GO term counting

hguo

October 3, 2018

```
# read the network. Yeast PIN is used here
network <- read.csv("../Data/yeast.pin.csv", header=T, stringsAsFactors=F)
geneA <- network$geneA
geneB <- network$geneB

## the GO terms for biological processes (BP)
## In this slim, each GO terms was shared by at least 20 genes
GOcategory.file <- read.csv("../Data/yeast.bp.term.csv",header=TRUE, stringsAsFactors=F)
go.cat <- GOcategory.file$GO.term

# Gene-GO term file
# Note in yeast PIN we use the systematic namds of all genes
GOterm.file <- read.csv("../Data/yeast.bp.gene.term.csv", header=T, stringsAsFactors=F)
GO.gene <- GOterm.file$System
GO.term <- GOterm.file$GO.term

# Define matrices
dim <- length(go.cat)
A <- matrix(0, nrow=dim, ncol=dim)
B <- matrix(0, nrow=dim, ncol=dim)

for (i in 1:length(geneA)) {
  goA <- GO.term[which(GO.gene %in% geneA[i])]
  for (j in 1:length(goA)) {
      m <- which(go.cat %in% goA[j])
      goB <- GO.term[which(GO.gene %in% geneB[i])]
      for (k in 1:length(goB)) {
          n <- which(go.cat %in% goB[k])
          A[m,n] <- A[m,n] + 1
          A[n,m] <- A[n,m] + 1
      }
   }
}

## the diagonal terms are double counted need be divided by 2
for (m in 1:dim) {
    for (n in 1:dim){
        if (m == n) {
           B[m,n] = A[m,n]/2
        }  else {
           B[m,n] = A[m,n]
        }
    }
}

B <- log10(B+1)
```

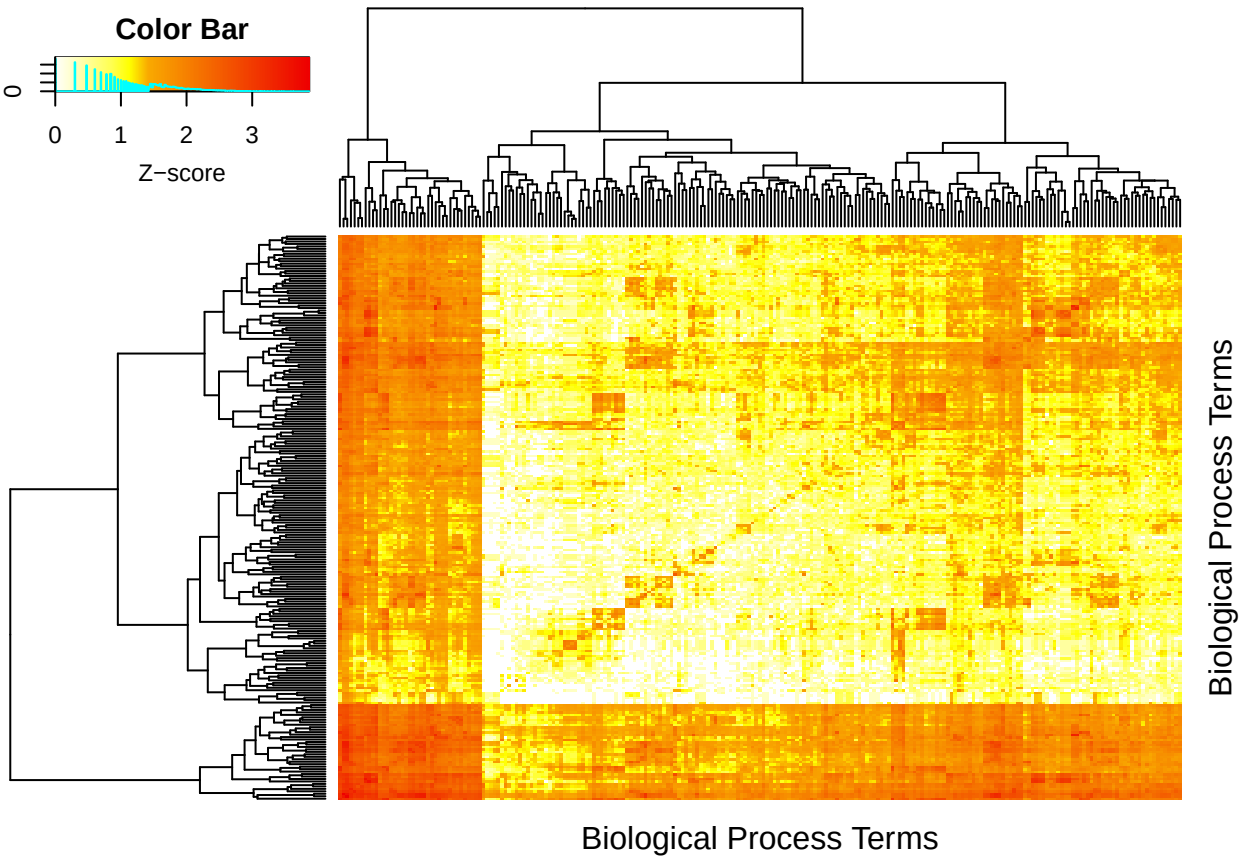
```
write.table(B, file="yeast.bp.matrix.csv",sep=",",row.names=F,col.names=F, quote=F)

# the final output will be a n*n matrix, where n is the total number of
# GO terms (229 in the yeast BP)
```

```
library('gplots')
```

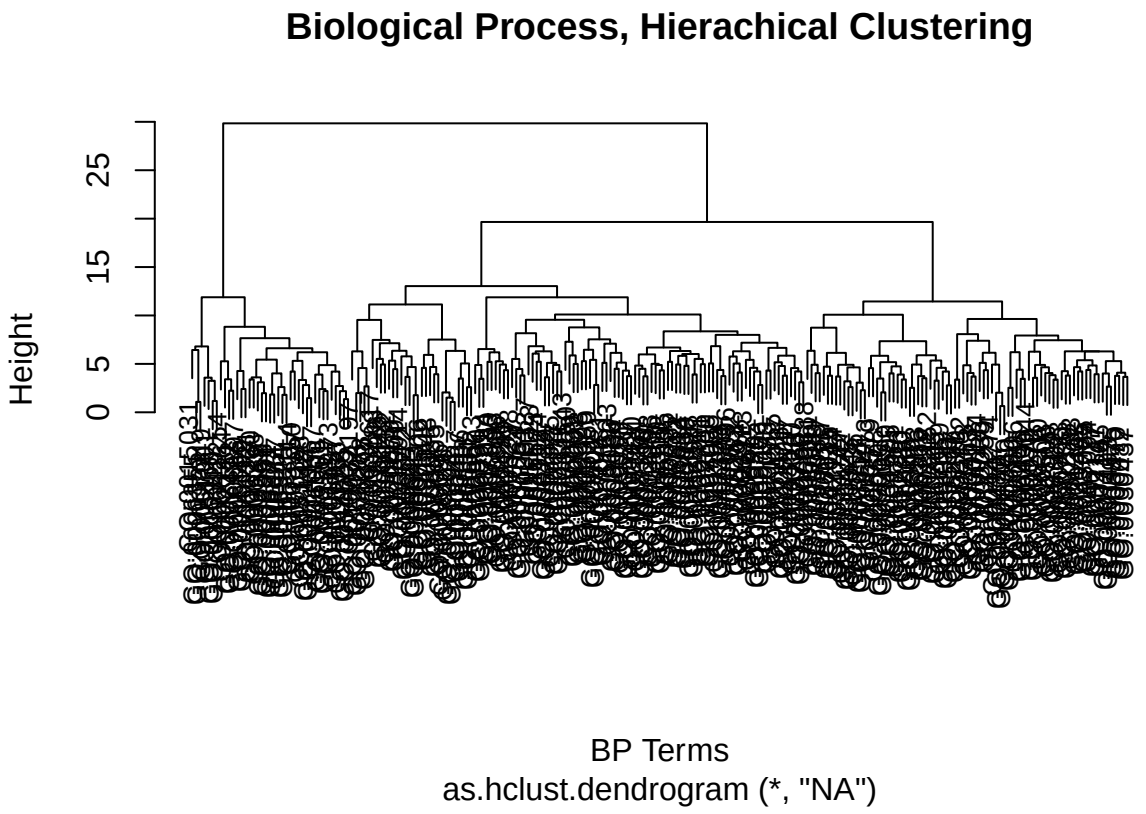
```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
# plotting the heatmap of the matrix
# define the col/row for the matrix B (see above)
colnames(B) <- go.cat
rownames(B) <- go.cat

# using quantiles for color scheme
quantile.B <- quantile(B, probs=seq(0,1,1/3))
colors = c(seq(0,(quantile.B[2]-0.01), length=100),
           seq((quantile.B[2]+0.01),(quantile.B[3]-0.01), length=100),
           seq((quantile.B[3]+0.01), max(B), length=100))
my_pallete <- colorRampPalette(c("white", "yellow", "red2"))(n=299)

#png(filename = "yeast.bp-bp.png",width=6, height=6, res=1200, unit="in")
hm <- heatmap.2(B, col=my_pallete, breaks=colors,
          trace='none', offsetRow=0, offsetCol=0,
          ylab="Biological Process Terms", xlab="Biological Process Terms",
          margins = c(2,2), key.title="Color Bar", key.xlab="Z-score", key.ylab=NA,
          labCol=NA, labRow=NA,
          scale="none", dendrogram="both", symbreaks=T, symm=F, symkey=F)
```




```
#dev.off()

hc <- as.hclust(hm$colDendrogram)

#pdf(file="yeast.bp-bp.hierachical.clustering.pdf, width=10, height=5, paper='special")
plot(hc, xlab="BP Terms", main="Biological Process, Hierachical Clustering", cex=.8)
```


Biological Process, Hierachical Clustering
Height
0
5
15
25
BP Terms
as.hclust.dendrogram (*, "NA")

```
#dev.off()
```