Date: November 24, 2015
Course: GBIO0009

HW 3b – Co-expression Networks

Coexpression networks are natural concept to biologists as many genes also linked via biological pathways. In this homework we will build co-expression networks using microarray expression data. Then we will extract modules (sub-networks) and analyze them both biologically and in relation to a trait. The homework aim is to solidify the concepts learned on Weighted Gene Coexpression Network Analysis (WGCNA) introduced during practical session.

Due date: Dec 8th, 2015

Introduction

In this homework we will use mice microarray expression data. The source tissue is mouse liver. The dataset contains total of 3600 gene expression profiles. These were filtered from the original over 20,000 genes by keeping only the most variant and most connected probes (i.e. genes).

```
library(WGCNA)
library(cluster)
options(stringsAsFactors = FALSE)

#switch to the directory containing req files

setwd("") #put your path here (if necessary)

#Read in the female liver data set
femData = read.csv("LiverFemale3600.csv")

#print column names
colnames(femData)
```

How many mice do we have in this dataset? Hint, not all columns refer to mice IDs

```
#Remove gene information and transpose the expression data
datExprFemale=as.data.frame(t(femData[, -c(1:8)]))
names(datExprFemale)=femData$substanceBXH
rownames(datExprFemale)=names(femData)[-c(1:8)]
```

In datExprFemale data frame what rows and columns represent?

```
# Read in the probe annotation
GeneAnnotation=read.csv(file="GeneAnnotation.csv")
```

The probe name MMT00073501 comes from which microarray platform (e.g. Affymetrix, Illumina, etc)? What is the average length of its probes? To which gene symbol and gene name it maps to?

```
# Now we read in the physiological trait data
traitData = read.csv("ClinicalTraits.csv")
dim(traitData)
names(traitData)
# use only a subset of the columns  (i.e. traits)
```

```
allTraits=traitData[,c(2, 11:15, 17:30, 32:38)]
names(allTraits)
dim(allTraits)
```

The allTraits data frame contains physiological and clinical traits data. For how many mice do we have clinical data? How many traits do we have in this data frame? What does the 1st column of this dataframe represents?

```
# Order the rows of allTraits so that
# they match those of datExprFemale:
Mice=rownames(datExprFemale)
matchedRows = match(Mice, allTraits$Mice)
datTraits = allTraits[matchedRows,]
```

After matching and ordering of the allTraits and datExprFemale data frames based on Mice IDs, what is are the new dimensions of the new datTraits data frame

```
# Choose a set of soft thresholding powers
powers=c(1:10) # in practice this should include powers up to 20.
# choose power based on SFT criterion
sft=pickSoftThreshold(datExprFemale,powerVector=powers)
# SFT index as a function of different powers
plot(sft$fitIndices[,1],-sign(sft$fitIndices[,3])*sft$fitIndices[,2],
      xlab="Soft Threshold(power)",ylab="SFT, signed R^2",type="p",
      main=paste("Scale independence"))
text(sft$fitIndices[,1],-sign(sft$fitIndices[,3])*sft$fitIndices[,2],
      labels=powers,col="red")
# this line corresponds to using an R^2 cut-off of h
abline(h=0.90,col="red")
```

Which power () would you select for further analysis? What "SFT.R.sq" value means? What scale free network topology means? Paste the plot you obtained.

How well at the selected power the resulting co-expression network follows scale free topology? Provide the R^2 value. (fill in highlighted are by missing code)

```
#use the power selected from previous step
k=softConnectivity(datE=datExprFemale, power= …)
scaleFreePlot(k, main="Check scale free topology\n")


# We now calculate the weighted adjacency matrix
A = adjacency(datExprFemale, power = 7)
#Alternatively use this function, the result is identical
#A= abs(cor(datExprFemale, use="p")^7)

# Turn adjacency into a measure of dissimilarity
distA=1-A
hierA=hclust(as.dist(distA), method="average" )
# Plot the resulting clustering tree together with the true color
assignment
sizeGrWindow(10,5);
```

```
colorStaticA=as.character(cutreeStaticColor(hierA, cutHeight=.99,
minSize=20))
#dynamic tree cutting
branch.number=cutreeDynamic(hierA,method="tree");
colorDynamicA=labels2colors(branch.number)

plotDendroAndColors(hierA, colors=data.frame(colorStaticA,
colorDynamicA), dendroLabels = FALSE, hang = 0.03,
     main = "Gene hierarchical clustering dendrogram" )
```

Now you should see the hierarchical clustering dendrogram. Paste it below. What tree brunches represent? Which tree cutting method produces more clusters? How many clusters you obtain in case of static and dynamic tree cutting algorithms?
Remember color grey means that these genes do not belong to any cluster
Hint: look at the *colorStaticA* and *colorDynamicA* character vectors and count occurrence of unique colors (i.e. modules). I suggest convert these vectors to factors and then run summary function

```
# Calculate eigengenes
MEList=moduleEigengenes(datExprFemale, colors=colorStaticA)
signif(cor(MEList$eigengenes, use="p"), 2)

MEs = MEList$eigengenes
# Add the weight to existing module eigengenes
#this is the body weight
weight = as.data.frame(datTraits$weight_g)
names(weight)="weight"
MET=orderMEs(cbind(MEs,weight))
# Plot the relationships among the eigengenes and the trait
plotEigengeneNetworks(MET,"",marDendro=c(0,4,1,2),
marHeatmap=c(3,4,1,2),cex.lab=0.8,xLabelsAngle=90)

results <- signif(cor(weight,MEList$eigengenes, use="p"),2)
results[,order(results,decreasing=T)]
```

Now we calculated module eigengenes and had clustered them via the dendrogram. What is the connection between eigengene and module (see notes and [1]). Paste the obtained image. The lower part of the plot shows ordered heatmap displaying correlations between modules. Looking from the top of the dendrogram (i.e. root, level 0), how many main brunches do you see? Weight trait is closest to which modules (name four module colors)? Why do we see the diagonal line made of squares in the heat plot? Do you see a red rectangle around "weight". Do the module colors correspond to the dendrogram above layout (yes or no)?

```
results <- signif(cor(weight,MEList$eigengenes, use="p"),2)
results[,order(results,decreasing=T)]
```

List Pearson correlation values of the four modules that are most similar (associated) to the weight (i.e. our selected trait). Do these results match to previously obtained dendrogram and heat map? What does negative correlation means in terms of similarity (i.e. does this mean high module similarity or no)? Look at the dendrogram to see what negative correlation means

```
# Choose a module assignment
moduleColorsFemale=colorStaticA
# Define numbers of genes and samples
nGenes = ncol(datExprFemale)
nSamples = nrow(datExprFemale)
# Recalculate MEs with color labels
MEs0 = moduleEigengenes(datExprFemale,moduleColorsFemale)$eigengenes
MEsFemale = orderMEs(MEs0)
modTraitCor = cor(MEsFemale, datTraits[-1], use = "p")
modTraitP = corPvalueStudent(modTraitCor, nSamples)
#Since we have a moderately large number of modules and traits,
#a suitable graphical representation will help in reading
#the table. We color code each association by the correlation value:
# Will display correlations and their p-values
textMatrix = paste(signif(modTraitCor, 2), "\n(",
signif(modTraitP, 1), ")", sep = "")
dim(textMatrix) = dim(modTraitCor)
par(mar = c(5, 12, 1, 1))
# Display the correlation values within a heatmap plot
labeledHeatmap(Matrix = modTraitCor, xLabels = names(datTraits[-1]),
yLabels = names(MEsFemale), ySymbols = names(MEsFemale),
colorLabels =FALSE,colors=greenWhiteRed(50),textMatrix=textMatrix,
setStdMargins = FALSE, cex.text = 0.6, zlim = c(-2,2),
main = paste("Module-trait relationships"))
```

Now that we tested all other traits with respect to modules, which traits have good correlation with obtained modules? Name a few? Which module has highest correlation to the traits?

```
#if not KEGGprofile is not installed run:
#source("http://bioconductor.org/biocLite.R")
#biocLite("KEGGprofile")

library(KEGGprofile)  #if not installed use
library(biomaRt)
ensembl=useMart("ensembl")
listDatasets(ensembl)
mart<- useDataset("mmusculus_gene_ensembl", useMart("ensembl"))
listFilters(mart)


module_genes=femData[colorStaticA=="black","gene_symbol"]

enterezIDs = getBM(attributes="entrezgene", filters = "wikigene_name", values =
module_genes, mart = mart)
result <- find_enriched_pathway(enterezIDs, returned_pvalue = 0.01, returned_genenumber
= 1, specis="mmu")
result[[1]][, c(1, 5)]
```

Which top 10 pathways are significantly enriched? Do they biologically link to the traits that were identified in previous question (just think a bit about biology and give your opinion? Look for articles in PubMed or GoogleScholar that link your trait and enriched pathway (there are several)?

Suggested references

[1] Langfelder, Peter, and Steve Horvath. "Eigengene networks for studying the relationships between co-expression modules." *BMC systems biology* 1.1 (2007): 54.