**Title**

Datelife:

Leveraging databases to study the time frame of origin of species

Leveraging databases to study the time frame of lineage divergence

Mining databases to get closer to a time tree of life

**Authors**

Sánchez-Reyes Luna L., O’Meara B.

**Introduction**

Date of origin/ time of origin of lineages/ time of diversification events/ time

Along with phylogenetic relationships

Constitute the basic information for lineage diversification research (such as the tempo and mode of speciation, extinction and even migration if we have geographical data).

A time frame of lineage origin can be obtained directly from the fossil record. But we commonly rely on ages inferred with molecular dating methods (explain why?).

These types of data have been accumulating in the last years, and there is a large availability of fossil and molecular-based dates of origin data and of phylogenetic relationships in data repositories such as dryad, treebase or open tree of life (How many trees? How many dated trees?)

With new methods such as total evidence and revbayes (fossilized birth-death), studies might include living and fossil lineages.

Also, the amount of time of origin and phylogenetic relationships data is increasing steadily because of better data sharing practices, more and better methods for molecular dating…

Molecular dates are a useful source of data for diversification and biodiversity research but available data have not been exploited because:

Data is in different repositories and formats

Lineage names are different among studies and difficult to reconcile

Taxonomy is also different among studies and difficult to reconcile

Data curation is necessary at some point (is this always true?). At least, the research community views it as an important or even crucial step before data analysis.

It is important to use available data on time frame of lineage origin because:

To know the state of dating for a group of interest

What range of estimated ages exist already? Are fossil and molecular time frames coherent? (e.g., Magallon et al. 2015)

To construct a time tree of life

Science communication, improve scientific discussions, time-framing other events of importance in other research areas.

Recent work on this area (i.e., supersmart and, which others?) aims to:

Generate new dates using all available DNA sequence information.

Performs one global analysis.

Problems or downsides: This might be time consuming for large groups

What lead to datelife development?

Describe public and research necessities covered by datelife…

Importance of datelife:

It allows rapidly obtaining time frame of lineage origin/divergence from already published studies, which are ideally constructed using robust information, such as sequence data and curated fossil calibrations.

Can rapidly construct dates from sequence data if available in BOLD for a set of lineages

Allows direct comparison of dates obtained with different markers available in BOLD (in plants and fungi in particular)

When lineages are not present in any chronograms and do not have sequence data, it can makeup branch lengths with different methods and add them following a reference tree.

It can perform tree dating on a tree with branch lengths proportional to substitution rate using query dates as calibration points (UseAllCalibrations function)

**Description of Datelife**

(BOLD and OToL species names are homogeneous?)

DateLife is a service for getting phylogenetic trees with branch lengths proportional to absolute time from public data repositories. At the moment, it only searches for chronograms in Open Tree of Life repository. It works through the R package datelife (documentation link), a web interface <http://www.datelife.org/query/> and an API (still not up, right?). It is a part of Phylotastic project.

It takes a set of lineage names given by the user, in the form of a listing or of a phylogeny in newick format.

Names can be binomial species names, clade names, common names (will we implement it?).

It can also use the taxon name resolution service from The Open Tree of Life implemented with rotl R package, which allows

**Benchmark:** Testing DateLife computing performance

1. Speed with different amount of lineages and type of analysis

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of lineages | Tol cache search  EstimateDates() | Bold tree  GetBoldOToLTree() | Bold chronogram  GetBoldOToLTree() | Dating  UseAllCalibrations() |
| 3 |  |  |  |  |
| 10 |  |  |  |  |
| 100 |  |  |  |  |
| 1 000 |  |  |  |  |
| 10 000 |  |  |  |  |
| 100 000 |  |  |  |  |

Maybe a graph on computing times…

1. Speed of web interface and of r package (in computers with different capacities?)

**Biological example:** Testing DateLife accuracy

Bird (or reptile) chronograms

Look for all chronograms containing any birds

Or, look for chronograms containing basal lineages

Determine which clade of birds has the more chronograms (have been dated more times) and use that as biological example

**Discussion**

Potential applications demonstrated here.

Improvements, short and long-term.

**Conclusions**

**Availability**

**Supplementary Material**

**Funding**

NSF grant 1458603

NESCent

Open Tree of Life

University of Tennessee, Knoxville

**Acknowledgements**

**References**