相似性分析(analysis of similarities, ANOSIM)，是一种用于分析高纬度数据组间相似性的非参数检验方法。它首先通过变量计算样本间距离(或者说相似性)，后计算关系排名，最后通过排名进行置换检验判断组间差异是否显著不同与组内差异。在生态统计中，可以使用ANOSIM分析 (往往同时配合排序分析来使用，对于ANOSIM，更常与NMDS排序分析放一起说明问题)，查看不同环境的群落组成结构差异是否显著，组间群落差异是否显著不同与组内差异。

这里以16S扩增子测序所得的细菌群落数据为例，展示R包vegan进行ANOSIM分析检验群落结构组成差异的一般过程。

OTU, 运算的分类单位operational taxonomic unit缩写OTU，指在数量分类方面作为对象的分类单位之总称，有种，变种，个体等。在用群体的时候，根据相似系数值和由任意标准去归纳整理有可能的，因为也有与历来分类单位的等级(rank)不一致的情况，所以使用了这个术语。

OTU，是指系统发生分析中的一个外部节点，是一个假定的分类单元；常见于动植物类系统发育分析。例如6 clone/2 OUTs是指要进行系统分析的个体，即每2个待分析的个体包含6个克隆。

在微生物的培养分析中经常用到，通过提取样本的总基因组DNA，利用16S rRNA或ITS的通用引物进行PCR扩增，通过测序以后就可以分析样品中的微生物多样性，那么如何区分这些不同的序列呢，这个时候就需要引入operational taxonomic units，一般情况下，如果序列之间，比如不同的16S rRNA序列的相似性大于98%就可以把它定义为一个OTU，每个OTU对应一个不同的16S rRNA序列，也就是每个OUT对应于一个不同的细菌(微生物)种。

通过OTU分析，就可以知道样本中的微生物多样性和不同微生物的丰度。同时通过OTU的稀有度分析，还可以看出你的测序量是否足够反应样品中的大部分微生物种，如果稀有度曲线趋于平稳就说明你的分析结果已经包含了样品中大部分微生物种，如果稀有度还在向上，就说明你的测序量不够，没有包括样品中的大部分微生物种。

`otu_table_anosim.txt`为OTU丰度表格，内容如下，每一列尾一个样本，每一行为一种OTU，交叉区域为每种OTU在各样中的丰度：

	#OTU_num	s1.1	s1.2	s1.3	s1.4	s1.5	s1.6	s1.7	s1.8	s2.1	s2.2
2	OTU_2	0.609009775	0.629515512		0.70558861		0.673820654		0.682320442		0.57
3	OTU_3	0.039842754	0.055779856		0.026030599		0.035274118		0.028155546		0.04
4	OTU_4	0.037292818	0.122821929		0.022949426		0.047705057		0.0219932		0.03
5	OTU_5	0.002549936	0.001487463		0.001806205		0.001168721		0.001699958		0.00
6	OTU_7	0.001168721	0.000637484		0.001274968		0.001381215		0.00159371		0.00
7	OTU_8	0.000106247	0.000212495		0.000318742		0.000318742		0.000106247		0
8	OTU_9	0.000637484	0.000212495		0.000424989		0.000531237		0		0.000318742

`bary_anosim.txt`为提前计算得到的样本距离矩阵文件(这里展示的是样本间的Bary-curtis距离)，其内容展示如下，每一列为一个样本，每一行为一个样本，交叉区域为样本间的Bary-curtis距离(取值范围为0-1,越接近于1表明样本间细菌组成差异越大)：

		s1.1	s1.2	s1.3	s1.4	s1.5	s1.6	s1.7	s1.8	s2.1	s2.2	s2.3
2	s1.1	0	0.165108358		0.175201854		0.138865264		0.164045883		0.156714815	
3	s1.2	0.165108358	0		0.192626413		0.151508695		0.18657032		0.210050981	
4	s1.3	0.175201854	0.192626413		0		0.113897135		0.092010177		0.193476389	
5	s1.4	0.138865264	0.151508695		0.113897135		0		0.10752229		0.178814265	
6	s1.5	0.164045883	0.18657032		0.092010177		0.10752229		0		0.167020799	
7	s1.6	0.156714815	0.210050981		0.193476389		0.178814265		0.167020799		0	

`group_anosim.txt`为样本分组信息，其内容如下，第一列(names)为各样本名称，第二列为各样本的分组信息，即这些样本所属的采样地点(s1, 地点1；s2，地点2；s3...以此类推，names列中的样本名称的顺序一定要和OTU表或者距离矩阵中的样本名称顺序一致)：

```
names   site
s1.1    s1
s1.2    s1
s1.3    s1
s1.4    s1
s1.5    s1
s1.6    s1
```

使用vegan包进行ANOSIM分析检验群落结构差异

可导入已经计算好的样本距离矩阵文件，也可以使用OTU丰度表格文件，同时导入样本分组文件。当读入数据为距离矩阵时，需要将读取的数据框转化为dist类型，便于后续的函数识别；当读入本示例的OTU表格时，则首先要进行转置操作，即要求所得数据框格式：每一行为一个样本，每一个列为物种信息。

读入文件，现有距离矩阵

```
dis <- read.delim("bray_anosim.txt",row.names=1,sep="\t",stringsAsFactors=F,check.names=F)
```

```
dis <- as.dist(dis)
```

或者直接使用OTU丰度表

```
otu <- read.delim("otu_table_anosim.txt",row.names=1,sep="\t",stringsAsFactors=F,check.names=F)
```

```
otu <- data.frame(t(otu))
```

样本分组文件

```
group <- read.delim("group_anosim.txt",sep="\t",stringsAsFactors=F)
```

整体水平(简要展示基本流程)

首先在所有分组水平上，使用ANOSIM检验整体差异，即查看来自5个不同采样地点所获得的土壤细菌群落结构组成在整体上是否不具备一致性，各组间群落差异是否显著不同于各组内的差异。

导入vegan包，并使用vegan包中的anosim()函数执行ANOSIM分析。

```
library(vegan)
```

ANOSIM分析(所有分组比较，即整体差异)

1. 若是已经提供好了距离矩阵，则直接使用现有的距离矩阵进行分析即可，根据group$site这里列样本分组信息进行ANOSIM分析，随机置换检验999次

```
anosim_result_dis <- anosim(dis, group$site, permutation=999)
```

Analysis of similarities (ANOSIM) provides a way to test statistically whether there is a significant difference between two or more groups of sampling units. Function `anosim` operates directly on a dissimilarity matrix. A suitable dissimilarity matrix is produced by functions `dist` or `vegdist`. The method is philosophically allied with NMDS ordination (`monoMDS`), in that it uses only the rank order of dissimilarity values.

2. 若是使用OTU丰度表，则需要计算时指定所依据的距离类型，这里依然使用Bray-Curtis距离

```
anosim_result_otu <- anosim(otu, group$site, permutation=999, distance='bray')
```

3. 或者首先根据丰度计算样本距离，再将所计算距离数据作为输入

```
dis1 <- vegdist(otu, method="bray")
```

```
anosim_result_dis1 <- anosim(dis1, group$site,permutation=999)
```

使用anosim()命令进行ANOSIM分析时，指定样本间距离数据("dis")，以及各样本所属的分组信息("group"数据框的"site"列，为了避免识别错误，不要使用纯数字作为分组名称)，且要注意分组信息中的分组名称与距离矩阵中样本名称要按顺序对应(需预先排列好二者的顺序)，permutations参数用于指定随机置换检验的次数，这里设这为999次(视数据量而定，数据量越大，置换次数也应当增大)

查看结果，上述3种方法计算得到的内容一致：

```
summary(anosim_result_dis)
```

```
names(anosim_result_dis)
```

```
> summary(anosim_result_dis)

Call:
anosim(x = dis, grouping = group$site, permutations = 999)
Dissimilarity:

ANOSIM statistic R: 0.3483
      Significance: 0.001

Permutation: free
Number of permutations: 999

Upper quantiles of permutations (null model):
   90%    95%  97.5%    99%
0.0526 0.0742 0.0939 0.1148

Dissimilarity ranks between and within classes:
        0%    25%   50%    75% 100%   N
Between  8 231.75 415.5 607.25  780 640
s1       1  17.75 166.5 227.00  360  28
s2       2  60.50 145.5 257.25  578  28
s3      17 100.75 238.5 438.25  761  28
s4      26 290.75 462.0 595.50  725  28
s5       4 123.00 437.5 544.25  680  28
```

这里，可主要关注两国重要统计值，R值(ANOSIM statistic R)和p值(Significance)

R值可以得出组间与组内比较的差异程度，其取值范围(-1,1)；R>0，说明组间差异显著，R<0，说明组内差异大于组间差异，R值的绝对值越大表明相对差异越大。p值越低表明这种差异检验结果越显著，一般以0.05为显著性水平界限

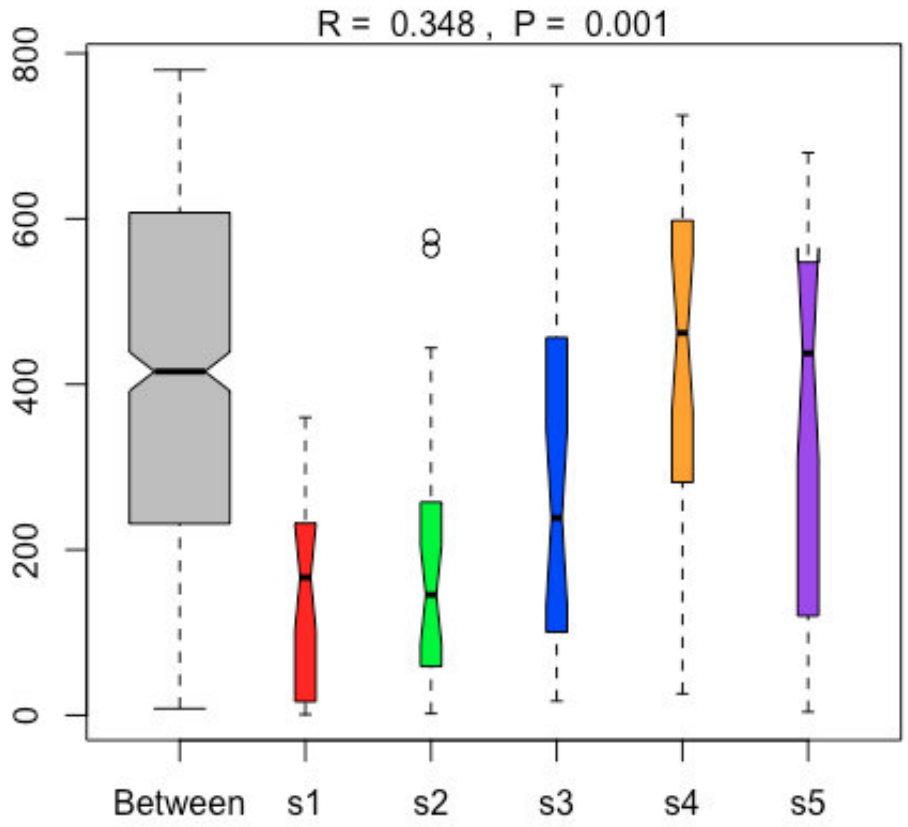
```
> names(anosim_result_dis)
[1] "call"         "signif"       "perm"         "permutations" "statistic"    "class.vec"
[7] "dis.rank"     "control"
> anosim_result_dis$statistic
[1] 0.3483259
> anosim_result_dis$signif
[1] 0.001
```

可使用plot()命令简单展示统计结果

```
pdf(paste("anosim.all.pdf",sep=""),width=10,height=5)
```

```
plot(anosim_result_dis,col=c("gray","red","green","blue","orange","purple"))
```

```
dev.off()
```



ANOSIM分析基于两两样本之间的距离值排序获得的秩，这样任一两两组的比较可以获得三个分类的数据。以箱线图的形式展示组间与组内的秩的分布，横坐标表示所有样品(between)以及各分组(本例为s1,s2,s3,s4,s5)，纵坐标表示距离(本示例使用Bray-Curtis距离)的秩。当between组相对与其他每个分组的秩较高时，则表明组间差异大于组内差异。

两组间比较(展示一个批量处理循环流程)

我们已知在整体水平上，组间群落组成是存在差异的。那么，究竟是哪些之间的差异所导致的呢？在这些物种组成具有显著性的不同群落中，哪些群落之间的差异较小，而哪些群落之间具有更大的差异呢？此外，有时我们也并不关注整体水平，而更多是想关注某几个特定分组之间的群落差异是否显著，这时候也需要我们将这些特定分组的样本挑选出，单独进行表。

现在使用ANOSIM分析探索两两环境(两两采样地)之间的土壤细菌群落结构组成的差异，包括是否具有显著性，以及组间的变异程度等。

推荐使用OTU丰度作为输入数据，每次筛选分组后重新计算样本距离，而不是读入现有的距离矩阵后，在矩阵中筛选样本行列。这样可避免由于样本数减少可能导致的距离变动而造成的误差(有时候因样本数的减少，导致了实际存在的物种数也相应地减少，这样重新计算的样本间距离与之前的相比，可能会有不同)

ANOSIM分析，使用循环处理，进行小分组比较，如两组间

推荐使用OTU丰度表作为输入数据，每次筛选分组后重新计算样本距离，避免由于样本数减少可能导致的距离变动而造成误差

```
group_name <- unique(group$site)
dir.create("anosim_two",recursive=T)
anosim_result_two <- NULL

for( i in 1:(length(group_name)-1)){
  for(j in (i+1):length(group_name)){
    group_ij <- subset(group,site %in% c(group_name[i],group_name[j]))
    otu_ij <- otu[group_ij$names,]
    anosim_result_otu_ij <- anosim(otu_ij,group_ij$site,permutations = 999,distance = "bray")

    if(anosim_result_otu_ij$signif <= 0.001)Sig <- "***"
    else if(anosim_result_otu_ij$signif <= 0.01)Sig <- "**"
    else if(anosim_result_otu_ij$signif <= 0.05)Sig <- "*"
    else Sig <- NA

    anosim_result_two <- rbind(anosim_result_two,
                               c(paste(group_name[i],group_name[j],sep="/"),
                                 "Bray-Curtis",anosim_result_otu_ij$statistic,
                                 anosim_result_otu_ij$signif,Sig))
    pdf(paste("anosim_two/anosim.",group_name[i],"_",group_name[j],".pdf",sep=""),
        width=7,height=5)
    plot(anosim_result_otu_ij,col=c("gray","red","blue"))
    dev.off()
  }
}

anosim_result_two <- data.frame(anosim_result_two,stringsAsFactors = F)
names(anosim_result_two) <- c("group","distance","R","P_value","Sig")
```

这里，在每一步循环中挑选出两个分组，并进行两两组间的ANOSIM分析。读入数据OTU表，筛选出对应的样本数据并作为输入，距离类型使用Bray-Curits距离，999次置换检验。

补充说明

在进行ANOSIM分析检验群体结构差异的同时，常结合排序分析(如PCA,PCoA, NMDS等，对与ANOSIM更推荐配合NDMS)的可视化展示，一同说明问题，使结果更具说服力。

通过排序图，观察群落样本的坐标分布，可判断不同分组之间是否具有明显区分(组间差异)，同一分组样本之间的离散程度(组内差异)，哪些分组在排序图中所处的区域较为一致(组间相似)，哪些分组与其他分组相比在图中单独位于一个较远的区域(差异较大的分组)，哪些样本明显偏离了平均位置(离群点)，各排序轴主要与那些环境因素存在潜在的相关性(群落差异主要受到哪些环境的影响)及各排序轴的解释量高低(模型解释度)等。

再结合ANOSIM的检验结果，验证以上观测，作总结。