title: "Untitled" output: html_document bibliography: library.bib

csl: systematic-biology.csl

Title

Datelife:

Leveraging databases to study the time frame of origin of species/lineages

Leveraging databases to study the time frame of lineage divergence

Mining databases to get closer to a publicly available time tree of life

Authors

Sánchez-Reyes Luna L., O’Meara Brian C.

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 data on time of origin and phylogenetic relationships is increasing steadily because of better data sharing practices, more and better methods for molecular dating, (what else?)

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.

Data curation is largely based on taxonomic knowledge.

It is important to use available data on time frame of lineage origin: 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 2015). To construct a time tree of life. For 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. Perform one global analysis using all available information. Problems or downsides: This might be time consuming for large groups

General issues with dating techniques Sauquet et al. 2012. Testing the Impact of Calibration on Molecular Divergence Times Using a Fossil-Rich Group: The Case of Nothofagus (Fagales) What can datelife do better, or palliate?

What led 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 reconstruct divergence dates for a set of lineages using sequence data from the Barcode Of Life Database (BOLD, http://www.boldsystems.org/). and the synthetic Open Tree of Life (OToL, https://tree.opentreeoflife.org), which relies on the taxonomic knowledge. -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 molecular 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 (chronograms) 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 (add link to documentation), 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. Lineage names can be binomial species, clades, common names (will we implement it?). It can also use the taxon name resolution service from The Open Tree of Life implemented with the rotl R package, which "corrects misspelled names and authorities, standardizes variant spellings, and converts nomenclatural synonyms to accepted names (Boyle et al. 2013)", increasing the probability of the query lineage names to correctly find a match in the chronogram databases.

It detects all chronograms containing at least two query lineages. For each of them, it constructs a subset chronogram containing only matching query lineages (query chronogram). The user chooses to get only the references from the original chronograms from which the query chronograms were constructed, a list of all the query chronograms in newick or phylo format, or a single chronogram summarized from all query chronograms using a supertree method (sdm) or the median of branch lengths (how to better explain this??). It also shows a summary of query lineages not found in any or some chronograms (missing taxa). Then, the user can also choose to add one or all missing taxa to all or some query chronograms by following a reference tree. Different methods to make up these missing branch lengths are implemented and can also be determined by the user: bladj, mrbayes or birth-death models. Explain each...

Finally, a tree from query lineages with branch lengths equivalent to substitution rates can be constructed using barcode markers available through the Barcode of Life Database (BOLD) and following a reference tree, which can be specified by the user. By default it uses the synthetic Open Tree of Life. This molecular tree can then be dated with chronosMPL from the ape package, treePL, PATHd8, mrbayes.

To estimate node ages:

chronosMPL uses mean path length method from Britton et al. 2002

Taxa with no sequence data can also be added following the reference tree using the same methods described before.

How to treat negative branch lengths.

The dates can be used as calibration points for larger trees.

Benchmark: Testing DateLife computing performance

a) Speed with different amount of lineages and types of analysis

---

Number of lineages Tol cache search Bold tree Bold chronogram Dating

```
                    EstimateDates()     GetBoldOToLTree()    GetBoldOToLTree()
UseAllCalibrations()
```

3

10

100

1 000

10 000

100 000

Maybe a graph on computing times…

a) Speed of web interface and of r package (in computers with different capacities?) ```{r include=FALSE}

```

**Biological example: Testing DateLife accuracy**

Bird (or reptile) chronograms, too long time… finches is good nothofagus

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** DateLife can be used through its current website http://www.datelife.org/query/ Or through phylotastic web portalhttp://phylo.cs.nmsu.edu:3000/ DateLife can also be used locally through its R package, which can be installed from CRAN or from the github repository using the devtools R package with the command devtools::install_github("phylotastic/datelife") in R. DateLife source code is available in the following github repository:

**Supplementary Material**

Supplementary material, including code files and online-only appendices, can be found in the Dryad data repository at

**Funding**

Funding was provided by NSF grant 1458603
```

NESCent

Open Tree of Life

University of Tennessee, Knoxville

Acknowledgements

References

References