Tree Estimation with phangorn

Amanda Kowalczyk

July 30, 2021

RERconverge includes tree-building functions that perform maximum likelihood branch length estimation given a fixed tree topology and alignments for each sequence of interest. These functions are built directly on phangorn functions `pml` and `optim.pml`, including arguments for parameters passed directly to those functions. For more details on those functions, refer to phangorn documentation.

Input file specification

Tree building functions require two inputs: a master tree topology and alignments from which to estimate branch lengths.

An example master tree file is included at *extdata/mastertree.tree* Example alignment files are included at *extdata/ExampleAlignments/*

```
library(RERconverge)
rerpath = find.package('RERconverge')
mastertreefn=paste(rerpath,"/extdata/mastertree.tree",sep="")
alignmentfn=paste(rerpath,"/extdata/ExampleAlignments/",sep="")
outputfn=paste(rerpath,"/extdata/ExampleTrees.trees",sep="")
```

Run phangorn tree building

The function `estimatePhangornTreeAll` estimates branch lengths for all sequences included in the specified alignment directory. This process is relatively slow - for example, it takes a couple minutes per gene to estimate branch lengths for most genes. The user must specify, at minimum, the alignment director, a master tree file, and a desired output file. Default function behaviors assume alignments are amino acid sequences in fasta format, and other arguments should be specified for other file and sequence types.

- `alndir`: filepath to the directory that contains alignments. Alignment format may be any type specifiable to the phangorn `read.phyDat` function (phylip, interleaved, sequential, clustal, fasta, or nexus)
- `treefile`: filepath to master tree text file in Newick format
- `output.file`: filepath to desired location to save estimated trees. Trees are written in Newick format in a single text file, the proper format to supply to the RERconverge `readTrees` function.
- `format`: string specifying the type of alignment file contained in `alndir`. Defaults to "fasta", and options include "phylip", "interleaved", "sequential", "clustal", "fasta", and "nexus".
- `type`: string specifying sequence type, passed on to phangorn function `read.phyDat`. Defaults to "AA", and options include "DNA", "AA", "CODON", and "USER".
- `submodel`: string specifying the substitution model to use when estimating tree branch lengths. Defaults to "LG", and options include "JC", "F81", "K80", "HKY", "SYM", and "GTR" - see phangorn documentation for additional options.

- `...`: other parameters, such as those specifying model fit parameters, are passed on to phangorn functions `pml` and `optim.pml`.

The code below takes a few mintues to run

```
estimatePhangornTreeAll(alndir=alignmentfn, treefile=mastertreefn, output.file=outputfn)
```

Note that default argument specification is appropriate for amino acid alignments in fasta format and uses the LG substitution model. This may also be specified by including arguments format="fasta", type="AA", and submodel="LG".

For DNA sequences, the general time reversible model (GTR) is a popular substitution model. When using `estimatePhangornTreeAll`, specify this model with the arguments type="DNA" and submodel="GTR".