Dating with constraints

A tutorial on McmcDate

Dominik Schrempf

January 13, 2023

In this tutorial, we are going to date a phylogenetic tree with constraints. That is, we are going to estimate the ages of the ancestral nodes of a phylogenetic tree with node age calibrations and node order constraints. In general, the rough pipeline is:

1. Prepare a multi sequence alignment, and decide on a rooted tree.
2. Estimate the distributions of branch lengths measured in substitutions per unit time.
3. Prepare auxiliary data such as node age calibrations or node order constraints.
4. Date the phylogenetic tree using McmcDate.

1 Provision of sequence data and a rooted tree

Here, we are going to use data from eukaryotes (Strassert et al. 2021).

2 Phylogenetic inference with Phylobayes

- 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.
- 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)

3 Preparation of node calibrations and node order constraints

- Node order calibrations (Yang and Rannala 2005).
- Relative node order constraints (Szöllősi et al. 2022).
- McmcDate can also brace nodes (Appendix A).

4 Dating with McmcDate

- McmcDate is a Haskell program (Appendix B).

A Node braces

B Internals of McmcDate

- Based on mcmc.
- Based on elynx-tree.
- Explain code a bit (I guess mostly proposals).

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.

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.

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.

Szöllősi, 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.