Ancestral sequence reconstruction with phangorn (Version 2.2.0)

Klaus P. Schliep*

April 3, 2017

1 Introduction

These notes describe the ancestral sequence reconstruction using the *phangorn* package [2]. *phangorn* provides several methods to estimatea ancestral character states with either Maximum Parsimony (MP) or Maximum Likelihood (ML).

2 Parsimony reconstructions

To reconstruct ancestral sequences we first load some data and reconstruct a tree:

```
> library(phangorn)
> library(magrittr)
> fdir <- system.file("extdata/trees", package = "phangorn")
> primates <- read.phyDat(file.path(fdir, "primates.dna"), format = "phylip", type
> tree <- pratchet(primates, trace=0) %>% acctran(primates)
> parsimony(tree, primates)
[1] 746
```

For parsimony analysis of the edge length represent the observed number of changes. Reconstructing ancestral states therefore defines also the edge lengths of a tree. However there can exist several equally parsimonious reconstructions or states can be ambiguous and therefore edge length can differ. "MPR" reconstructs the ancestral states for each (internal) node as if the tree would be rooted in that node. However the nodes are not independent of each other. If one chooses one state for a specific node, this can restrict the choice of neighbouring nodes (figure 2). The function acctran (accelerated transformation) assigns edge length and internal nodes to the tree [3].

*mailto:klaus.schliep@gmail.com

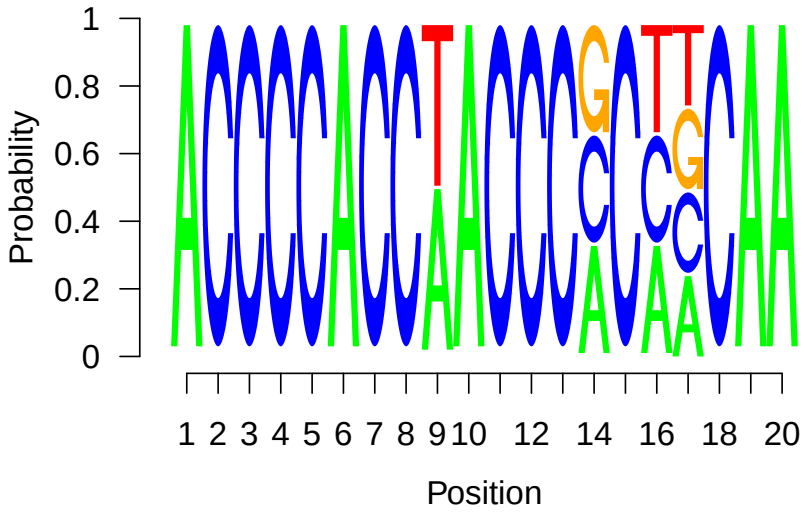

Figure 1: Representation for the reconstruction of the first 20 characters for the root node.

```
> anc.acctran = ancestral.pars(tree, primates, "ACCTRAN")
> anc.mpr = ancestral.pars(tree, primates, "MPR")
```

All the ancestral reconstructions for parsimony are based on the fitch algorithm and so far only bifurcating trees are allowed. However trees can get pruned afterwards using the function `multi2di` from *ape*.

```
> tmp <- require(seqLogo)
> if(tmp) seqLogo( t(subset(anc.mpr, getRoot(tree), 1:20)[[1]]), ic.scale=FALSE)

> par(mfrow=c(2,1))
> plotAnc(tree, anc.mpr, 17)
> title("MPR")
> plotAnc(tree, anc.acctran, 17)
> title("ACCTRAN")
```

3 Likelihood reconstructions

phangorn also offers the possibility to estimate ancestral states using a ML. The advantages of ML over parsimony is that the reconstruction accounts for different edge lengths. So far only a marginal construction is implemented (see [4]).

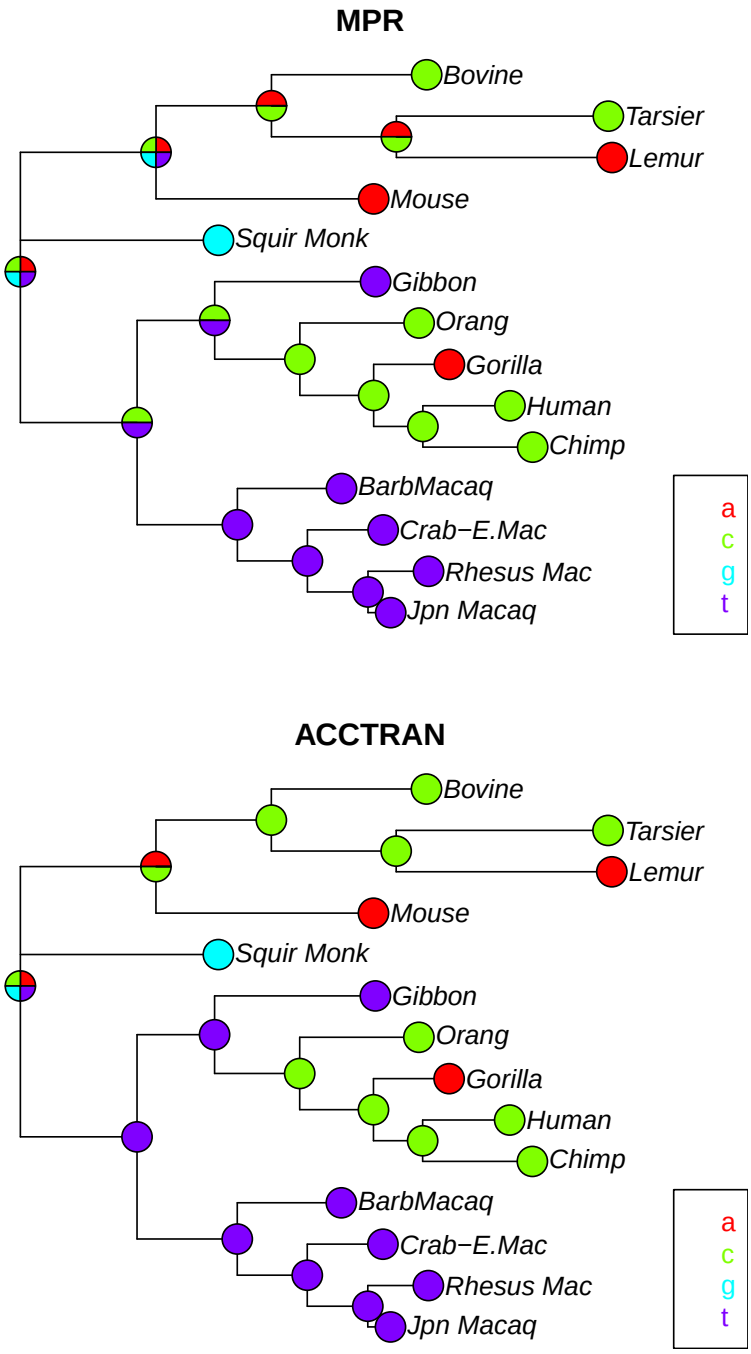

Figure 2: Ancestral reconstruction for one character for the "MPR" and "ACCTRAN" reconstruction. When nodes contain several colours reconstruction is not unique!

```
> fit = pml(tree, primates)
> fit = optim.pml(fit, model="F81", control = pml.control(trace=0))
```

We can assign the ancestral states according to the highest likelihood ("ml"):

$$P(x_r = A) = \frac{L(x_r = A)}{\sum_{k \in \{A,C,G,T\}} L(x_r = k)}$$

and the highest posterior probability ("bayes") criterion:

$$P(x_r = A) = \frac{\pi_A L(x_r = A)}{\sum_{k \in \{A,C,G,T\}} \pi_k L(x_r = k)},$$

where $L(x_r)$ is the joint probability of states at the tips and the state at the root x_r and π_i are the estimated base frequencies of state i. Both methods agree if all states (base frequencies) have equal probabilities.

```
> anc.ml = ancestral.pml(fit, "ml")
> anc.bayes = ancestral.pml(fit, "bayes")
```

The differences of the two approaches for a specific site (17) are represented in figure3.

```
> par(mfrow=c(2,1))
> plotAnc(tree, anc.ml, 17)
> title("ML")
> plotAnc(tree, anc.bayes, 17)
> title("Bayes")
```

References

[1] Emmanuel Paradis. *Analysis of Phylogenetics and Evolution with R.* Springer, New York, 2006.

[2] Klaus Peter Schliep. phangorn: Phylogenetic analysis in R. *Bioinformatics*, 27(4):592–593, 2011.

[3] D.L. Swofford and W.P. Maddison. Reconstructing ancestral character states under wagner parsimony. *Math. Biosci.*, 87:199–229, 1987.

[4] Ziheng Yang. *Computational Molecular evolution.* Oxford University Press, Oxford, 2006.

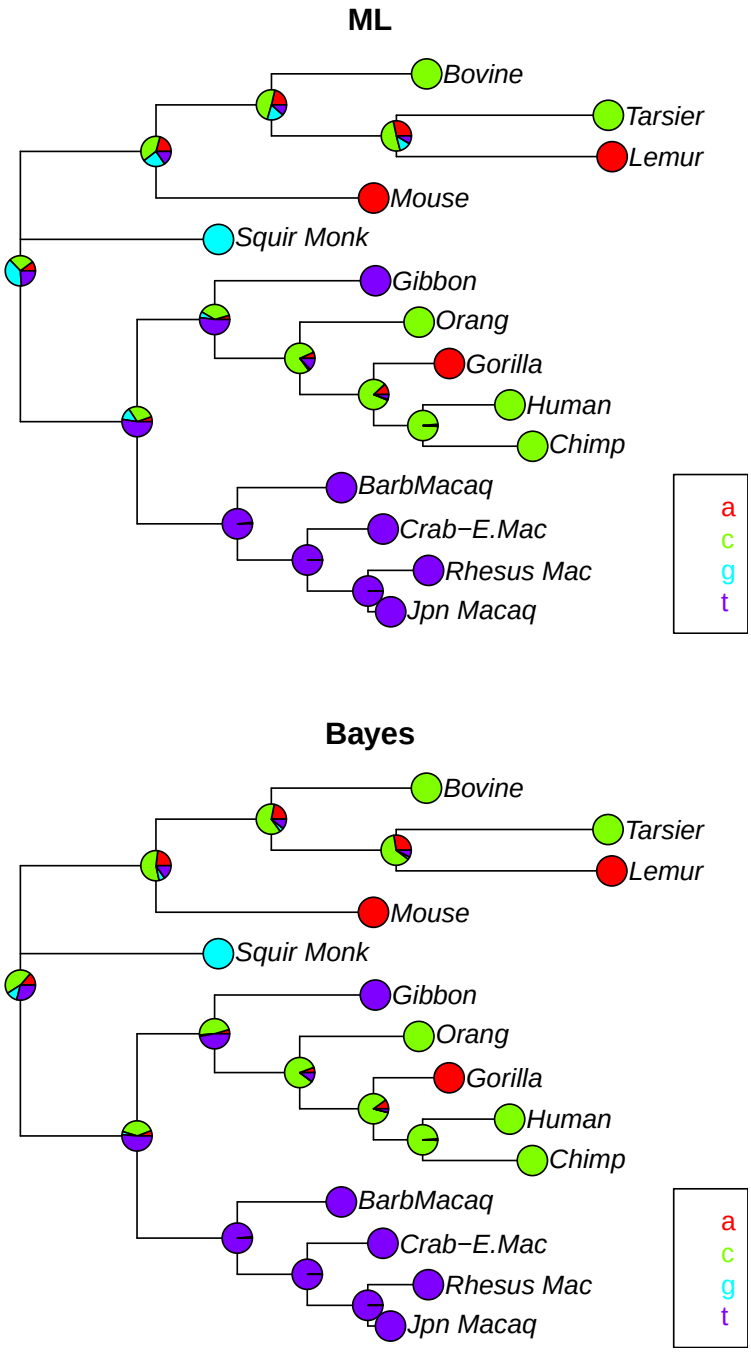

Figure 3: Ancestral reconstruction for fig.2 the using the maximum likelihood and (empirical) Bayesian reconstruction.

4 Session Information

The version number of R and packages loaded for generating the vignette were:

- R version 3.3.3 (2017-03-06), `x86_64-pc-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=C`, `LC_MONETARY=en_US.UTF-8`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, grDevices, graphics, grid, methods, stats, utils
- Other packages: ape 4.1, magrittr 1.5, phangorn 2.2.0, seqLogo 1.40.0
- Loaded via a namespace (and not attached): Matrix 1.2-8, Rcpp 0.12.10, fastmatch 1.1-0, igraph 1.0.1, knitr 1.15.1, lattice 0.20-35, nlme 3.1-131, parallel 3.3.3, quadprog 1.5-5, stats4 3.3.3, tools 3.3.3