COMMENTARY

Anciently duplicated genes reduce uncertainty in molecular clock estimates

Olga Zhaxybayeva[1]
Department of Biological Sciences, Dartmouth College, Hanover, NH 03755

Deciphering the histories of lineages as they have unfolded over billions of years, and placing them in the context of the known fossil record and Earth's biogeochemical events, is a challenging task in evolutionary biology. Shortly after discovering the molecular structure of DNA and determining the first amino acid sequences of proteins, it was recognized that the evolutionary distances among lineages could be estimated from the changes observed in nucleotide and amino acid sequences (substitutions). Combined with the hypothesis that steady accumulation of substitutions over time is analogous to the "ticking" of a clock [and hence the moniker "molecular clock" (1)], the relative divergence times among lineages could be estimated from the genetic distances among them. Estimating the times of the major evolutionary events in the history of Earth's biosphere could be made when the branches of the evolutionary tree are calibrated with available fossil evidence, allowing conversion of the relative divergence times to the absolute estimates (2). Realization that the molecular clock "ticks" unevenly within and across genes, among different lineages and over time, as well as a scarce and often controversial fossil record, ultimately led to a development of a new class of Bayesian "relaxed clock" models (3). These models do not require all lineages to evolve at the same rate and allow incorporation of heterogeneous (as well as vague) prior information from multiple genes, a large number of taxa, and multiple calibration points (3). Despite these methodological advances, dating various past evolutionary events has continued to produce conflicting estimates and wide uncertainty intervals. In PNAS, Shih and Matzke (4) introduce a clever approach to decrease the uncertainty associated with the inferred dates via additional constraints provided by anciently duplicated gene families.

In the course of evolution, many gene families undergo expansions through gene duplication. These expansions are considered to be among the driving forces behind evolutionary innovations, because duplicated gene copies (known as paralogs) can acquire new functions. Some duplication events happened in the distant past, before the time of the last universal common ancestor of the three domains of life. Analyses of two such ancient paralogous gene families, elongation factors Tu/1α and catalytic and noncatalytic subunits of ATPase, produced a significant breakthrough in our understanding of the relationship between the three domains of life (5, 6). On phylogenetic trees of these families, an internal branch separates the two paralogs. Therefore, one paralog can be used as an outgroup for the other, and thus be used to root the Tree of Life. An additional feature of the anciently duplicated gene's phylogeny—that any lineage divergence since the gene duplication event is represented by multiple nodes on the gene's phylogenetic tree—promises to provide yet another advance in our understanding of the past evolution of life. Shih and Matzke (4) recognize that for ancient paralogs the nodes corresponding to the lineage diversifications had to diverge at the same time, and therefore can provide additional time constraints in relaxed clock analyses.

Shih and Matzke's "proof of concept" analysis establishes the potential of duplicated gene families to refine disputed dates of major evolutionary events.

To incorporate these extra constraints, Shih and Matzke (4) introduce two strategies that they dub "cross-calibration" and "cross-bracing" (Fig. 1). In cross-calibration, for both paralogs an identical prior distribution for divergence times is assigned to all nodes involved in the same diversification event. However, the age values for the individual nodes at each step of Bayesian analysis are picked from the distribution independently and, therefore, can differ because of the uncertainty embraced by the distribution. A stricter cross-bracing approach requires individual nodes to have identical ages at each step of Bayesian analysis, making them covary during the analysis. The authors hypothesized that using either cross-calibration or cross-bracing should decrease the uncertainty of inferred node ages.

To test their hypothesis, Shih and Matzke (4) focus on the evolutionary history of ATPase gene family. Presence of multiple nuclear and organellar copies of catalytic and noncatalytic subunits of ATPase genes in eukaryotes allowed the authors to introduce several constraints for node ages. Cross-calibrated and cross-braced analyses resulted in a significant decrease in the uncertainty of inferred node ages and branch rates in comparison with those inferred using the noncatalytic subunit of ATPase alone, demonstrating an improvement in the overall accuracy of the relaxed molecular clock analysis.

Shih and Matzke's (4) "proof of concept" analysis establishes the potential of duplicated gene families to refine disputed dates of major evolutionary events. To illustrate this promise, the authors examine the nodes on the ATPase gene tree that correspond to three past evolutionary events: the acquisition of mitochondria and plastids by eukaryotic lineages, the diversification of eukaryotes, and the timing of the last universal common ancestor. Dating of these events was revisited in numerous previous studies, and the estimates vary and remain controversial. Although the inferred dates in the present study produce plausible scenarios in line with available biogeochemical evidence, it is hard to evaluate whether discrepancies with comprehensive multigene studies that involve larger numbers of fossil calibration points, such as a recent study by Parfrey et al. (7), are because of cross-calibration, limited phylogenetic information present in a single gene, or the reduced number of calibration points. Shih and Matzke (4) acknowledge that before extending cross-bracing analysis to multiple genes and abundant calibration

Author contributions: O.Z. wrote the paper.

The author declares no conflict of interest.

See companion article on page 12355.

[1]E-mail: olgazh@dartmouth.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1310930110

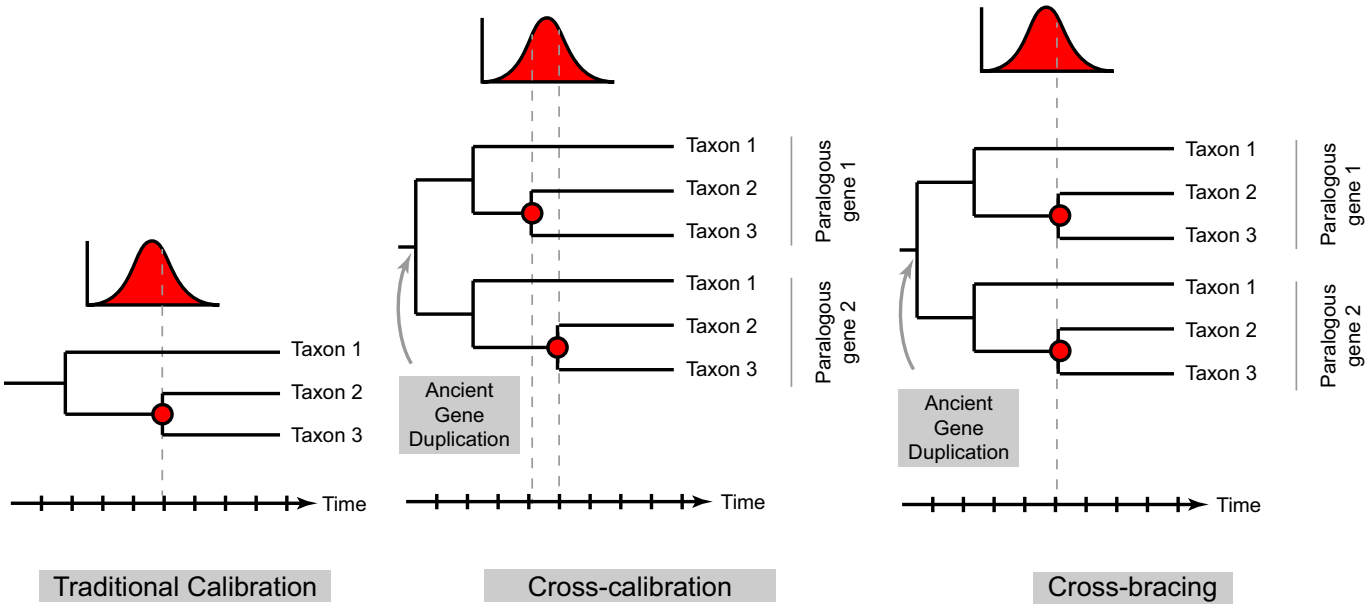

Fig. 1. Illustration of node age assignment in traditional, cross-calibrated, and cross-braced relaxed clock Bayesian analyses. Each panel represents a single step in the analysis. Red circles on phylogenetic tree mark the nodes corresponding to the same diversification event. In a gene family that underwent duplication, the tree will have two such nodes. The depicted probability distribution represents prior information about possible times for the diversification event. In a traditional relaxed clock analysis of a gene family without paralogs, the estimate is selected from the distribution (gray dashed line). In a cross-calibrated analysis of a gene family with paralogs, a diversification event is represented by one prior distribution from which the ages of the two nodes are drawn independently, thus introducing an additional constraint into the posterior age estimates. Under even stricter cross-bracing approach, the two nodes are assigned the same age. Modified from figure S1 in Shih and Matzke (4).

points, some of the technical challenges need to be addressed. The current relaxed clock software lacks appropriate settings to fully exploit the utility of a cross-bracing approach and is too slow. Possibly as a result, no significant decrease in the uncertainty of a cross-braced over a cross-calibrated approach was observed, despite the additional methodological strictness of the former.

Nevertheless, the proposed technique opens opportunities for more accurate future molecular clock analyses. In addition to ATPases and elongation factors, a genome-wide screen suggested the existence of over 150 gene families that putatively underwent duplication events before diversification of the three domains of life (8). Many of these gene families will not have the same evolutionary histories because of extensive horizontal gene transfer that have dramatically impacted the evolution of prokaryotes (9). Simulations have shown that horizontal gene transfer can affect coalescence times of the nodes of a gene tree (8), which could in turn influence the age estimates of the same nodes in different genes. Compounded with the saturation of substitutions and incomplete lineage sorting, dating events that go back billions of years will remain a tricky business. Still, multigene cross-braced analyses of anciently duplicated gene families show promise and may lead to more accurate estimates for the timing of the major biological inventions.

1 Zuckerkandl E, Pauling L (1962) Molecular disease, evolution, and genetic heterogeneity. *Horizons in Biochemistry*, eds Kasha M, Pullman B (Academic, New York), pp 189–225.
2 Donoghue PC, Benton MJ (2007) Rocks and clocks: Calibrating the Tree of Life using fossils and molecules. *Trends Ecol Evol* 22(8):424–431.
3 Welch JJ, Bromham L (2005) Molecular dating when rates vary. *Trends Ecol Evol* 20(6):320–327.
4 Shih PM, Matzke NJ (2013) Primary endosymbiosis events date to the later Proterozoic with cross-calibrated phylogenetic dating of duplicated ATPase proteins. *Proc Natl Acad Sci USA* 110:12355–12360.
5 Gogarten JP, et al. (1989) Evolution of the vacuolar H+-ATPase: Implications for the origin of eukaryotes. *Proc Natl Acad Sci USA* 86(17):6661–6665.
6 Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. *Proc Natl Acad Sci USA* 86(23):9355–9359.
7 Parfrey LW, Lahr DJ, Knoll AH, Katz LA (2011) Estimating the timing of early eukaryotic diversification with multigene molecular clocks. *Proc Natl Acad Sci USA* 108(33):13624–13629.
8 Zhaxybayeva O, Lapierre P, Gogarten JP (2005) Ancient gene duplications and the root(s) of the tree of life. *Protoplasma* 227(1):53–64.
9 Olendzenski L, Gogarten JP (2009) Evolution of genes and organisms: The tree/web of life in light of horizontal gene transfer. *Ann N Y Acad Sci* 1178(1):137–145.