R语言计算Alpha多样性指数

本篇以某16S扩增子测序所得细菌群落数据为例，简介使用R计算几种在微生物群落分析中常用的Alpha多样性指数。

关于下文出现的各Alpha多样性指数的概念，可[点击查看简介](http://blog.sciencenet.cn/blog-3406804-1179809.html)。

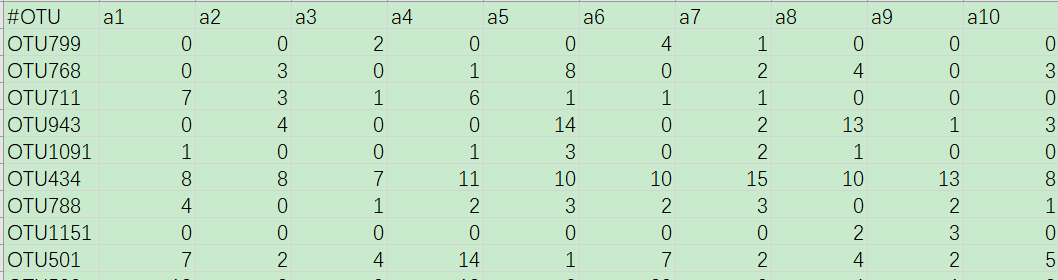
示例文件、R脚本等，已上传至百度盘，提取码xim2。

<https://pan.baidu.com/s/1fUiG1WMK6T8cpAQg4trXDg>

## 示例文件说明

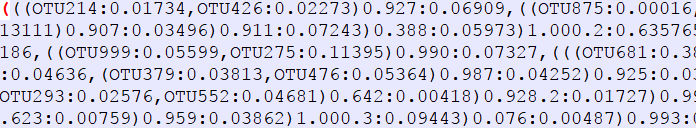
文件“otu\_table.txt”为某16S扩增子测序所得OTU丰度表格，其内容展示如下。

每一列为一个样本，每一行为一种OTU，交叉区域为每种OTU在各样本中的丰度。

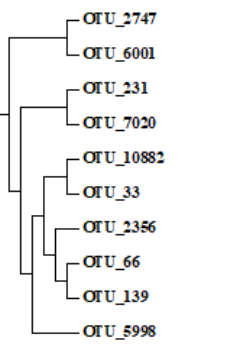


文件“otu\_tree.tre”为使用各OTU代表序列构建的进化树（即OTU水平的16S进化树）。

纯文本模式（如使用记事本）打开查看该进化树文件，内容长这样，即进化树文本文件的标准样式：



使用进化树可视化工具（例如MEGA）查看进化树，就是我们通常见到的这种了：



接下来使用R，基于上述两个文件，计算微生物群落的Alpha多样性（主要是丰度表，每一种Alpha多样性指数都会用到它；进化树仅用在谱系多样性指数的计算中）。

## R语言计算多种常见的Alpha多样性指数

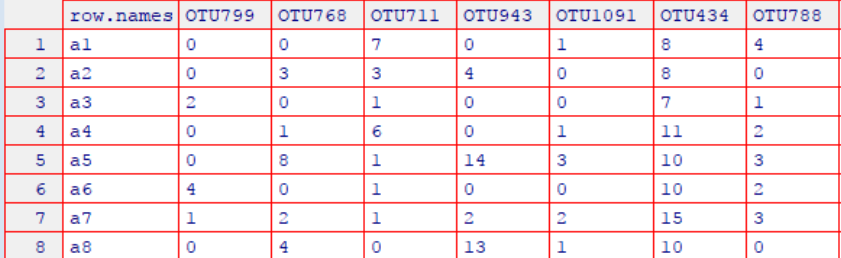
首先读入OTU丰度表。

#读入物种数据

otu <- read.delim('otu\_table.txt', row.names = 1, sep = '\t', stringsAsFactors = FALSE, check.names = FALSE)

otu <- t(otu)

读取后作个转置，使得每一行为一个样本，每一列为一种OTU。当然如果OTU丰度表在一开始就已经转置过了，读取后就可以直接使用了。



vegan包是生态群落分析的常用包，该包提供了计算多种Alpha多样性指数的命令，加载vegan包使用。常用的仅基于物种丰度数据即可得到的多样性指数，例如物种丰富度（Richness）、Shannon指数、Simpson指数、均匀度、Chao1指数、ACE指数等，均可基于vegan包中的命令计算得到。

谱系多样性（即PD\_whole\_tree）无法基于vegan包计算出，因为该指数比较特殊，除了物种丰富度外还需给定进化树文件。寻找在R中计算谱系多样性的方法也是找了许久才找到，参考网址：<https://daijiang.name/en/2014/05/04/notes-func-phylo-book-1/>。链接中提到picante包中提供了计算谱系多样性的命令，加载使用。

library(vegan)

library(picante)

#注：picante 包加载时默认同时加载 vegan，因此可省略“library(vegan)”这一步。

**几种常见Alpha多样性指数的计算命令**

常用的Alpha多样性指数计算命令如下所述。

主要使用到了vegan包中的diversity()命令（主要用于计算Shannon和Simpson）、estimateR()命令（主要用于计算Chao1和ACE），以及picante包中的pd()命令（计算PD\_whole\_tree）。关于这几个命令的详情，可分别使用?diversity()、?estimateR()、?pd()等查看帮助。

关于下文出现的各Alpha多样性指数的概念，可[点击查看简介](http://blog.sciencenet.cn/blog-3406804-1179809.html)。

##物种丰富度 Richness 指数

richness <- rowSums(otu > 0)

#或

richness <- estimateR(otu)[1, ]

##Shannon（以下 Shannon 公式的对数底数均设使用 e，在 R 中即表示为 exp(1)）

#Shannon 指数

shannon\_index <- diversity(otu, index = 'shannon', base = exp(1))

#Shannon 多样性

shannon\_diversity <- exp(1)^shannon\_index

#Shannon 均匀度（Pielou 均匀度）

pielou <- shannon\_index / log(richness, exp(1))

##Simpson

#Gini-Simpson 指数（我们平时常用的 Simpson 指数即为 Gini-Simpson 指数）

gini\_simpson\_index <- diversity(otu, index = 'simpson')

#经典 Simpson 指数（使用频率比较低）

simpson\_index <- 1 - gini\_simpson\_index

#Invsimpson 指数（Gini-Simpson 的倒数）

invsimpson\_index <- 1 / gini\_simpson\_index

#或

invsimpson\_index <- diversity(otu, index = 'invsimpson')

#Simpson 多样性

simpson\_diversity <- 1 / (1 - gini\_simpson\_index)

#Simpson 均匀度（equitability 均匀度）

equitability <- 1 / (richness \* (1 - gini\_simpson\_index))

##Chao1 & ACE

#Chao1 指数

chao1 <- estimateR(otu)[2, ]

#ACE 指数

ace <- estimateR(otu)[4, ]

##goods\_coverage 指数

goods\_coverage <- 1 - rowSums(otu == 1) / rowSums(otu)

##谱系多样性（与上述相比，还需指定进化树文件）

tree <- read.tree('otu\_tree.tre')

pd\_whole\_tree <- pd(otu, tree, include.root = FALSE) #测试时发现有根树和无根树的PD\_whole\_tree计算结果是一样的，但是无根树的计算会更快

**构建组合函数一步计算多种Alpha多样性指数**

熟悉了计算各Alpha多样性指数的命令后，接下来我们考虑将它们整合在一起，方便一步得到多个Alpha多样性指数结果。

整合上述命令，定义一个统一的函数alpha()，包含丰富度、Shannon熵指数、Simpson指数（Gini-Simpson指数）、Pielou均匀度、Chao1指数、ACE指数、goods\_coverage、PD\_whole\_tree共8种常见的Alpha多样性指数。

#定义函数

library(picante) #picante 包加载时默认同时加载 vegan

alpha <- function(x, tree = NULL, base = exp(1)) {

est <- estimateR(x)

Richness <- est[1, ]

Chao1 <- est[2, ]

ACE <- est[4, ]

Shannon <- diversity(x, index = 'shannon', base = base)

Simpson <- diversity(x, index = 'simpson') #Gini-Simpson 指数

Pielou <- Shannon / log(Richness, base)

goods\_coverage <- 1 - rowSums(x == 1) / rowSums(x)

result <- data.frame(Richness, Shannon, Simpson, Pielou, Chao1, ACE, goods\_coverage)

if (!is.null(tree)) {

PD\_whole\_tree <- pd(x, tree, include.root = FALSE)[1]

names(PD\_whole\_tree) <- 'PD\_whole\_tree'

result <- cbind(result, PD\_whole\_tree)

}

result

}

现在我们直接使用定义好的命令alpha()，就可以一步得到多种Alpha多样性指数了。

#加载 OTU 丰度表和进化树文件

otu <- read.delim('otu\_table.txt', row.names = 1, sep = '\t', stringsAsFactors = FALSE, check.names = FALSE)

otu <- t(otu)

tree <- read.tree('otu\_tree.tre')

#不包含谱系多样性，无需指定进化树；Shannon 公式的 log 底数我们使用 2

alpha\_all <- alpha(otu, base = 2)

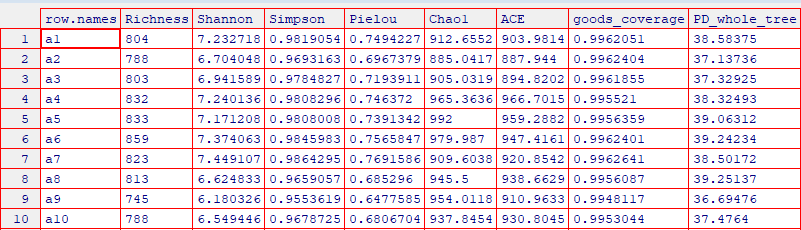
#包含谱系多样性时，指定进化树文件；Shannon 公式的 log 底数我们使用 2

alpha\_all <- alpha(otu, tree, base = 2)

#输出保存在本地

write.csv(alpha\_all, 'alpha.csv', quote = FALSE)

查看包含谱系多样性的统计结果“alpha\_all”，如下所示，对于每个样本，计算了多种Alpha多样性指数（一共定义了8种）供后续分析使用。



## 扩展，qiime程序计算Alpha多样性指数

若是对于16S/18S/ITS测序的微生物群落分析，qiime程序也是一个不错的选择。虽然个人感觉没有R方便，因为它需要将OTU表转化为biom文件，麻烦……

在linux命令行中调用qiime程序的“alpha\_diversity.py”来完成Alpha多样性指数的计算。

cp otu\_table.txt otu\_table.tsv

#在 otu\_table.tsv 开头添加一行“# Constructed from biom file”，以便将 otu\_table.tsv 转为 qiime 可识别样式

sed -i '1i\# Constructed from biom file' otu\_table.tsv

#otu\_table.tsv 转换为 otu\_table.biom

biom convert -i otu\_table.tsv -o otu\_table.biom --table-type="OTU table" --to-json

#alpha 多样性指数计算，包含 7 种指数

alpha\_diversity.py -i otu\_table.biom -o alpha.txt -t otu\_tree.tre -m observed\_otus,shannon,simpson,chao1,ace,goods\_coverage,PD\_whole\_tree

#如果不计算 PD\_whole\_tree 指数，就无需指定进化树文件 otu\_tree.tre

#得到的结果文件 alpha.txt 就是了

结果文件“alpha.txt”的内容如下，包含了我们计算的7种Alpha多样性指数。这里“observed\_species”就是丰富度richness指数。

