Quick Start Guide for 'LRTT' Package

Chao ZHOU, Tao WANG

3/26/2018

Department of Bioinformatics and Biostatistics, Shanghai Jiao Tong University, Shanghai, China

SJTU-Yale Joint Center for Biostatistics, Shanghai Jiao Tong University, Shanghai, China

1 Overview

Microbes play an important role in the world, both for our humans life and our environments. Here we developed one differential anbudance analysis incorporating the phylogeny which described by the phylogenetic relationship, log ratio tree test **LRTT**. The follow steps show you how to use this packages, if any problem please free free to contact Chao ZHOU at Supdream8@sjtu.edu.cn

2 Workflow

2.1 Install packages

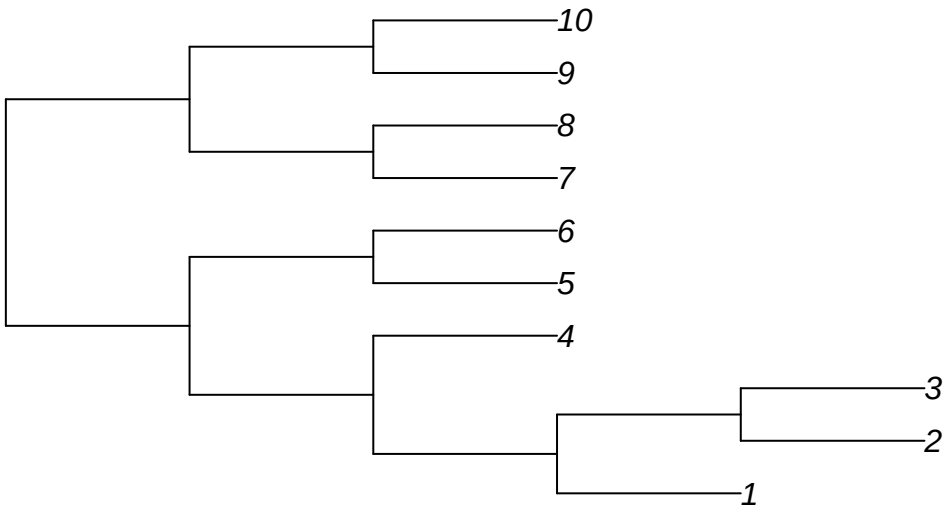
It is easiest to use the 'devtools' packages to install the 'LRTT'.After installed 'LRTT', you can use R help to see functions or data about this packages.You can also get the rar document from web to install it on local. There are other serveral packages will be used in follow steps.

```
#install.packages("devtools")
#library(devtools)
#install_github("ZRChao/LRTT")

library(LRTT)
library(help = LRTT)

library(ape)
library(dirmult)
library(mvtnorm)
```

2.2 Data pro-precess

OTU table have to corresponds to the tree structure. Firstly, we introduce four different simulation methods, you can try any one you like. You can also skipped this step to use the example data we provided as `load(Sim.data)` or package MiSPU gives.Four simulation situation methods, `BIT.Sim` and `DTM.Sim` is based on the tree structure and return the count data includes the leaves and internal nodes of the tree; `LNM.Sim` and `ANCOM.Sim` just correspond to the differential OTU and return the OTU table. But we have bulid the index matrix function Taxa.index to return the relationship between leaves and internal nodes, the rows of the matrix is the leaves (OTU) and columns are internal nodes, value 1 of the matrix means there are connected by branches otherwise is 0. Use this matrix, we can multiple it to the OTU table to get the internal nodes count. which have the same function as `caper::clade.members.list*`.

```
p <- 10
tree <- Tree.Sim(p)      # the same effect as ape:rtree
plot.phylo(tree, type = "phylogram")
```




```
taxa.index <- Taxa.index(p, tree)
taxa.index               # the first column is the root
```

```
##       11 12 13 14 15 16 17 18 19
##  [1,]  1  1  1  1  0  0  0  0  0
##  [2,]  1  1  1  1  1  0  0  0  0
##  [3,]  1  1  1  1  1  0  0  0  0
##  [4,]  1  1  1  0  0  0  0  0  0
##  [5,]  1  1  0  0  0  1  0  0  0
##  [6,]  1  1  0  0  0  1  0  0  0
##  [7,]  1  0  0  0  0  0  1  1  0
##  [8,]  1  0  0  0  0  0  1  1  0
##  [9,]  1  0  0  0  0  0  1  0  1
## [10,]  1  0  0  0  0  0  1  0  1
```

To simulation, based on the tree structurewe have to give probability on each branch which must statisfy that sums equal to 1. So after this setting, we can calculate the probability multiple along the branch to get the probability for the leaves which also statisfy all the probablity on the leaves sums to 1, otherwise something wrong. Set the dif.taxa have different probability for his children. Then, we can check which OTU is differential one.

```
set.seed(2)
dif.taxa <- sample(tree$edge[, 1], 1)
prob <- Prob.branch(tree, seed = 1, dif.taxa)

prob.m1 <- Prob.mult(p, tree, prob[, 1])
prob.m2 <- Prob.mult(p, tree, prob[, 2])

dif.otu <- which(prob.m1 != prob.m2)
dif.otu
```

```
## [1] 1 2 3
```

Based on the above parameters, we can simulate different count matrix with different distribution: Binomial Tree, Dirichlet Tree Multinomial which is based on the tree structure and will return the count of leaves and internal nodes; There can be some degenetation on the leaves which is small probability things. With the

above settings and parameters, we can do next four simulations.

```
## BIT and DTM will return all count data
data.bit <- BIT.Sim(p, seed = 2, N = 20, tree = tree, prob_control = prob[, 1],
                    prob_case = prob[, 2])
data.dtm <- DTM.Sim(p, seed = 2, N = 20, tree, prob[, 1], prob[, 2], theta = 0.1)
dim(data.bit)
```

```
## [1] 40 18
```

```
dim(data.dtm)
```

```
## [1] 40 18
```

```
## LNM and ANCOM will return only OTU count data
data.lnm <- LNM.Sim(p, seed = 2, N = 20, dif.otu)
data.ancom <- ANCOM.Sim(p, seed = 2, N = 20, dif.otu)
dim(data.lnm)
```

```
## [1] 40 10
```

```
dim(data.ancom)
```

```
## [1] 40 10
```

```
data.alllnm <- cbind(data.lnm %*% taxa.index[, -1], data.lnm)
dim(data.alllnm)
```

```
## [1] 40 18
```

You also can skipped this step to use the example data we provided as or package `MiSPU` gives.

```
data("Sim.data")
summary(Sim.data)
```

```
##             Length Class      Mode
## BIT         100    data.frame list
## DTM         100    data.frame list
## LNM         100    data.frame list
## ANCOM       100    data.frame list
## tree          6    phylo      list
## dif.otu      24    -none-     numeric
## dif.taxa      5    -none-     numeric
## prob.branch   2    data.frame list
## group       100    -none-     numeric
```

2.3 Tree Ratio test

Based on the tree, we use the least log ratio transform to do differential analysis. Once the parents is has different probability to his children, which means their children's probability is different, so the ratio must be different between of different group. So, we just do log ratio test of each brothers to decides whether it is differential or not.

```
data.bit <- as.matrix(Sim.data$BIT)
grouplabel <- Sim.data$group
dim(data.bit)
```

```
## [1] 100 100
```

```
tree <- Sim.data$tree
p <- min(tree$edge[, 1]) -1
p
```

```
## [1] 100
```

```
taxa.index <- Taxa.index(p, tree)
colnames(data.bit) <- as.character(1:p)
all.bit <- cbind(data.bit %*% taxa.index , data.bit)

tree.results <- Tree.ratio(p, tree, taxa.index, all.tab = all.bit , group = grouplabel)
str(tree.results)
```

```
## List of 4
##  $ taxa.pvalue: num [1:99] 0.1843 0.0933 0.5605 0.2286 0.4985 ...
##  $ otu.dif    : Named logi [1:100] FALSE FALSE FALSE FALSE FALSE FALSE ...
##   ..- attr(*, "names")= chr [1:100] "1" "2" "3" "4" ...
##  $ alltab     : num [1:100, 1:199] 165931 64850 28016 187687 66052 ...
##   ..- attr(*, "dimnames")=List of 2
##   .. ..$ : NULL
##   .. ..$ : chr [1:199] "101" "102" "103" "104" ...
##  $ taxa.dif   : chr [1:4] "122" "128" "137" "154"
```

The results of of `Tree.ratio` include data after pruned `alltab`, internal node's pvalue `taxa.pvalue`, differential internal nodes `taxa.dif` which after adjustment on each multiple test, `dif.otu` which show is differential one. However, we have to noticed that one child OTU on the leaves of the tree is differential one, it must because his ancestor have different probability for his children, while once ancenstor have differential probability for his children, his children will not be differential one. So here, we have to check the differential OTU finded `Tree.ratio`.

```
tree.dectected <- Tree.ratio.back(p, tree.ratio = tree.results, taxa.index,
                                  otutab = data.bit, group = grouplabel)
tree.dectected
```

```
##  [1] "18" "25" "26" "27" "35" "36" "37" "38" "39" "53" "54" "55" "56" "57"
```

which return the differential OTU by correction step. Some times you may interested the differential internal nodes which can see `result$taxa.dif`.Compare to other methods here, we plot the venn plot.

```
# t detected
t.pv <- apply(data.bit/rowSums(data.bit), 2, function(x)
            return(t.test(x[grouplabel==1], x[grouplabel==2])$p.value))
t.pv.adj <- p.adjust(t.pv, method = "fdr")
t.detected <- which(t.pv.adj <= 0.05)

# zig detected
```

Reference

Chao ZHOU, Tao WANG. Differential Abundance Analysis for Microbiome data Incorporating phylogeny. 2018 (Under review).

Mandal, Siddhartha, et al. "Analysis of composition of microbiomes: a novel method for studying microbial composition." Microbial ecology in health and disease 26.1 (2015): 27663.

Xia, Fan, et al. "A logistic normal multinomial regression model for microbiome compositional data analysis." Biometrics 69.4 (2013): 1053-1063.

Torben Tvedebrink (2010). Overdispersion in allelic counts and theta-correction in forensic genetics. Theoretical Population Biology, 78(3), 200-210.