# Accounting for and understanding species relationships in mixed models.

*Max.R.Brown*

Ashworth Laboratories, University of Edinburgh.

*ABSTRACT*

Phylogenetic trees are becoming more reliable and more readily available for many more species than ever before. When using a model that contains many species, it may be desirable to account for the relatedness of species and this usually explains a non-trivial amount of variation in the model. Secondly, it may be desirable to calculate a signal gives insight into the pattern of trait evolution, visually aided by mapping species traits to phylogenies. Here I show how the R package MCMCglmm (1) can be used to incorporate phylogenies into linear mixed effect models.

4 March 2020

This short introduction is not exhaustive and does not contain any R code per-se. It merely highlights important steps that need to be taken to incorporate phylogenies into a mixed model framework using a particular R package. Examples of the code will be very shortly available on my GitHub account (https://github.com/euphrasiologist) with perhaps a more explicit tutorial to come.

**1. Importing phylogenies into R**

Luckily, R has plenty of functionality for parsing and visualising phylogenies. These are usually imported as a "phylo" object, created by the ape R package (2). Once in R, phylo objects can be manipulated in many ways and plotted using generic functions in ape. I find it is a nice idea to plot the phylogeny and become familiar with it, as later on this can be useful in interpreting the results. Phylogenies can be parsed into R most commonly in the Newick or Phylip formats. Other formats may be supported and imported using functions built by the community.

**2. Ultrametric trees**

One important attribute of a phylogeny is whether the it is ultrametric or not. An ultrametric tree is one where all the tips are equidistant from the root, and often represents some kind of molecular clock with an assumption that rates of mutation are equal across the tree. This makes a difference when we come to calculating the variance that the phylogeny explains in the model.

**3. Tree building considerations**

Some taxa, for example plants, have well resolved phylogenies for internal, old branches. In studies with limited sampling of a diverse set of species, the internal nodes can be constrained to a known phylogeny, an example here (3). This makes any phylogenetic inference more robust. Secondly, if different sequences have differing mutation rates (for example chloroplast DNA vs nuclear DNA) then the slowly mutating sequences can be aligned across all species and the faster mutating ones within genera or families. These "gapped alignments" can then be passed to software to infer a species tree. A robust phylogenetic software for implementing all of the above is IQ-TREE (4).

**4. Parsing a phylogeny to MCMCglmm**

Firstly I would like to emphasise that the tip labels of the phylogeny all need to match the species names in your dataset. MCMCglmm cannot run without this. I have found that changing the node labels to NULL is important. Edge lengths of the phylogeny cannot be zero, and if they are, I have found that you can change them to a very small number (e.g. 1e-10), this solves the problem. Once all these changes are made, the phylo object can then be given to the MCMCglmm::inverseA() function. Two arguments of this function are particularly important. One is "nodes" - here you will have to specify a character vector ("TIPS" or "ALL") to the argument. If tips only is specified, not only is the function slower to run, and the eventual model also slower (by orders of magnitude), but in the end you only get posterior modes of the species effect estimates and not one for every node. My recommendation would be to always have the argument as "ALL" unless you have particular reason. The second important argument is "scale". If your tree is ultrametric, then you can go ahead and specify "scale = TRUE". If not, the oppostie should be specified and noted that some calculations will have to be changed later.

After taking into account the above, the tree can now be passed to the MCMCglmm::MCMCglmm() function. Fixed effects can be specified in any way but the random effects are different. What I like to do is to create a duplicate column of the species/taxa and call it "animal". So there are now two columns in the data, your "species" column which deals with the between species effects and your "animal" column which deals with the effect of phylogenetic non-independence. Most simply, both "species" and "animal" can then be specified as single parameter random effects. The MCMCglmm::MCMCglmm() function has an argument "ginverse" which takes the output of MCMCglmm::inverseA() as a named list, where the name is identical to the column specifying the phylogenetic effects. Here, this name would be "animal".

## 5. Interpretation of the phylogenetic effects

Once the model has been run, the model object will contain iterations drawn from the posterior distributions of the fixed and random effects. These can be accessed through the $Sol (Solutions) and $VCV (Variance-Covariance) slots. If the "pr" argument in the MCMCglmm::MCMCglmm() function was TRUE, then all of the point estimates for the species level and phylogenetic random effects will have been saved in the $Sol slot. Running the summary() function on the model object will produce a pretty output of much relevant information. For the purposes here, two sections are particularly important. These are the "G-structure" which gives a summary of the random effects specified and the R-structure which gives information about the residual variance in the model. From this output, one may already be able to see if there is an effect of phylogeny by eyeing up the posterior means of the variance estimates.

## References

1. Hadfield, J. D., “Mcmc Methods For Multi-Response Generalized Linear Mixed Models: The Mcmcglmm R Package,” *Journal Of Statistical Software* **33**(2), pp. 1-22 (2010).
2. Paradis, E., Claude, J., and Strimmer, K., “Ape: Analyses Of Phylogenetics And Evolution In R Language,” *Bioinformatics* **20**(2), pp. 289-290 (2004).
3. Heckenhauer, J., Abu Salim, K., Chase, M. W., Dexter, K. G., Pennington, R. T., Tan, S., Kaye, M. E., and Samuel, R., “Plant Dna Barcodes And Assessment Of Phylogenetic Community Structure Of A Tropical Mixed Dipterocarp Forest In Brunei Darussalam (Borneo),” *Plos One* **12**(10) (2017).
4. Nguyen, L. T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q., “Iq-Tree: A Fast And Effective Stochastic Algorithm For Estimating Maximum-Likelihood Phylogenies,” *Molecular Biology And Evolution* **32**(1), pp. 268-274 (2015).