Dating with constraints
A tutorial on McmcDate

Dominik Schrempf

January 13, 2023

Contents

Abstract

McmcDate can date a phylogenetic tree with constraints. That is, it can estimate the ages of the ancestral nodes of a phylogenetic tree with node age calibrations, node order constraints, and node braces. Mcmc-Date is fast because it approximates the phylogenetic likelihood with a multivariate normal distribution.

In this tutorial, we are going to date a phylogenetic tree of eukaryotes. We are going to use node age calibrations from fossils, and node order constraints from two possible successions of plastid endosymbiosis events.

This tutorial is a work in progress.

1 Definitions and explanations

We use the term *(rooted) tree* to denote a directed acyclic graph with node labels and branch lengths, and in which not more than one branch connects any two nodes. Usually, nodes correspond to (ancestral) species, and branch lengths to sequence distance or time. We use the term *(rooted) topology* to denote a (rooted) tree without information about branch lengths. We put the word *rooted* into parentheses, because it is sometimes omitted. The root of the tree or topology is the only node with all branches pointing away to other nodes, and is the oldest node. Leaves are extant nodes with a single branch pointing towards them.

For a given topology, *dating the tree*[1] corresponds to finding branch lengths for this topology which are measured in absolute time units (for example, in Million years), and which describe the data in the *best* way. *Best* can mean different things, and we will carefully analyze what *best* means in our case. Sometimes, all we can do is date a tree with relative time units (Section 2.1).

Similarly, we use the term *unrooted tree* to denote an undirected acyclic graph with node labels and branch lengths, and in which not more than one branch connects any two nodes. Finally, we use the term *unrooted topology* to denote an unrooted tree without information about branch lengths. We need unrooted trees and topologies because we will use reversible substitution models (for a review, see Yang 2006) to infer unrooted trees with branch lengths measured in expected number of substitutions. Reversible substitution models are unable to discriminate between the two directions in time.

2 Dating a tree

Dating a tree is difficult because we want to estimate branch lengths measured in absolute time units. However, an alignment only contains information about distances measured in expected number of substitutions. In particular, it does not contain information about the evolutionary rates nor about the actual elapsed absolute time units. That is, for a given branch,

$$\begin{aligned} d \text{ in [expected number of substitutions]} = \\ r \text{ in [expected number of substitutions per year]} \cdot t \text{ in [years],} \end{aligned} \tag{1}$$

where d is the branch length measured in expected number of substitutions, r is the evolutionary rate on this branch, and t is the branch length measured in absolute time units. The situation is severe because r and t are confounded! If we multiply t with a positive number c and divide r by the same positive number c, the distance d stays constant. Since the likelihood only depends on the distance d we can not discriminate between parameters

$$t' = t \cdot c, \tag{2}$$
$$r' = r/c, \tag{3}$$

for *any* $c \in (0, \infty)$!

---

[1] We should probably say *dating the topology* but this phrase is not used.

2.1 Node age calibrations

This means we need auxiliary data to constrain the ages of at least some internal nodes. We call constraints on the ages of internal nodes *node age calibrations*. Without node age calibrations, all dated trees are equally likely! Since most phylogenetic dating methods are Bayesian, and impose prior functions on all of their parameters, the values of the posterior function of different trees still differ in a deceiving way. However, in the absence of node age calibrations, the differences in the values of the posterior function are exclusively caused by the chosen prior functions, and not by the information we have obtained from the data.

Even more, in order to limit the maximum ages of the nodes, we require at least one node age calibration with a maximum boundary. The closer in terms of absolute time units the calibrated node with a maximum boundary is to the root, the better. Ideally, we have a maximum age boundary for the root itself, but sometimes the root age boundaries are unknown.

I can not repeat this often enough: **Without proper node age calibrations**, and in particular, **without at least one maximum age boundary**, **dating a tree in absolute time units is impossible**. Without a single maximum node age boundary, we can still date the tree using branch lengths measured in *relative time units*. In particular, we achieve this by fixing the tree height to be 1.0. McmcDate automatically falls back to dating in relative time when no maximum node age boundary is found. If other programs date trees in absolute time units without a maximum age boundary, do not trust the results!

3 Pipeline

The pipeline for dating a tree is:

1. Prepare a multi sequence alignment and an unrooted topology.

2. For this alignment and unrooted topology, infer a distribution of unrooted trees with branch lengths measured in expected number of substitutions.

3. Find a rooted topology to date. Prepare auxiliary data such as node age calibrations or node order constraints.

4. Date the topology with McmcDate. In particular, infer a distribution of rooted trees with branch lengths measured in absolute time units. If the node age calibrations are insufficient, we can still use relative time units.

The following sections describe the steps of the pipeline in detail.

4 Step 1: Alignment and unrooted topology

We are going to date a topology of eukaryotes (Strassert et al. 2021). The full published data set contains three alignments:

1. The alignment with the highest number of taxa comprises 733 eukaryotes with 62723 amino acids.

2. The authors used the first alignment to infer a tree with IQ-TREE (Minh et al. 2020) so they can filter taxa in an informed way (for example, keep slow evolving taxa). The reduced data set comprises 136 operational taxonomic units with 73460 amino acids.

3. A small data set with 63 operational taxonomic units and 73460 amino acids for tree inference with Bayesian models.

Here, we focus on the second alignment with 136 taxa. We use shorter taxon names and I have converted the file from FASTA format to Phylip format which is required by Phylobayes (Lartillot et al. 2013). I provide the alignment `strassert-136taxa.phy` in the data subfolder of this tutorial:

```
1   data/strassert-136taxa.phy
```

For the phylogenetic inference in the next step, we also need an unrooted topology. I have unrooted the tree in Figure 3 in Strassert et al. (2021). Similar to the alignment, I provide the unrooted topology in the data subfolder of this tutorial:

```
1   data/strassert-136taxa.unrooted.tre
```

Please also see the README in the data subfolder.

5   Step 2: Phylogenetic inference

McmcDate is fast because it uses a clever way to approximate the phylogenetic likelihood similar to McmcTree (Yang and Rannala 2005). McmcTree approximates the surface of the phylogenetic likelihood using Taylor expansion (Reis and Yang 2011). That is, in a prior analysis we compute the maximum likelihood together with the gradient and the Hessian matrix which we use in a subsequent analysis to date the tree.

Here, we use a Bayesian method to estimate the posterior distribution. In particular,

- Use Phylobayes (Lartillot et al. 2013).

- Decide on evolutionary model depending on the size of the data set and the computational requirements. Recommended models from preferred but slow and complex to fast and simple: GTR+CAT+G4, LG+CAT+G4, LG+EDM64+G4, LG+C60+G4, LG+G4.

We specify an evolutionary model with exchangeabilities EX, and across-site compositional heterogeneity model ASCH as `EX+ASCH`. All discussed evolutionary models used for simulations as well as inferences implicitly use discrete gamma rate heterogeneity with four components.

- GTR model (Tavaré 1986).

- CAT model (Lartillot and Philippe 2004).

- Gamma rate variation model (Yang 1993).

- LG model (Le and Gascuel 2008).

- EDM model (Schrempf et al. 2020).

- C60 model (Quang et al. 2008)

6   Step 3: Rooted topology and auxiliary data

Strassert et al. (2021) discuss one unrooted topology with two possible root positions. (1) The root separates amorphea from diaphoretickes and excavates (Figure 3 in Strassert et al. 2021), and (2) the root separates amorphea and excavates from diaphoretickes. Here, we choose option (1) which is the more plausible one.

- Node order calibrations (Yang and Rannala 2005).

- Relative node order constraints (Szöllõsi et al. 2022).

- McmcDate can also brace nodes (Appendix A).

7   Step 4: Dating with McmcDate

- McmcDate is a Haskell program (Appendix B).

A   Node braces

B   Internals of McmcDate

McmcDate is a Haskell program.

Recommend cabal, but there is also stack (-s) option.

The wrapper script used in this tutorial tries to make a good compromise between usability and customizability. It exposes some, but not all functionality of McmcDate. Most notably,

- Based on mcmc.

- Based on elynx-tree.

- Explain code a bit (I guess mostly proposals).

B.1   Haskell modules

Modules containing definitions specific to the analysis are in the app subfolder of the McmcDate repository.

B.1.1   More important modules

Definitions Proposals and monitors, configuration.

State State space. If you try to understand what is going on, or if you want to change analysis settings, this should be your starting point.

B.1.2 Other modules

Hamiltonian Hamiltonian proposal.

Main Functions to prepare the data, run and continue the Metropolis-Hasting-Green algorithm.

Monitor Prior specific monitoring functions.

Options Handle command line options.

Probability Prior and likelihood functions.

Tools Miscellaneous tools.

References

Lartillot, N. and H. Philippe (2004). "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process." In: *Molecular Biology and Evolution* 21.6, pp. 1095–1109. DOI: `10.1093/molbev/msh112`.

Lartillot, N., N. Rodrigue, D. Stubbs, and J. Richer (2013). "PhyloBayes MPI: Phylogenetic Reconstruction with Infinite Mixtures of Profiles in a Parallel Environment." In: *Systematic Biology* 62.4, pp. 611–615. DOI: `10.1093/sysbio/syt022`.

Le, S. Q. and O. Gascuel (2008). "An improved general amino acid replacement matrix." In: *Molecular Biology and Evolution* 25.7, pp. 1307–1320. DOI: `10.1093/molbev/msn067`.

Minh, B. Q., H. A. Schmidt, O. Chernomor, D. Schrempf, M. D. Woodhams, A. von Haeseler, and R. Lanfear (2020). "IQ-TREE 2: New models and efficient methods for phylogenetic inference in the genomic era." In: *Molecular Biology and Evolution* 37.5, pp. 1530–1534. DOI: `10.1093/molbev/msaa015`.

Quang, L. S., O. Gascuel, and N. Lartillot (2008). "Empirical profile mixture models for phylogenetic reconstruction." In: *Bioinformatics* 24.20, pp. 2317–2323. DOI: `10.1093/bioinformatics/btn445`.

Reis, M. dos and Z. Yang (2011). "Approximate Likelihood Calculation on a Phylogeny for Bayesian Estimation of Divergence Times." In: *Molecular Biology and Evolution* 28.7, pp. 2161–2172. DOI: `10.1093/molbev/msr045`.

Schrempf, D., N. Lartillot, and G. Szöllősi (2020). "Scalable empirical mixture models that account for across-site compositional heterogeneity." In: *Molecular Biology and Evolution*. DOI: `10.1093/molbev/msaa145`.

Strassert, J. F. H., I. Irisarri, T. A. Williams, and F. Burki (2021). "A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids." In: *Nature Communications* 12.1. DOI: `10.1038/s41467-021-22044-z`.

Szöllõsi, G. J., S. Höhna, T. A. Williams, D. Schrempf, V. Daubin, and B. Boussau (2022). "Relative Time Constraints Improve Molecular Dating." In: *Systematic Biology* 71.4, pp. 797–809. DOI: `10.1093/sysbio/syab084`.

Tavaré, S. (1986). "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences." In: *Lectures on Mathematics in the Life Sciences* 17, pp. 57–86.

Yang, Z. (1993). "Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites." In: *Molecular Biology and Evolution*. DOI: 10.1093/oxfordjournals.molbev.a040082.

Yang, Z. and B. Rannala (2005). "Bayesian Estimation of Species Divergence Times Under a Molecular Clock Using Multiple Fossil Calibrations with Soft Bounds." In: *Molecular Biology and Evolution* 23.1, pp. 212–226. DOI: 10.1093/molbev/msj024.

Yang, Z. (2006). *Computational molecular evolution.* Vol. 284.