DEsiR: Discrimination Estimation in R

Kevin Healy, Thomas Guillerme & Andrew L Jackson

8 April 2016

This package estimates Trophic Discrimination Factors (TDF) based on the imputation function within the `MCMCglmm` package and includes the error associated with building phylogenetic trees using the `MulTree` package.

Installation

To install `DEsiR` its dependency `MulTree` package must first be installed directly from GitHub using the following:

```
if(!require(devtools)) install.packages("devtools")
install_github("TGuillerme/mulTree", ref = "master")
```

Following this you can then install `DEsiR` directly from GitHub using the following:

```
install_github("healyke/DEsiR", ref = "master")
```

And load in the data

```
library(mulTree)
library(DEsiR)
```

Read in the data

```
#read in the data
DEsiR.data <-read.csv(file=system.file("extdata", "DEsiR_data.csv", package = "DEsiR"),
 header=T, stringsAsFactors = F)

head(DEsiR.data)
```

```
##             species     habitat taxonomic.class tissue diet.type
## 1 Rattus_norvegicus terrestrial        mammalia  liver herbivore
## 2 Rattus_norvegicus terrestrial        mammalia  liver carnivore
## 3 Rattus_norvegicus terrestrial        mammalia  liver  omnivore
## 4 Rattus_norvegicus terrestrial        mammalia  liver  omnivore
## 5 Rattus_norvegicus terrestrial        mammalia  blood herbivore
## 6 Rattus_norvegicus terrestrial        mammalia  blood carnivore
##   source.iso.13C source.iso.15N delta13C delta15N
## 1          -25.3            4.8      3.1      1.2
## 2          -16.2           12.3      3.1      0.6
## 3          -24.5            7.1      1.5      3.2
## 4          -16.7            7.4      2.2      3.0
## 5          -25.3            4.8      1.5      3.6
## 6          -16.2           12.3      0.7      3.7
```

```
#read in the phylogenetic information
#first the mammal trees
mammal_trees <- read.tree(system.file("extdata",
                                      "3firstFritzTrees.tre",
                                      package = "DEsiR"))
#than the bird trees
bird_trees   <- read.tree(system.file("extdata",
                                      "3firstJetzTrees.tre",
                                      package = "DEsiR"))

#combine them together using the tree.bind function from the mulTree package
combined_trees<-tree.bind(x = mammal_trees, y = bird_trees, sample = 2, root.age = 250)
```

As may we want to include the error associated with building phylogenies into our analysis we take a sample of the possible trees (See Healy et al 2014). In this case we combine the mammal and bird phylogenies at a rooted age of 250 mya and take a sample of two possible trees.

Testing the new data: `set.tef.est`

In order to estimate a trophic enrichment factor for a new species we need to check that the species is already present in our phylogeny and check what data is available for the new species.

set.tef.est checks for the presence of the following data: tissue type ("blood", "claws", "collagen", "feather", "hair", "kidney", "liver", "milk", "muscle"), habitat ("terrestrial", "marine"), diet.type ("carnivore", "herbivore", "omnivore", "pellet"), and the isotopic ratios of the food sources for carbon (source.iso.13C) and nitrogen (source.iso.15N).

```
#####function that checks the dat for some species we want to estimate TEF for
new.data.test <- setTefEst(species = "Meles_meles",
                           habitat = "terrestrial",
                           taxonomic.class = "mammalia",
                           tissue = "blood",
                           diet.type = "omnivore",
                           source.iso.13C = c(-24.1),
                           source.iso.15N = c(7.0),
                           tree = combined_trees)
```

```
## Meles_meles present in phylogeny
```

If the species is not in the phylogeny already, such as say the Komodo dragon (Varanus komodoensis), or we are missing values as donated by NA, say tissue type, we get warning messages to indicate what is missing from our data.

```
## Warning in setTefEst(species = "Varanus_komodoensis", habitat = "terrestrial", : tissue levels do no
##      from Healy et al 2016
```

```
## Varanus_komodoensis not present in phylogeny
```

Formatting the new data: `tef.mul.clean`

We now need to format the data by combining both the isotopic data already available within the package and the data from the new species and matching these species to the included phylogeny. We also include what isotope we want to estimate a trophic discrimination value for (either "carbon" or "nitrogen").

```
tdf_data_c <- tefMulClean(new.data = new.data.test,
                          data = DEsiR.data,
                          species.col.name = "species",
                          tree =
                          combined_trees,
                          taxonomic.class = "mammalia",
                          isotope = "carbon")
```

```
## The following taxa were dropped from the analysis:
##  Tachyglossus_aculeatus Zaglossus_bruijni Zaglossus_bartoni Ornithorhynchus_anatinus Anomalurus_beec
```

We now have a mulTree class object, which is required by the imputation analysis. It contains the matched phylogenies, in this case two phylogenies

```
tdf_data_c$phy
```

```
## 2 phylogenetic trees
```

and a dataset containing TDF and related data with the new species for which you want to estimate a trophic enrichment factor at the top with a NA for either delta13C or delta15N depending on isotope.

```
head(tdf_data_c$data)
```

```
##                sp.col     habitat taxonomic.class tissue diet.type
## 1         Meles_meles terrestrial        mammalia  blood  omnivore
## 120 Rattus_norvegicus terrestrial        mammalia  liver herbivore
## 2   Rattus_norvegicus terrestrial        mammalia  liver carnivore
## 3   Rattus_norvegicus terrestrial        mammalia  liver  omnivore
## 4   Rattus_norvegicus terrestrial        mammalia  liver  omnivore
## 5   Rattus_norvegicus terrestrial        mammalia  blood herbivore
##     source.iso.13C delta13C           animal
## 1            -24.1       NA       Meles_meles
## 120          -25.3      3.1 Rattus_norvegicus
## 2            -16.2      3.1 Rattus_norvegicus
## 3            -24.5      1.5 Rattus_norvegicus
## 4            -16.7      2.2 Rattus_norvegicus
## 5            -25.3      1.5 Rattus_norvegicus
```

Running the analysis: `tef.mul.clean`

With the data formatted as a mulTree object we now need to set up the model. In this case we will run the full model to estimate delta13C with the fixed factors of source isotopic ratio, diet type and habitat type.

```
formula.c <- delta13C ~ source.iso.13C + diet.type + habitat
```

and random terms that includes the "animal" term which is required to include phylogeny into the analysis, sp.col to allow multiple species entries into the analysis and account for variation within a species, and tissue type.

```
random.terms = ~ animal + sp.col + tissue
```

As we rely on Bayesian imputation to estimate the missing value we also need to specify a prior, in this case we use a non-informative prior as recommended in the MCMCglmm guidelines.

```
prior <- list(R = list(V = 1/4, nu=0.002),
              G = list(G1=list(V = 1/4, nu=0.002),
                       G2=list(V = 1/4, nu=0.002),
                       G3=list(V = 1/4, nu=0.002)))
```

along with the number of iterations to run the chain (nitt), the burn-in (burnin), the sampling thinning (thin), the number of chains to run (no.chains). (See MCMCglmm guidelines.)

```
nitt <- c(1200000)
burnin <- c(200000)
thin <- c(500)
no.chains <- c(2)
```

We need to check that our MCMC chains are converging so we use the Gelman and Rubin diagnostic to check the convergence and also check that the estimated parameters have an effective sample size >1000.

```
convergence =  c(1.1)
ESS = c(1000)
```

As the function exports the model output to avoid memory issues within R make sure you have set the working directory to somewhere appropriate.

```
getwd()
```

```
## [1] "/Users/kevinhealy/Desktop/Enrichment_factors/DEsIR/doc"
```

```
list.files()
```

```
## [1] "Introduction-to-DEsiR.html" "Introduction-to-DEsiR.pdf"
## [3] "Introduction-to-DEsiR.Rmd"
```

finally we can run the analysis. This model will take approximatly 5 min but for larger nitt or larger samples of phylogentic trees it will take longer.

```
TDF_est.c <- tefMcmcglmm(mulTree.data = tdf_data_c,
                         formula = formula.c,
                         random.terms = random.terms,
                         prior = prior,
                         output = "test_c_run",
```

```
                          nitt = nitt,
                          thin = thin,
                          burnin = burnin,
                          no.chains = no.chains,
                          convergence =  convergence,
                          ESS = ESS)
```

```
##
## 2016-04-11 - 17:02:55: MCMCglmm performed on tree 1
## Convergence diagnosis:
## Effective sample size is > 1000: TRUE
## 2180.613; 2679.903; 2000; 2000; 3581.691; 2000; 2000; 2000; 1857.273; 2348.312; 1862.905; 2000
## All levels converged < 1.1: TRUE
## 1.000181; 1.001642; 0.999773; 1.000469; 1.002796; 1.000472
## Individual models saved as: test_c_run-tree1_chain*.rda
## Convergence diagnosis saved as: test_c_run-tree1_conv.rda
##
## 2016-04-11 - 17:08:03: MCMCglmm performed on tree 2
## Convergence diagnosis:
## Effective sample size is > 1000: TRUE
## 1844.997; 1992.549; 2000; 2000; 2000; 2000; 3133.528; 2000; 2000; 2000; 1821.501; 2149.526
## All levels converged < 1.1: TRUE
## 1.000375; 1.000362; 1.000712; 1.000802; 1.001246; 1.000602
## Individual models saved as: test_c_run-tree2_chain*.rda
## Convergence diagnosis saved as: test_c_run-tree2_conv.rda
##
## 2016-04-11 - 17:08:03: MCMCglmm successfully performed on 2 trees.
## Total execution time: 9.48836 mins.
## Use read.mulTree() to read the data as 'mulTree' data.
## Use summary.mulTree() and plot.mulTree() for plotting or summarizing the 'mulTree' data.
```

tefMcmcglmm now runs the selected amount of chains (no.chains) for each of the sampled phylogeny and exports the resulting MCMC chains to the working directory. Hence in this case we have run two chains for each of the two sampled trees so there should be four files in your working directory ending with something like run-tree1_chain2.rda.

```
list.files()
```

```
## [1] "Introduction-to-DEsiR.html"  "Introduction-to-DEsiR.pdf"
## [3] "Introduction-to-DEsiR.Rmd"   "test_c_run-tree1_chain1.rda"
## [5] "test_c_run-tree1_chain2.rda" "test_c_run-tree1_conv.rda"
## [7] "test_c_run-tree2_chain1.rda" "test_c_run-tree2_chain2.rda"
## [9] "test_c_run-tree2_conv.rda"
```

These are the full MCMCglmm model outputs for each of the chains run and can be imported back into the R for full inspection if required using read.mulTree. Notice also the two other files ending with something simlar to run-tree2_conv. These give a discriotion on the convergance diagnostics of the chains for each tree. The results of these diagnostics for each tree are also printed in R after running tefMcmcglmm returing whether the Effective sample size exceded ESS for all estimated paramaters (with the number for each parameter given as a list), and whether al chains converged for each parameter.

All these models can be seperatly imported into R using read.mulTree(). However, since we are only intrested in the estimated TDF for our species tefMcmcglmm only imports the imputed posterior distribution of the estimated TDF for our species.

```r
names(TDF_est.c)
```

```
## [1] "tef_estimates" "tef_global"
```

```r
summary(TDF_est.c$tef_global)
```

```
##
## Iterations = 1:8000
## Thinning interval = 1
## Number of chains = 1
## Sample size per chain = 8000
##
## 1. Empirical mean and standard deviation for each variable,
##    plus standard error of the mean:
##
##           Mean             SD       Naive SE Time-series SE
##        2.83183        1.48905         0.01665        0.01665
##
## 2. Quantiles for each variable:
##
##    2.5%     25%     50%     75%   97.5%
## -0.1576  1.8615  2.8651  3.7967  5.6965
```

```r
hist(TDF_est.c$tef_global)
```