made4 Multivariate analysis of microarray data

John Barlow

April 23, 2018

references:

I used the following resources to learn about using made4

1. Bioconductor made4

https://www.bioconductor.org/packages/release/bioc/html/made4.html (https://www.bioconductor.org/packages/release/bioc/html/made4.html)

http://www.bioconductor.org/packages//2.7/bioc/vignettes/made4/inst/doc/introduction.pdf (http://www.bioconductor.org/packages//2.7/bioc/vignettes/made4/inst/doc/introduction.pdf)

2. Culhane AC, Thioulouse J, Perriere G, Higgins DG.(2005) MADE4: an R package for multivariate analysis of gene expression data. Bioinformatics 21(11):2789-90. note the introduction and reference manual can be opened from within R

“made4 is useful for Multivariate data analysis and graphical display of microarray data. Functions include between group analysis and coinertia analysis. It contains functions that require ade4 package.”

Korin found this package because we are particularly interested in the coinertia analysis function to explore potential trends between gene abundancy data sets.

before the preliminaries

recent versions of made4 and associated packages were built under R version 3.4.4, so like me you might need to upgrade from an older R version or else when you download made4 it will throw a warning message - I don’t know if updating was absolutely necessary, but I did not chance it as my last R update was about 6 months old

I found a package to make the process of upgrading R a bit easier - the installr package - it seemed to work well and included an option to upgrade packages when it ran

you need to run installr in the R Gui console, not from R studio here is a link on updating R with installr

ttp://bioinfo.umassmed.edu/bootstrappers/bootstrappers-courses/courses/rCourse/Additional_Resources/Updating_R.html

(http://bioinfo.umassmed.edu/bootstrappers/bootstrappers-courses/courses/rCourse/Additional_Resources/Updating_R.html)

you can install the installr package from RStudio

```
install.packages("installr")
```

then close RStudio and open RGui in RGui open the installr library and run updateR

```
library(installr)
updateR()
```

the preliminaries

download made4 package - made4 package uses ade4 package, scatterplot3d package, RColorBrewer and gplots

once your R version is up to date, then on to the preliiminaries of installing made4 and associated packages - ade4 and scatterplot3d appear to install automatically when you install made4

```
install.packages("made4")
```

Korin suggests one needs to install made4 as below from the biocLite not via install packages, but install packages seemed to work for me (although it loaded an older version and to get the most recent version I had to download direct from the made4 bioconductor site

```
source("https://bioconductor.org/biocLite.R")
biocLite("made4")
biocLite("BiocUpgrade")
```

confirm that the 3 downloaded files are in your current working directory

```
getwd()
#move files into current wd or if needed set wd
setwd("insert address for your particular wd")
```

load made4 - associated packages ade4, RColorBrewer, gplots, and scatterplot3d load automatically when you load made4

```
library(made4)
```

Data structure constraints for made4:

1. with some functions for testing relationships between 2 data sets is that the datasets must contain the same samples ("either the rows or the columns of a matrix must be matchable").

read in the 3 data files

```
res_mat <- read.csv("res_mat_abun.csv")
bac_mat <- read.csv("bac_mat_abun.csv")
vf_mat <- read.csv("vf_mat_abun.csv")
```

2. made4 needs the abundance values to be numeric, and for some reason a number of values are as integers in our csv files - so code we figured out to correct this

check the structure of the data files

```
str(res_mat)
#so clean thus up by setting all values to numeric, except not the gene names
res_mat_all <- res_mat[,-c(1)]
res_mat_all[] <- lapply(res_mat_all[], as.numeric)
rownames(res_mat_all) <- res_mat$Name
head(res_mat_all)
str(res_mat_all)
```

3. also, a number of functions throw an error if any sites have all 0 values for the abundance of observed genes e.g. the HOSP, WOCA, and SWCA for ARGs all have zero abundance for ARGs (see Figure 3 ARG freq by site.html (fig3_arg_freq_by_site.html)).

here is the code Korin figured out to resolve this issue

```
none <- lapply(res_mat_all, function(x) all(x == 0))
which(none == "TRUE")
# remove those 6 columns
res_mat2 <- res_mat_all[,-c(4,10,11,15,21,24)]
head(res_mat_all)
str(res_mat_all)
```

repeat for the other 2 files, removing the same 6 columns from each file

```
bac_mat_all <- bac_mat[,-c(1)]
bac_mat_all[] <- lapply(bac_mat_all[], as.numeric)
rownames(bac_mat_all) <- bac_mat$Name
bac_mat2 <- bac_mat_all[,-c(4,10,11,15,21,24)]
head(bac_mat2)
str(bac_mat2)

vf_mat_all <- vf_mat[,-c(1)]
vf_mat_all[] <- lapply(vf_mat_all[], as.numeric)
rownames(vf_mat_all) <- vf_mat$Name
vf_mat2 <- vf_mat_all[,-c(4,10,11,15,21,24)]
head(vf_mat2)
str(vf_mat2)
```

data files now ready to use

made4 functions

made4 is well documented - to view documentation

```
browseVignettes("made4")
```

the package also contains a data set of gene expression data

made4 has an overview function which generates a boxplot, histogram and hierachial tree of the data

```
overview(res_mat2)
```

the ord function makes it east to run ordination methods to explore structure of a data matrix - the methods include correspondence (coa, the default), non-symmetric correspondence analysis (nsc) or principal component (pca).

```
res_coa<-ord(res_mat2, type= "coa")

summary(res_coa$ord)
```

and plot the results

```
plot(res_coa)

plotgenes(res_coa)
plotarrays(res_coa)
#including generating in 3d
do3d(res_coa$ord$li)
do3d(res_coa$ord$co)
```

the support documents provide package and demo code for visualizing the 3d plots dynamically (can be rotated).