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

Keywords:
ancestral state resconstruction
transfer learning
predictive ecology

What is this paper?

The goal of this paper is to introduce PTL to a broad audience, extend the framework from being specific to latent representation of networks (Strydom *et al.* 2022a, b) to arbitrary traits, and provide guidelines for when PTL is appropriate and how to validate its prediction using both simulated and empirical data

1 What is Phylogenetic Transfer Learning

The goal of PTL is to take a species pool for which a given trait is only *partially* observed, and inpute the value of that trait for the rest of the species based on the evolutionary relatedness of each species.

1.1. Ancestral State Reconstruction

Ancestral State Reconstruction (ASR; also called ancestral character reconstruction, character estimation, or character mapping) is a core topic in phylogenetics (Joy *et al.* 2016). 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 estimate the value of that trait at some historical point in the phylogeny (typically at the node representing the MRCA of a clade of interest).

To do this, one assumes a statistical model of evolution (and often tries to infer the best among a set of candidate models). Depending on whether the trait of interest is discrete or continuous, the models are typically discrete-space Markov chains (where the transition matrix is a target of inference) or, in the continuous setting, either Brownian Motion or an Ornstein-Uhlenbeck (OU) process (where the parameters of the model are the target of inference). The former is models neutral evolution and the latter is used when there is hypothesized selection on the trait.

The methodology used to fit models has followed the historical progression from maximum parsimony models (a naive approach that favors as few evolutionary changes as possible in discrete traits) to maximum likelihood estimation. Modern methods revolve on using Bayesian methods, which has the direct benefit of including uncertainty estimates in infered ancestral states, and potentially uncertainty in the true topology of the phelogeny itself (Huelsenbeck & Bollback 2001).

Typically rate of trait evolution is learned as a single value across the whole phylogeny, and branch length enables 'amount' of evolution. Although in principle a hierarchical* model could be used to infer both a global rate of evolution and rate values specific to each branch (*sadly 'hierarchical' in this is sense different than 'hierarchical' as it is used in Huelsenbeck & Bollback (2001), which refers to different tree models but a single set of parameters across all branches—this is because these models are constructed in the context of DNA substition rates, which are assumed to be fixed).

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

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.

2 Substance of the paper

The main substance of this paper is to provide guidelines on when PTL is robust, and diagnostics to validate PTL estimates. This will be done in two parts: **(1)** using simulated phylogenies and trait values to compare efficacy of PTL imputation across known "ground-truth" evolutionary dynamics, and **(2)** using published ASR datasets, with known values for all extant species, to test imputation efficacy empirically.

2.1. Simulated phylogenies First goal is to test predictive efficacy of PTL under various parameterizations of the "ground-truth" evolutionary dynamics, e.g. - Rate of speciation - Rate of evolution - Trait dimensionality & correlation - Frequency of "puntuactions" (i.e. pertubations to trait value at a speciation event)

and second to compare efficacy based on different properties of the data, e.g. - Proportion of species with trait values - Predictive efficacy vs. distance to MRCA w/ data - Is there 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

2.2. 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, and do crossvalidation where we drop the trait values for ~20% of the species and inpute them with PTL, and compare.

2.3. Questions 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)
- What diagnostics can we use to be confident a PTL inputated trait is statistically robust?

Possible Applications

The core idea of PTL is to fill in gaps for data-sparse processes in ecology, so naturally the applications are going to be focused.

Link prediction in networks

This is the inciting question for which the idea was conceived(Strydom *et al.* 2022a). Interactions are hard to sample (**Catchen202MisLin?**). Not much to say here that isn't in (Strydom *et al.* 2022b).

Forecasting species range shifts

Many projections of species range shifts under climate change are based on statistical associations between historical species observations and climatic conditions. The gap between so-called correlative vs. mechanistic species distribution models (Shabani *et al.* 2016) is of critical interest for forecasting species ranges, but robust mechanistic understandings about what and why climatic conditions limit where a species' range require detailed sampling and potentially experimental conditions (Lee-Yaw *et al.* 2016), which are difficult to scale. PTL could potentially alleviate this by giving good proxies of the physiological limitations of species ranges for more species.

Connectivity and movement ecology

Reliable information about species movement is sparse, and PTL could fill this gap (Catchen *et al.* 2023).

Model species and ecological monitoring

There are a lot of species on Earth. Monitoring them all would be hard. Can we use single species a proxies for larger groups of species? Maybe a little. PTL can guide us on what good proxy species are.

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.

Lee-Yaw, J.A., Kharouba, H.M., Bontrager, M., Mahony, C., Csergő, A.M., Noreen, A.M.E., *et al.* (2016). A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits. *Ecology Letters*, 19, 710–722.

Shabani, F., Kumar, L. & Ahmadi, M. (2016). A comparison of absolute performance of different correlative and mechanistic species distribution models in an independent area. *Ecology and Evolution*, 6, 5973–5986.

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.

Strydom, T., Bouskila, S., Banville, F., Barros, C., Caron, D., Farrell, M.J., *et al.* (2022b). Graph embedding and transfer learning can help predict potential species interaction networks despite data limitations.