths are obtained from a combination of the dates of the two known fossils in millions of years, the unknown date of divergence of the quartet and the two rates of evolution in substitutions per nucleotide site per million years ago. Following the notation of Fig. 1, the expressions for the branch lengths of this rooted tree in substitutions per site are: AX = BX = $\mu_X t_X$; XZ = $\mu_X(t_Z - t_X)$; CY = DY = $\mu_Y t_Y$; YZ = $\mu_Y(t_Z - t_Y)$. For each quartet, we obtain the values of t_Z, μ_X, and μ_Y that maximizes the likelihood. To obtain confidence intervals about the estimated date of divergence, we found the value of t_Z, either side of this maximum, that gave a log likelihood 1.92 less than the maximum log likelihood. This procedure derives from the expectation that twice the difference in log likelihood will be χ^2 distributed with 1 df at the 5% significance level (14).

To perform the rate constancy test, we simulated 1,000 sequence data sets for each quartet under the rate-constrained model. These were then analyzed to produce a distribution of the log likelihood differences between the rate-constrained and free-rate models. If the log likelihood difference observed for the real quartet was greater than that of 95% of the simulations, we rejected that quartet as failing to fit a rate-constrained model (15). Only quartets for which the rate-constrained model was not rejected were used to produce date estimates. The simulations were performed by using the program SEQ-GEN (16). Details of tests of the accuracy of the maximum likelihood quartet method and its robustness to rate heterogeneity are given elsewhere (11).

RESULTS

The majority of the quartets were rejected (Fig. 2), implying that rate heterogeneity is widespread among metazoan taxa for these sequences and demonstrating the value of testing each comparison separately rather than basing the expectation of clock-like evolution on inferred general trends. Date estimates from quartets with no detectable rate variation between pairs were obtained for two nodes, the echinoderm-vertebrate split and the protostome-deuterostome split (Fig. 2). We note that the branching order of phyla on the "tree of life" is still controversial (we have used the tree from ref. 17 for illustrative purposes because it is largely based on morphology) but all versions express these two fundamental dichotomies regardless of differences elsewhere in the tree. The confidence intervals are wide, as expected for molecular date estimates (8), therefore these results cannot be used to put an absolute date on the origin of phyla. However, the confidence intervals permit us to exclude the divergence of these lineages at or near the base of the Cambrian and all are older than 680 Ma, predating the recent metazoan fossil discoveries. Although we produce direct date estimates for only two nodes, we can combine these dates with a knowledge of metazoan phylogeny to infer the presence of other metazoan lineages in the Precambrian. All sister taxa to the dated nodes, such as Brachiopoda, Porifera, and Cnidaria (17), must also have been present before the Vendian period (600 Ma), despite the lack of earlier fossils.

The difference in the date estimates from our two sets of sequences illustrates the importance of basing estimates on several sequences and may reflect differences in the fit between model and data for different types of sequences. The evolution of protein-coding genes is likely to be more accurately modeled than rRNA. One aspect of the evolution of rRNA that is not adequately modeled is the evolution of paired sites (18, 19), the effect of which is likely to be an increase in the count of independent changes between two sequences, thus causing an overestimation of time since divergence. The shorter confidence intervals of mitochondrial estimates may result from longer sequences and a better fit between model and data.

Our findings are in broad agreement with some earlier molecular studies (20–22), despite differences in method and sequences used. Ayala *et al.* (23) reanalyzed the molecular data presented in ref. 21 and obtained a divergence date estimate of 704 Ma (which, although it substantially predates the Cambrian period, they claim to be compatible with the Cambrian explosion hypothesis). The authors obtained estimates by using a gamma model of rate heterogeneity between sites and compared them to those using a model where sites evolve at equal rates. Although both models gave similar date estimates, the degree of rate heterogeneity assumed (a shape parameter of 2.0 for all genes) was much less than those estimated previously (for review, see ref. 24). Underestimating the degree of rate heterogeneity between sites will tend to underestimate the dates of divergence (16, 25) caused by a nonlinear effect of heterogeneity on branch length.

The Cambrian explosion hypothesis rests on the interpretation of the sudden appearance of metazoans in the fossil record as evidence for a rapid radiation of animals late in the Proterozoic. Our results indicate that more metazoan lineages may have been present in the Precambrian than were unequivocally recorded in the fossil record. Therefore, the absence of appropriate Precambrian fossils is not necessarily evidence for an explosive evolutionary radiation near the beginning of the Cambrian period.

Pre-Vendian animals are likely to have been small and soft-bodied, because known fossils of Precambrian animals lack hard parts (26) and thus would have had poor fossilization potential. Fossilized skeletons appear for the first time in the

Echinodermata-Vertebrata
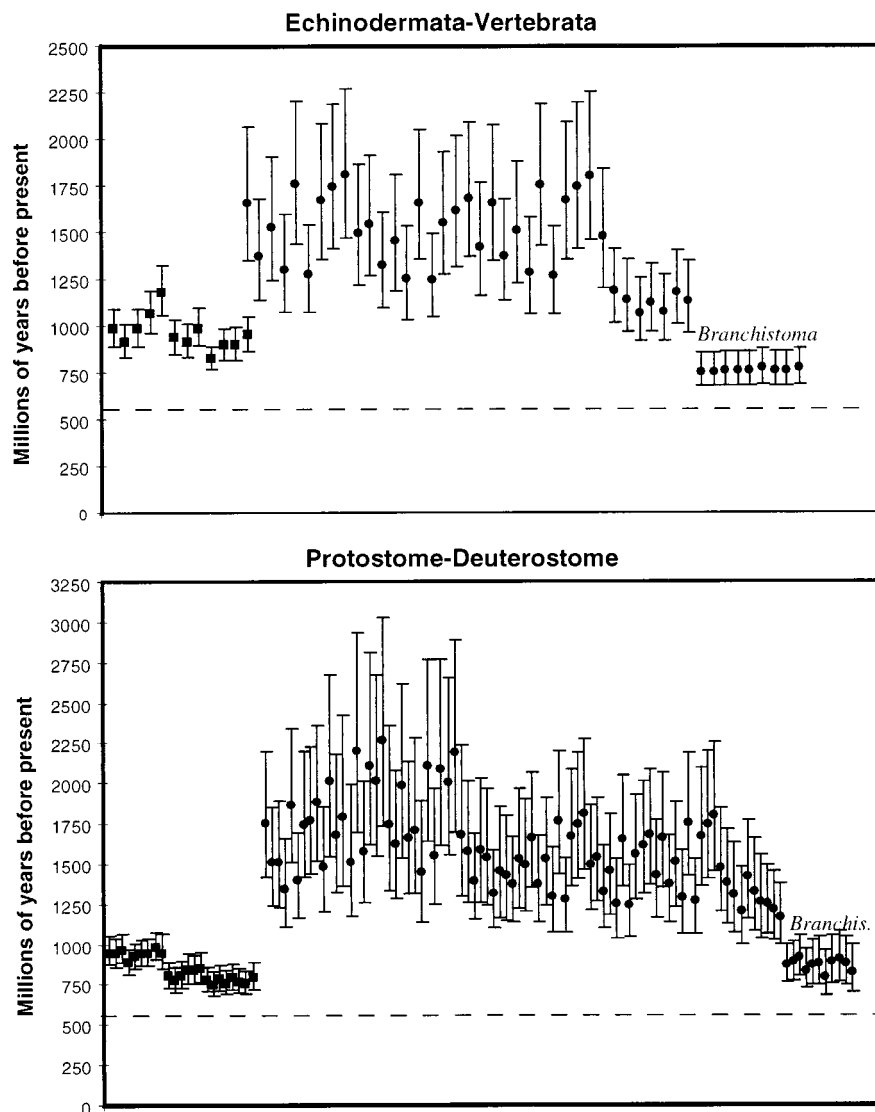

Protostome-Deuterostome

FIG. 2.   Date estimates with 95% confidence intervals, given by quartets that passed the rate constancy test for the mitochondrial (■) and 18S (●) alignments, for two deep nodes: the split between vertebrates and echinoderms, and the split between deuterostomes (vertebrates and echinoderms) and protostomes (all other pairs: see footnote). The date estimates are nonindependent because the quartets share many of the same branches of the phylogenetic tree. This nonindependence prevents the estimates from being combined statistically. The dashed line marks the Precambrian-Cambrian boundary at 545 Ma. Thirty-four percent of mitochondrial quartets and 22% of 18S quartets passed the rate constancy test. The 18S quartets containing *Branchistoma* (Cephalochordata) are significant outliers: these are small soft-bodied animals whose fossil record is likely to be poor, hence young fossil dates may overestimate the rate and produce recent divergence dates (21).

early Cambrian, their independent evolution in many separate lineages (27) possibly triggered by chemical changes in the oceans following post-glacial ocean upwelling (28). It seems probable that metazoan diversity is recorded for the first time in the Cambrian because of a combination of ideal fossilization conditions and the advent of hard parts, or larger bodies, or both, that made many animal lineages "visible" in the fossil record (7). The results from molecular studies are consistent with the interpretation of fossil evidence for a Precambrian history of modern phyla, either by new discoveries or through the reinterpretation of previous finds. Several early representatives of modern phyla are now tentatively identified from the Ediacaran period, including molluscs, cnidarians, arthropods, echinoderms, and annelids (29–32).

Molecular dating provides a means of testing hypotheses about the timing of origin of lineages, provided that molecular dates are presented with realistic estimates of confidence intervals. The method presented here is a significant advance on previous studies (20–23), because (*i*) the confidence intervals reflect the stochastic error in the substitution process; (*ii*)

assumptions of rate heterogeneity between lineages are explicitly tested within the maximum likelihood framework; and (*iii*) multiple rates of molecular evolution are included in the model. Although we cannot provide precise estimates of the origin of metazoan phyla, we can use our results to confidently reject the Cambrian explosion hypothesis, which rests on a literal interpretation of the fossil record and assumes that special evolutionary phenomena, capable of producing profound differentiation in a short period, operated in the Cambrian but not before or since. By contrast, the Precambrian phylogenetic fuse hypothesis assumes no more than we already know to be reasonable: that lineages can diverge gradually over time and that the fossil record contains gaps that can greatly reduce the chances of finding fossils for certain periods or particular types of organisms.

1. Bell, M. A. (1997) *Trends Ecol. Evol.* **12,** 1–2.
2. Conway Morris, S. (1997) *Curr. Biol.* **7,** 71–74.
3. Cooper, A. & Fortey, R. A. (1998) *Trends Ecol. Evol.* **13,** 151–156.
4. Li, C.-W., Chen, J.-Y. & Hua, T.-E. (1998) *Science* **279,** 879–882.
5. Xiao, S. Zhang, Y. & Knoll, A. H. (1998) *Nature (London)* **391,** 553–558.
6. Valentine, J. W., Erwin, D. H. & Jablonski, D. (1996) *Dev. Biol.* **173,** 373–381.
7. Fortey, R. A., Briggs., D. E. G. & Wills, M. A. (1996) *Biol. J. Linn. Soc.* **57,** 13–33.
8. Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. (1996) in *Phylogenetic Inference*, eds. Hillis, D. M., Moritz, C. & Mable, B. K. (Sinauer, Sunderland, MA), pp. 407–514.
9. Cooper, A. & Penny, D. (1997) *Science* **275,** 1109–1113.
10. Felsenstein, J. (1981) *J. Mol. Evol.* **17,** 368–376.
11. Rambaut, A. & Bromham, L. D. (1998) *Mol. Biol. Evol.* **15,** 442–448.
12. Hasegawa, M., Kishino, H. & Yano, T. (1985) *J. Mol. Evol.* **21,** 160–174.
13. Yang, Z. (1994) *J. Mol. Evol.* **39,** 306–314.
14. Yang, Z., Goldman, N. & Friday, A. (1995) *Syst. Biol.* **44,** 384–399.
15. Goldman, N. (1993) *J. Mol. Evol.* **36,** 182–198.
16. Rambaut, A. & Grassly, N. C. (1997) *Comput. Appl. Biosci.* **13,** 235–238.
17. Nielsen, C. (1995) *Animal Evolution: Interrelationships of the Living Phyla* (Oxford Univ. Press, Oxford).
18. Wheeler, W. C. & Honeycutt, R. L. (1988) *Mol. Biol. Evol.* **5,** 90–96.
19. Dixon, M. T. & Hillis, D. M. (1993) *Mol. Biol. Evol.* **10,** 256–267.
20. Runnegar, B. (1982) *Lethaia* **15,** 199–205.
21. Wray, G. A., Levinton, J. S. & Shapiro, L. H. (1996) *Science* **274,** 568–573.
22. Nikoh, N., Iwabe, N., Kuma, K., Ohno, M., Sugiyama, T., Watanabe, Y., Yasui, K., Shi-Cui, Z., Hori, K., Shimura, Y. & Miyata, T. (1997) *J. Mol. Evol.* **45,** 97–106.
23. Ayala, F. J., Rzhetsky, A. & Ayala, F. J. (1998) *Proc. Natl. Acad. Sci. USA* **95,** 606–611.
24. Yang, Z. (1996) *Trends Ecol. Evol.* **11,** 367–372.
25. Adachi, J. & Hasegawa, M. (1995) *J. Mol. Evol.* **40,** 622–628.
26. Conway Morris, S. (1993) *Nature (London)* **361,** 219–225.
27. Clarkson, E. N. K. (1993) *Invertebrate Palaeontology and Evolution* (Chapman and Hall, London).
28. Cook, P. & Shergold, J. (1984) *Nature (London)* **308,** 231–236.
29. Fedonkin, M. A. & Waggoner, B. J. (1997) *Nature (London)* **388,** 868–871.
30. Waggoner, B. J. (1996) *Syst. Biol.* **45,** 190–222.
31. Gehling, J. G. & Rigby, J. K. (1996) *J. Palaeont.* **70,** 185–195.
32. Conway Morris, S. (1989) *Science* **246,** 339–346.