Phylogenetic Transfer Learning

Michael D. Catchen [1,2]

[1] McGill University   [2] Québec Centre for Biodiversity Sciences

Correspondance to:

Michael D. Catchen — `michael.catchen@mcgill.ca`

Substance: Find a bunch of ASR datasets, do crossvalidation on tip states, test predictive efficacy on inputing 'hidden' species.

Premise: Imputing traits among extant species using ancestral state resconstruction (ASR) can be useful for a wide variety of purposes in ecology

Last revision: *May 6, 2023*

1 **What is this paper?**

2 What is Phylogenetic Transfer Learning

3 Ancestral State Reconstruction

4 *Ancestral State Reconstruction* (ASR; also called ancestral character reconstruction, character estimation, or

5 character mapping) is a core topic in phylogenetics (Joy *et al.* 2016).

6 The goal is, given an estimate of a phylogeny $\mathcal{P}$ and trait values T_i for each extant species in the phylogeny i, to

7 estimate the value of that trait at some historical point in the phylogeny (typically at the node representing the

8 MRCA of a clade of interest).

9 To do this, one assumes a statistical model of evolution (and often tries to infer the best among a set of candidate

10 models). Depending on whether the trait of interest is discrete or continuous, the models are typically

11 discrete-space Markov chains (where the transition matrix is a target of inference) or, in the continuous setting,

12 either Brownian Motion or an Ornstein-Uhlenbeck (OU) process (where the parameters of the model are the

13 target of inference). The former is models neutral evolution and the latter is used when there is hypothesized

14 selection on the trait.

15 The methodology used to fit models has followed the historical progression from maximum parsimony models

16 (a naive approach that favors as few evolutionary changes as possible in discrete traits) to maximum likelihood

17 estimation. Modern methods revolve on using Bayesian methods, which has the direct benefit of including

18 uncertainty estimates in infered ancestral states, and potentially uncertainty in the true topology of the

19 phelogeny itself (Huelsenbeck & Bollback 2001).

20 Typically rate of trait evolution is learned as a single value across the whole phylogeny, and branch length

21 enables 'amount' of evolution.

22 Although in principle a hierarchical* model could be used to infer both a global rate of evolution and rate values

23 specific to each branch (*sadly 'hierarchical' in this is sense different than 'hierarchical' as it is used in

24 Huelsenbeck & Bollback (2001), which refers to different tree models but a single set of parameters across all

25 branches—this is because these models are constructed in the context of DNA substitution rates, which are

26 assumed to be fixed).

Phylogenetic Transfer Learning

The core goal of Phylogenetic Transfer Learning (PTL) is to take a phylogeny $\mathscr{P}$ where the species pool consists of two types of species: (1) species with trait observations, which we denote $\mathscr{O}$ and call the *observed* species and (2) species *without* trait observatoins, which we denote $\mathscr{U}$ and call *unobserved*, and produce predicted trait values for the unobserved species $\mathscr{U}$.

PTL does this by using ASR to infer a parameterized model of evolution and ancestral state of the partially observed trait, and then to simulate the parameterized evolutionary model forward from the ancestral node to get an estimate of trait values (with uncertainty) for each species in U.

The *transfer learning* component in particular comes from the first use of this methods to inpute latent representations of species based on their position in food-webs (Strydom *et al.* 2022a, b), although the method if flexible enough to be applied to either latent or observed traits.

There are two possible models for PTL to be done in: (1) As in (Strydom *et al.* 2022a), the the evolutionary model is inferred only from the observed trait values O. (2) The evolutionary model from a trait for which there are observations available for the entire species pool. It may be the case that evolutionary dynamics inferred with auxillary available information for every species (e.g. the raw sequences from which the tree is built) could improve imputation accuracy.

Substance of the paper

1. Simulated phylogenies

   - Test predictive efficacy of PTL under various parameterization of the "ground-truth" evolutionary dynamics
     - Rate of speciation
     - Rate of evolution
     - Trait dimensionality & correlation
     - Frequency of "puntuactions" (i.e. pertubations to trait value at a speciation event)
   - Predictive efficancy under different data scenarios
     - Proportion of species with trait values

- – Predictive efficacy vs. distance to MRCA w/ data
  - – Do we have a set of traits for all species to infer evolutionary dynaimcs?
    - ∗ How correlated are evolution between traits for all species vs. traits we want to impute

2. Real data

Thankfully there are lots of ACR studies out there with data for each extant species. So, in short, find as many ACR studies as we can,

Relevant questions we want to address: - What phylogenetic scales can PTL give robust predictions - There is a clear upper limit (if the MRCA for set of species is too long ago, there will be massive variance in predicted state at the tip) - There is also a lower limit (if there is horizontal-gene-transfer and the phylogeny is not a reliable proxy for how traits are evolving) - Robustness to noise in trait measurements? - How does this very w/ amount and different types of selection - When is PTL overkill? - Weighted average of neighbors by distance as alternative (ack. David Rolnick for idea)

Possible Applications

The core idea of PTL is to fill in gaps for data-sparse processes in ecology.

- Link prediction
- SDMs: mechanistic niche predictions on subsets of species
- Connectivity (Catchen *et al.* 2023)

Catchen, M.D., Lin, M., Poisot, T., Rolnick, D. & Gonzalez, A. (2023). Improving ecological connectivity assessments with transfer learning and function approximation.

Huelsenbeck, J.P. & Bollback, J.P. (2001). Empirical and Hierarchical Bayesian Estimation of Ancestral States. *Systematic Biology*, 50, 351–366.

Joy, J.B., Liang, R.H., McCloskey, R.M., Nguyen, T. & Poon, A.F.Y. (2016). Ancestral Reconstruction. *PLOS Computational Biology*, 12, e1004763.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022a). Food web reconstruction through phylogenetic transfer of low-rank network representation. *Methods in Ecology and Evolution*, 13, 2838–2849.

79 Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022b). Graph embedding

80   and transfer learning can help predict potential species interaction networks despite data limitations.