

# DateLife Workflow

*Luna L. Sanchez Reyes*

*2019-10-08*

DateLife is a service for searching and processing information on ages of any number of taxa of interest, across chronograms available in public data repositories coming from published peer reviewed studies. It can also generate new taxon age information by linking several external services and tested algorithms.

It only requires a set of taxon names as input, in the form of a comma separated listing or vector, or of a phylogeny with taxon names on the tips. Taxon names can correspond to binomial species names or clades. When taxon names are clades, DateLife pulls all accepted species names within the clade (up to OTOL's limit of \_\_\_\_\_ species) from OTOL's reference taxonomy using a service of rphylotastic R package. Names belonging to subspecies or any other infraspecific category are treated as species. DateLife can process input names with the taxon name resolution service (TNRS), which corrects misspelled names or typos, and standardizes variation in spelling and synonyms [Boyle2013], increasing the probability to correctly find the queried taxa in the chronogram database. DateLife uses TNRS to compare names against OTOL's reference taxonomy using a service from the R package rotl [Michonneau2016].

DateLife's main function searches taxon names across the chronogram database specified by the user. At the moment, it queries chronograms from OTOL [Hinchliff2015] repository. DateLife identifies chronograms having at least two taxon names, and subset them to contain only the taxa of interest. It then stores taxon age information from each chronogram individually as a patristic matrix, named with the citation of the original study. This format allows a rapid summary in a number of different ways, including: 1) citations of the original studies containing the subset chronograms, 2) a list of mrca ages of subset chronograms, 3) a list of complete subset chronograms in newick or phylo format, 4) a table containing all information retrieved in html or R's data frame format, or 5) a single chronogram summarized from subset chronograms using the Super Distance Matrix (SDM) supertree construction approach [Criscuolo2006] or using the median of branch lengths.

DateLife also stores information on input taxon presence/absence across subset chronograms. Users can choose to add ages of missing taxa to subset chronograms in different ways, depending on the amount of knowledge they want to input or how much they want to be involved in the steps of the addition process. If users have no access to biological information (i.e., a character, DNA or protein matrix), missing taxa can be added to any chronogram simply at random, or by following taxonomic or phylogenetic knowledge from expert sources. There are a wide number of open reference taxonomies available, such as the Catalogue of Life [Roskov2017] or the NCBI taxonomy database [Federhen2012]. Expert phylogenies (with or without branch lengths) to be used as topological constraint (backbone) can also be obtained from a number of public repositories, such as OTOL [Hinchliff2015], TreeBASE [Piel2002] and Dryad (<https://www.datadryad.org/>). At the moment, DateLife only uses OTOL's synthetic tree and reference taxonomy as expert knowledge to automatically add missing taxa to chronograms. Alternatively, users can input a reference taxonomy or topological constraint of their choosing or making. If OTOL's synthetic tree is not satisfactorily resolved for the taxa of interest, DateLife can construct a sequence data matrix from DNA markers available from the Barcode of Life Database (BOLD; [Ratnasingham2007]), to attempt to further resolve polytomies. It will follow OTOL's synthetic tree as backbone. To use information from a topological constraint, DateLife calls the congruification method described in [Eastman2013] to find shared nodes between trees (congruent nodes). It then fixes their ages, and add ages to remaining nodes with a dating method that can be specified by the user. If users have access to biological data, they can input a tree with branch lengths proportional to relative substitution rates as topological constraint. In this case, age data from congruent nodes will be used as calibration points. Age data from several chronograms can be combined and congruified to be used as calibration points in a single analysis.

Several dating methods are implemented in DateLife. Branch Length Adjuster (BLADJ) is a simple algorithm to distribute ages of undated nodes evenly, which minimizes age variance in the chronogram [Webb2008].

DateLife implements BLADJ from the development R version of phylocom's R package [Webb2008], phylocomr (<https://github.com/ropensci/phylocomr>). It can only be used when there is a topological constraint with no branch lengths. PATHd8 is a non-clock, rate-smoothing method [Britton2007] to date trees. It is also called through R. treePL, is a semi-parametric, rate-smoothing, penalized likelihood dating method [Smith2012]. It is called through R. MrBayes program [Huelsenbeck2001; Ronquist2003] can be used when adding taxa at random, following a reference taxonomy or a topological constraint. It draws ages from a pure birth model, as implemented by Jetz and collaborators [Jetz2012]. DateLife calls MrBayes through an R function.

DateLife can also correct negative branch lengths in several ways.