Phylogenetic signal testing workflow

Tom Smith Jan 2023

Workflow to test for phylogenetic signal in a continuous trait (i.e. do evolutionarily similar species show similar values of the trait in question?).

Examples of our papers using these or similar methods:

- Kontopoulos et al. 2020 "Adaptive evolution shapes the present-day distribution of the thermal sensitivity of population growth rate" PLOS Biology.
- Smith et al. 2022 "Latent functional diversity may accelerate microbial community responses to temperature fluctuations" eLife.

Step 1: align sequences

First you need to create an alignment of sequences. Sequences should be in a fasta file format like this:

```
> sequence_1
AGTGGCCTGTCAA.... etc
> sequence_2
AGTGGCCTGTCAA.... etc

...
```

Sequences have a header begining with ">" followed by a name identifying the sequence, with the sequence on the following line. The file type will be `.fa` or `.fasta`.

Programs for multiple sequence alignment:

- MAFFT - a general tool for alignment: https://mafft.cbrc.jp/alignment/software/
- SINA - specifically align microbial 16S sequences to the SILVA database: https://www.arb-silva.de/aligner/

After alignment, trim sequences so there are no overhangs. The variation at the ends of the sequences may be due to sequencing accuracy, rather than differences between species. There are various free online tools to visualise alignments. Geneious is a nice tool for visualising and editing sequences, but it requires a paid license (though, you can get a limited trial version).

Step 2: Construct phylogeny

Now you want to infer a phylogeny based on the sequence alignment. Various tools, methods, different nucleotide substitution models, etc.

My "go-to" is RAxML, which infers a phylogeny based on maximum likelihood, which I tend to run with a GTR-gamma substitution model: https://cme.h-its.org/exelixis/web/software/raxml/

Interesting walk-through with coded examples of different substitution models: https://revbayes.github.io/tutorials/ctmc/

Instead of maximum likelihood, you can use Bayesian tools to infer a phylogeny, e.g.:

- BEAST: https://beast.community/
- MrBayes: https://nbisweden.github.io/MrBayes/

Step 2.1 (optional): Time-calibrate the phylogeny

You might want to time-calibrate the tree, so branch-lengths represent geological time, rather than sequence divergence.

There's a good discussion of this by Litsios & Salamin (2012) https://academic.oup.com/sysbio/article/61/3/533/1672348 which actually seems to argue that time-calibration is probaly not neccesary, and the inferrences about phylogenetic signal in traits may be more acurate on phylogenies based on sequence divergence.

Some relevant passages:

```
 In practice, time-calibrated trees (or chronograms) are nearly
always chosen over trees depicting molecular changes (or phylograms;
e.g., Crespi and Teo 2002; Jones et al. 2009a; Friedman et al. 2009;
Skinner and Lee 2010). The implicit argument for this practical
choice is that branch lengths estimated using the DNA markers
employed for phylogenetic inference should not be assumed to be
correlated with the rate of phenotypic evolution (Bromham et al.
2002).

 ....

With many of these variables under selection, it is expected that
variation will appear between organisms. As mutations are needed for
any novelty to evolve, a higher substitution rate will increase the
chance of appearance and subsequent evolution of a trait. Under this
paradigm, it is thus most probable that changes in many phenotypic
traits followed closely changes in the rate of DNA substitution and
not a clock-like evolutionary rate (Davies and Savolainen 2006; Smith
and Beaulieu 2009; Seligmann 2010). In such case, ancestral character
state reconstruction will without doubt be less accurate if done on
chronograms than on phylograms.
```

If you *do* want to time-calibrate the phylogeny, PLL-DPPDIV is a useful tool for this: https://cme.h-its.org/exelixis/web/software/dppdiv/index.html

For bacteria, we can't calibrate based on fossils, but we can constrain the divergence of different clades based on data from higher resolution molecular phylogenies. In the papers above we time calibrated based on divergence times given in TimeTree (http://timetree.org/).

Step 3: Test for Phylogenetic Signal

Once we have a phylogeny, we can map our continuous trait to it and test for phylogenetic signal. Good R packages for doing these tests are `ape` and `phytools`.

Basic tests of phylogenetic signal can be done, looking at the metrics Blomberg's K and Pagel's λ. A decent description of these is here: Munkemuller et al. (2012) "How to measure and test phylogenetic signal", *Methods in Ecology and Evolution*.

We can test for these metrics in R using the `phylosig()` function in `phytools`.

```
phylosig(tree,trait,test=TRUE) # Blomberg's K by default
phylosig(tree,trait,method="lambda",test=TRUE) # Pagel's lambda
```

Step 4: Ancestral state reconstruction

So we've tested for phylogenetic signal, afterwards we might want to produce a nice visualisation of evolution of the trait across the tree. To do this, we need to do an ancestral state reconstruction, i.e. estimate what the trait was at each internal node of the tree.

First we need to fit a model to estimate the ancestral states, this can be done with the `fastAnc()` function in `phytools`. Subsequently we can add a projection of this reconstruction on to the tree with `contMap()`.

Further reading

This is a great workshop to read through and run the code for for all sorts of phylogentic comparative methods: http://www.phytools.org/Cordoba2017/

Lessons particularly relevant to our phylogenetic signal and ancestral state reconstruction:

- Exercise 5: Fitting models of continuous character evolution http://www.phytools.org/Cordoba2017/ex/5/Cont-char-models.html
- Exercise 7: Ancestral state reconstruction for continuous characters http://www.phytools.org/Cordoba2017/ex/7/Anc-states-continuous.html