Generating node calibration information for divergence time estimation using `CladeDate`

Santiago Claramunt

2022-06-06

Introduction

`CladeDate` is an R package for the estimation of the age of a focal taxonomic group or clade based on its fossil record, with output that can be directly used in molecular clock calibrations and phylogenetic time-tree estimation. The method uses a sample of the fossil record of a clade to generate point estimates, confidence intervals, and probability density functions for the age of the clade.

Selecting clades and fossils

The first step is to identify a focal clade of interest. The most recent common ancestor of that clade corresponds to a calibration node in a phylogenetic tree (**Figure 1**). The focal clade must have strong support from previous phylogenetic analysis and must have a fossil record. In particular, the oldest fossil in the focal clade should be of high quality in terms of phylogenetic evidence of clade membership and geostratigraphic provenance (Parham et al. 2012). This is because the oldest fossil is the most influential as it defines a minimum age for the clade. In fact, the identification of a high-quality fossil can be the first step in the selection of a focal clade. Note that previous assignment of a fossil to a taxonomic group, even if deemed well established, does not guarantee that the fossil is a descendant of the most recent common ancestor of that taxonomic group represented in the phylogenetic tree, as it can be part of the stem group instead.

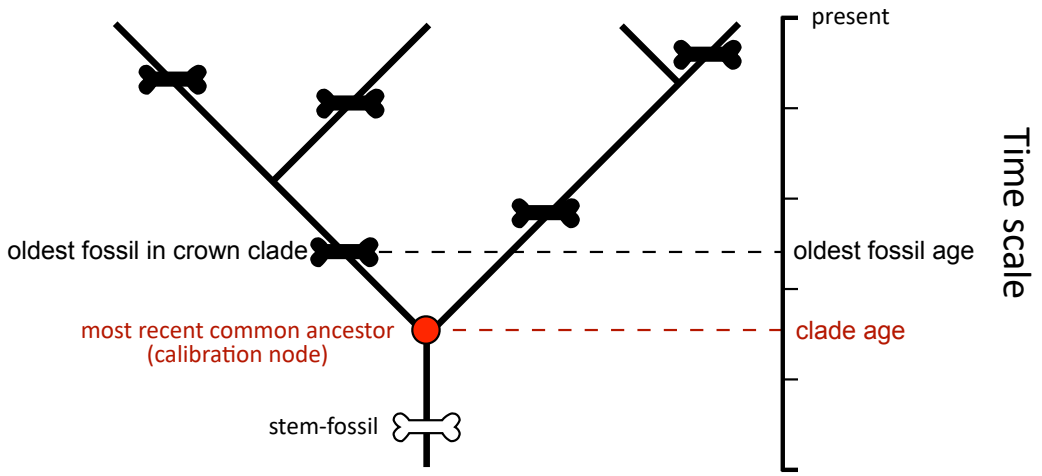


Figure 1. A hypothetical focal clade, relevant fossils, and their correspondence with the calibration ode and the time scale.

The second step is to collate information about the fossil record of the focal clade. However, not all fossils ever assigned to a clade may be included. In addition to the quality of the fossils and the certainty of taxonomic assignation, sample independence has been taken into account, as it will affect estimation. No more than one fossil from each fossil species and geological formation should be included, and, depending on the geographical scale, only one fossil from a particular region may be included to avoid non-independence arising from the history of biogeographic dispersal. For example, in a time-tree estimation for extant birds, only the first fossil occurrence of a clade in each continent was used (Claramunt & Cracraft 2015).

To demonstrate these principles, I will focus on estimating the age of the bird order Suliformes (frigatebirds, cormorants, and allies). The oldest fossil that can be unequivocally placed in the crown-clade Suliformes is *Limnofregata azygosternon* (Olson 1977), represented by a nearly complete and articulated skeleton with feather impressions from the Fossil Butte Member of the Green River Formation, Wyoming, USA (holotype USNM 22753). Derived characters suggest close affinities with Fregatidae (Olson 1977) and a cladistic analysis of extant and fossil Pelecaniformes, including 464 characters and 59 species, recovered *L. azygosternon* as a stem Fregatidae (Smith 2010). The age of the Fossil Butte member is well-known, with a radiometric date of 51.97 ± 0.09 Ma (Smith et al. 2010). *L. azygosternon* thus fulfills all requirements for a high-quality fossil for time-tree calibration (Parham et al. 2012).

Because the Suliformes is a cosmopolitan group, its biogeographic history may influence patterns of fossilization and recovery around the globe increasing spatial autocorrelation of fossil finds. Therefore, I will use the sample selected by Claramunt & Cracraft (2015) representing only the first fossil find of Suliformes on each continent.

Models for estimating clade age

`CladeDate` uses statistical methods for the estimation of the bounds of truncated distributions to estimate the age of a clade based on the temporal distribution of its fossil record. The most basic model assumes fossil recovery as a homogeneous Poisson process that generates a uniform distribution of fossil times $t_1 ... t_n$ between an appropriate base line (e.g. the present) and the age of the clade θ. For a base line in the present, the probability of the set of fossil times given the true clade age θ is $P(t_1 ... t_n \mid \theta) = 1/\theta^n$ (for $t < \theta$) (Wang & Everson 2007), and an unbiased point estimate of clade age is simply $t_n + t_n/n$ (Strauss & Sadler 1989, Solow 2003), in other words, adding the average time interval between fossil finds (t_n/n) to the oldest fossil age (t_n). An optimal confidence interval with confidence level 1 - a can be defined as $[t_n , t_n/a^{1/n}]$ (Strauss & Sadler 1989, Solow 2003, Wang et al. 2009). Base lines different from the present can easily be implemented by substituting t_i by $x_i = t_i$ – base line in the calculations. By default, `CladeDate` uses the youngest fossil in the set as a baseline so $x_i = t_i - t_1$ (Gingerich & Uhen 1998, Solow 2003).

Given the narrow uncertainty in the age of *Limnofregata*, we will ignore fossil age uncertainty for now and use the midpoint of fossil age intervals in a vector.

```
Fossils <- c(52.0, 48.0, 29.9, 25.7, 25.7, 25.0, 16.5)
```

We can then estimate the age of Suliformes using:

```
pdate(Fossils)
estimate    lower    upper
57.07143 52.00000 70.96152
```

Which returns a point estimate and 95% confidence bounds. The option `KStest = TRUE` can be used to test the assumption of uniformity via a Kolmogorov-Smirnov test.

```
pdate(Fossils, KStest = TRUE)
      Uniform distribution not rejected (Kolmogorov-Smirnov P = 0.43)
      estimate    lower    upper
      57.07143 52.00000 70.96152
```

In this case, the null hypothesis of uniformity is not rejected, and we can trust the estimates. Otherwise, a warning is printed. In any case, we can explore alternative estimation methods by changing the method option. The `"RobsonWhitlock"` method does not assume sample uniformity and uses only the two oldest fossils in the point estimator $t_n + (t_n - t_{n-1})$ (Robson & Whitlock 1964, Solow 2003), i.e. adding the last time interval duration to the age of the oldest fossil, and the approximate confidence interval is $[t_n, t_n + P/(1-P) (t_n - t_{n-1})]$.

```
pdate(Fossils, method = "RobsonWhitlock")
estimate    lower    upper
```

```
      56       52      128
```

The point estimate is now a bit younger but the uncertainty is much higher (as reflected in the upper bound), as expected from an estimator that uses fewer datapoints. An alternative estimator uses only the two oldest fossils but with the additional information that each fossil must come from a different descendant lineage: the “ghost-lineage” method (Norris et al. 2015). The resultant distribution is a log-logistic distribution with shape parameter 1 and scale parameter *x*/2 in which *x* is the temporal gap between these two fossils. In the Suliformes record, the second oldest fossil is *Masillastega rectirostris*, a fossil skull from the Messel oil shales that show clear affinities with the Suloidea (Mayr 2002) thus representing the sister group of the Fregatidae.

```
pdate(Fossils, method = "NorrisGhostLin")
estimate    lower    upper
      54       52       90
```

The result of adding this information is an estimate closer to the age of the oldest fossil, and reduced uncertainty. Two additional methods are just alternative parameterizations of methods already used. The `"Beta"` method is based on the fact that t_n/θ has a Beta(n,1) distribution (Wang *et al.* 2009) and uses the `qbeta` function for estimation, producing the same results as the `"StraussSadler"` method. The “penultimate gap” method of Norris et al. (2015) restricts the assumption of uniformity to the two oldest fossils that are used for estimation, does not assume knowledge of subclade affinities, and produces the same results as the `"RobsonWhitlock"` method.

Finally, the “optimal linear estimation” method does not assume sample uniformity yet uses more than just the two oldest fossils: it uses a weighted sum of the *k* oldest ages (Cooke 1980, Robert & Solow 2003). Optimal weights are derived from the fact that, despite the actual distribution of the entire fossil record, the distribution of the *k* oldest ages can be modelled with a Weibull distribution (Cooke 1980, Robert & Solow 2003). This method performs well under different scenarios of non-constant fossilization/recovery potential (Wang *et al.* 2016) and may be the method of choice for rich fossil records (n > 20) that are not distributed uniformly.

```
pdate(Fossils, method = "OLE")
estimate     lower     upper
 63.96709  52.00000 106.80009
```

In the case of the Suliformes, the number of fossils in the set is already fewer than the default for *k* (10) so it is not surprising that the method results in high uncertainty. Another alternative to deal with non-uniform fossil records is to evaluate whether the *k* oldest fossils can be assumed to be uniformly distributed (using a K-S test) and use the `"StraussSadler"` method on those (Cooke 1980).

Obtaining empirical distributions in the face of fossil age uncertainty

In most cases, fossil ages are not known exactly but only minimum and maximum temporal bounds are known, typically corresponding to bounds of standard geochronological units of the geological time scale or the result of chronostratigraphic analysis. The function `clade.date` produces clade age estimates under fossil age uncertainty using a Monte Carlo resampling procedure. The function takes a set of fossil ages as input in the form of a single vector with exact ages, or a two-column matrix indicating the upper and lower bound of the age of each fossil. It then generates an empirical calibration density using random pseudoreplicates generated using the quantile functions corresponding to each estimation method described above. For example, for Strauss & Sadler's (1989) model, $t_n/a^{1/n}$ is the quantile function of the age uncertainty distribution (although not explicitly stated in the literature, this can be demonstrated by the fact that t_n/θ has a Beta(n,1) distribution, Wang *et al.* 2009). Uncertainty in fossil age is modeled as a uniform distribution with bounds determined by lower and upper stratigraphic bounds. In each Monte Carlo replicate, one random age is sampled for each fossil from their corresponding uniform distributions (Figure 2). Then, a random clade age is generated for this pseudosample of fossil ages. The resultant empirical distribution combines the age uncertainty of each fossil with the uncertainty associated with the estimation of the clade's age based on the sample of fossils.

To demonstrate this function, we will estimate the age of the eupasseres, the main subgroup of Passeriformes including suboscine perching birds (Tyranni) and the songbirds (Passeri). The oldest fossil unequivocally representing eupasseres is *Wieslochia weissi*, a nearly complete skeleton found in the Rauenberg clay pits in Germany (Mayr & Manegold 2006). *Wieslochia* is thought to belong into the suborder Tyranni due to the presence a well-developed *processus procoracoideus* of the coracoid and a well-developed *tuberculum ligamenti collateralis ventralis* of the ulna (Mayr & Manegold 2004, 2006, Claramunt & Cracraft 2015), and was recovered as a stem Tyranni in a cladistic analysis (Ksepka *et al.* 2019). Calcareous nannofossils and dinoflagellate cysts indicate a fossil age in the intersection of NP23 and Subzone D14a (Maxwell et al. 2016), thus between 30.2 and 32.0 Ma (Speijer et al. 2020). The first occurrence of eupasseres in the fossil of all other continents was taken from Claramunt & Cracraft (2015). This time, we record minimum and maximum bounds of fossil ages in a matrix in which the rows are fossils, and the columns are minimum and maximum fossil ages, in that order. This matrix can be built by stacking the fossil ages using:

```
Fossils <- rbind(c(30.2, 32.0), c(23.0, 25.0), c(16, 19), c(15.5, 16.5),
c(13.6, 16.0), c(11.6, 16.0), c(5.3, 11.6))
```

Then, the main function is run and the result are stored in an object of class `"clade.date"`:

```
Calib <- clade.date(ages = Fossils, KStest = TRUE, n = 10000)
```

The execution requires only a couple of second. A summary function prints the results in a compact way:

```
summary.clade.date(Calib)

      Exact one-sample Kolmogorov-Smirnov test
data:  Mages
D = 0.19383, p-value = 0.9131
alternative hypothesis: two-sided

Quantiles:
   0%   50%   95%
30.23 33.48 42.97

Parameters of the lognormal function:
 offset meanlog   sdlog
30.2262  1.1792  0.8587
```

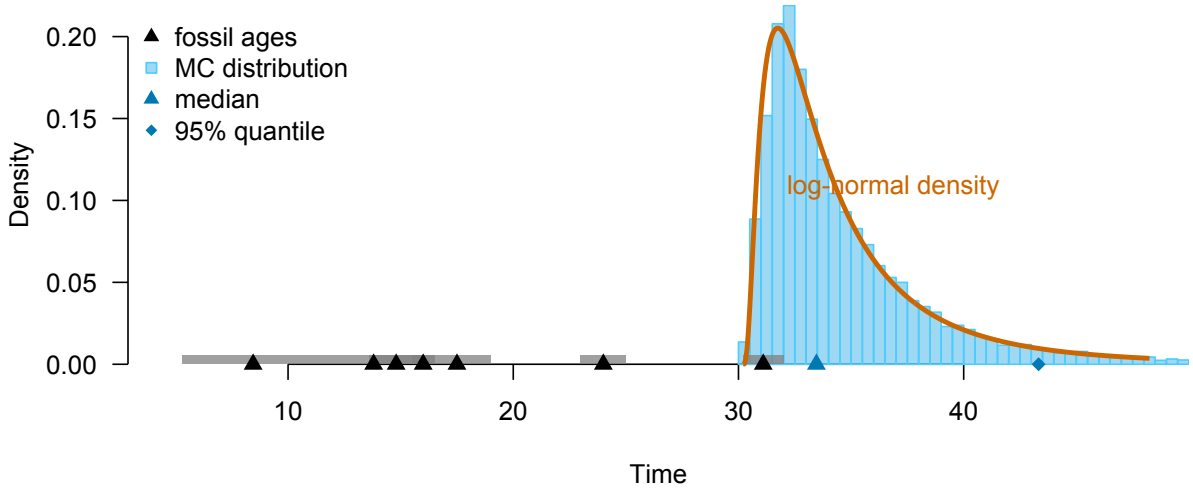
Because the fossil record does not depart significantly from a uniform distribution (Kolmogorov-Smirnov test $p > 0.05$), the use of the `"StraussSadler"` method (the default) is justified, but any of the other methods implemented in `pdate` described above can be used with `clade.date` by changing the `method` option.

In addition to relevant quantiles, `clade.date` reports the estimated parameters of a standard probability function fit to the Monte Carlo sample. `clade.date` fits standard probability functions commonly used in Bayesian time-tree estimation programs. Log-normal, gamma, and exponential densities used in MrBayes (Ronquist et al. 2012) and BEAST2 (Bouckaert et al. 2019) are fit using the `fitdistr()` function in the MASS package (Venables & Ripley 2002), whereas skew-normal and skew-student distributions used in MCMCtree (Yang 2007) are fit with specific functions in the `fGarch` package (Wuertz et al. 2017). In addition to requesting specific functions (i.e. `PDFfit = "lognormal")` the option `"best"` (the default) returns the best function among log-normal, gamma, and exponential models based on the Akaike Information Criterion. In this example, the MC distribution is better fit by a log-normal density function, which estimated parameters plus an offset (30.2 Ma, the minimum age of *Wieslochia* or the 0% quantile) can be used to parameterize a calibration density for eupasseres in BEAST2.

Other options in `clade.date` allow for controlling which quantiles are reported (default: `p = c(0, 0.5, 0.95)`), the number of pseudoreplicates (default: `n = 10000`), the output of individual replicate values (default: `repvalues = TRUE`), and the output of a plot (default: `plot = FALSE`).

A plotting function, plots fossil ages, empirical distributions, quantiles, and probability densities (Figure 3):

```
plot.clade.date(Calib)
```



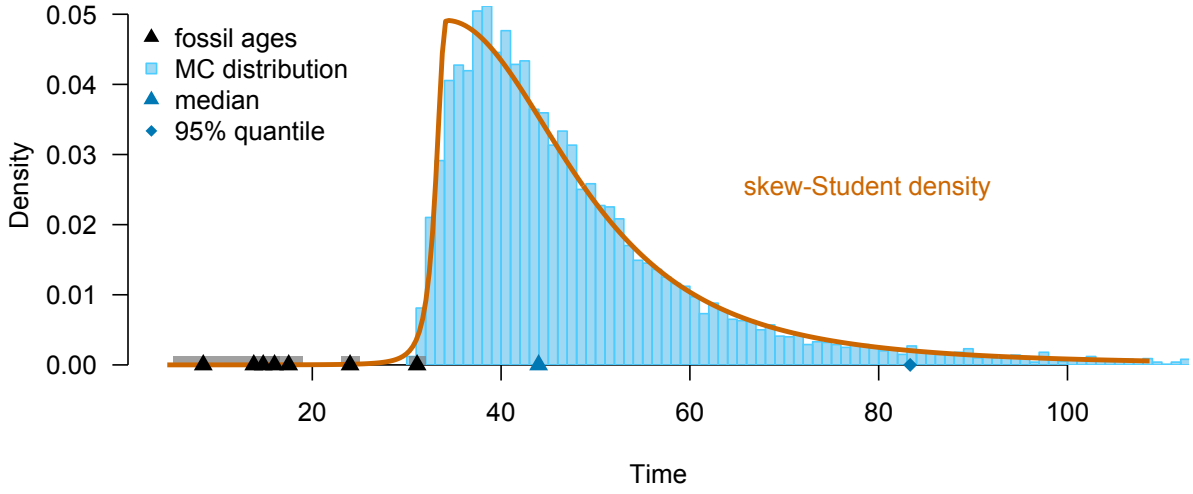
The summary function can also covert the parameters of the probability density function to the parameterization used by MrBayes:

```
summary.clade.date(Calib, param="mrbayes")

Parameters of the lognormal function (MrBayes format):
 offset   mean st.dev.
 65.66   69.01   4.79
```

The final example illustrates the use of the optimal linear estimation method and the fit of a skew-normal distribution that can be used in MCMCtree:

```
clade.date(ages = Fossils, method="OLE", n = 10000, PDFfit="skewStudent",
plot=TRUE)
```



References

Bouckaert R., Vaughan T.G., Barido-Sottani J., Duchêne S., Fourment M., Gavryushkina A., et al. (2019) BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Computational Biology*, *15*, e1006650.

Claramunt, S. & J. L. Cracraft. (2015). A new time tree reveals Earth history's imprint on the evolution of modern birds. *Science Advances*, 1, e1501005.

Cooke, P. (1980). Optimal linear estimation of bounds of random variables. *Biometrika* 67, 257-258.

Gingerich, P. D., & Uhen, M. D. (1998). Likelihood estimation of the time of origin of Cetacea and the time of divergence of Cetacea and Artiodactyla. Palaeontologia Electronica https://palaeo-electronica.org/1998_2/ging_uhen/text.pdf

Ksepka, D. T., Grande L., & Mayr, G. 2019. Oldest finch-beaked birds reveal parallel ecological radiations in the earliest evolution of passerines. *Current Biology* 29(4),657-663.

Maxwell, E. E., Alexander, S., Bechly, G., Eck, K., Frey, E., Grimm, K., Kovar-Eder, J., Mayr, G., Micklich, N., Rasser, M. and Roth-Nebelsick, A. (2016). The Rauenberg fossil Lagerstätte (Baden-Württemberg, Germany): A window into early Oligocene marine and coastal ecosystems of Central Europe. *Palaeogeography, Palaeoclimatology, Palaeoecology* 463, 238-260.

Mayr, G. (2002). A skull of a new pelecaniform bird from the Middle Eocene of Messel, Germany. *Acta Palaeontologica Polonica* 47, 507-512.

Mayr, G., & A. Manegold, A. (2006). New specimens of the earliest European passeriform bird. *Acta Palaeontologica Polonica* 51, 315–323.

Norris, R. W., Strope, C. L., McCandlish, D. M., & Stoltzfus, A. (2015). Bayesian priors for tree calibration: Evaluating two new approaches based on fossil intervals. *bioRxiv* 014340.

Olson, S. L. 1977 A Lower Eocene frigatebird from the Green River Formation of Wyoming (Pelecaniformes: Fregatidae).
Smithsonian Contributions to Paleobiololgy 35: 1–33.

Parham, J. F., Donoghue, P. C., Bell, C. J., Calway, T. D., Head, J. J., Holroyd, P. A., ... & Benton, M. J. (2012). Best practices for justifying fossil calibrations. *Systematic Biology*, *61*(2), 346-359.

Robson, D. S., & Whitlock, J. H. (1964). Estimation of a truncation point. *Biometrika 51*(1):33-39.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., & Huelsenbeck, J. P. (2012). MRBAYES 3.2: Efficient Bayesian phylogenetic inference and model selection across a large model space. Systematic Biology *61*, 539-542.

Smith, N. D. 2010 Phylogenetic analysis of Pelecaniformes (Aves) based on osteological data: implications for waterbid phylogeny and fossil calibration studies. PLoS One 5:e13354.

Solow, A. R. (2003). Estimation of stratigraphic ranges when fossil finds are not randomly distributed. *Paleobiology*, *29*, 181–185.

Speijer, R. P., Palike, H., Hollis, C. J., Hooker, J. J., & Ogg, J. G. (2020). The Paleogene period. Pp. 1087-1140 in Gradstein, F. M., Ogg, J. G., Schmitz, M. D., & Ogg, G. M. (eds.). *The geologic time scale*. Elsevier.

Strauss, D., & Sadler, P. M. (1989). Classical confidence intervals and Bayesian probability estimates for ends of local taxon ranges. *Mathematical Geology*, *21*, 411–427.

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S, Fourth edition*. Springer, New York.

Wang, S. C., & Everson, P. J. (2007). Confidence intervals for pulsed mass extinction events. *Paleobiology* 33(2):324-336.

Wang, S. C., Chudzicki, D. J., & Everson, P. J. (2009). Optimal estimators of the position of a mass extinction when recovery potential is uniform. Paleobiology, 35(3), 2009, pp. 447–459

Wang, S. C., Everson, P. J., Zhou, H. j., Park, D., & Chudzicki, D. J. (2016). Adaptive credible intervals on stratigraphic ranges when recovery potential is unknown. *Paleobiology* 42,(2):240-256.

Wuertz, D., Setz, T., Chalabi, Y., Boudt, C., Chausse, P., & Miklovac, M. (2017) fGarch: Rmetrics - Autoregressive Conditional Heteroskedastic Modelling. Version 3042.83.2. https://CRAN.R-project.org/package=fGarch

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology & Evolution*, *24*:1586–1591.