— Differential Expression and Gene Ontology —

Course in Practical Analysis of Microarray Data
Computational Exercises
May 2005, Manuela Hummel and Ulrich Mansmann

1.) Preliminaries.

a. To go through this exercise you need to have the libraries GOstats and hgu133a. For producing the graphs you also need the package Rgraphviz which only works on Unix.

```
> library(GOstats)
> library(hgu133a)
> library(Rgraphviz)
```

b. Load the `.RData`–file which contains the RMA–normalized expression set `data.rma` and the mean raw expression values in `mean.expression.res`.

```
> loadURL("http://compdiag.molgen.mpg.de/ngfn/data/2005/may/data.rma.RData")
```

2.) Differential expression.

a. Test for differential expression with a p–value adjustment by Benjamini–Yekutieli.

```
> cl <- pData(data.rma)[,"group"]
> t <- mt.teststat(exprs(data.rma), classlabel=cl)
> p <- 2 * pt(-abs(t), df = ncol(exprs(data.rma)) - 2)
> p.adj <- mt.rawp2adjp(p, proc = c("BY"))
```

b. Create an expression set with genes that have an adjusted p–value < 0.05 and a fold change of at least 2.

```
> ord <- order(p.adj$index)
> adj <- p.adj$adj[ord, "BY"]
> FC <- mean.expression.res[,3] / mean.expression.res[,4]
> eset <- data.rma[adj < 0.05 & FC > 2, ]
```

3.) Gene Ontology.

a. To explore whether the selected genes have a common molecular function we use the Gene Ontology.

```
> gn <- geneNames(eset)
> LLID <- unlist(mget(gn, hgu133aLOCUSID, ifnotfound=NA))
> go <- makeGOGraph(as.character(LLID), "MF", removeRoot=FALSE)
> go
```

b. Plot the GO graph (only Unix).

```
> att <- list()
> lab <- rep("", length(nodes(go)))
> names(lab) <- nodes(go)
> att$label <- lab
> plot(go, nodeAttrs=att)
```

c. We can now ask if there are more interesting genes at one node that one might expect by chance.

```
> hyp <- GOHyperG(unique(LLID), lib="hgu133a", what="MF")
> names(hyp)
```

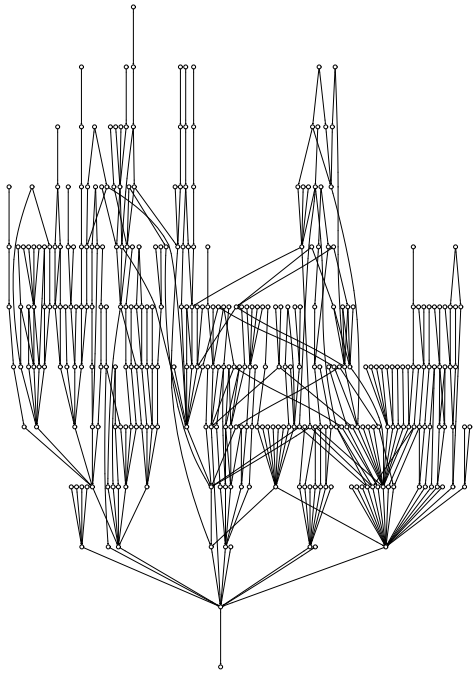

Figure 1: see exercise 3.b.

```
> go.pv <- hyp$pvalues[nodes(go)]
> go.pv <- sort(go.pv)
```

d. Create a table of those GO terms that have a p–value < 0.01. Show the GO IDs, GO terms (some terms are quite long and will therefore be abbreviated), p–values and the number of genes that are annotated at each term.

```
> sig <- go.pv[go.pv < 0.01]
> counts <- hyp$goCounts[names(sig)]
> terms <- getGOTerm(names(sig))[["MF"]]
> nch <- nchar(unlist(terms))
> terms2 <- substr(unlist(terms), 1, 25)
> terms3 <- paste(terms2, ifelse(nch > 25, "...", ""), sep="")
> matrix(c(names(terms), terms3, round(sig, 3), counts),
+ ncol=4, dimnames=list(1:length(sig),
+ c("GO ID", "Term", "p-value", "# Genes")))
```

```
   GO ID        Term                              p-value # Genes
1  "GO:0004871" "signal transducer activit..." "0"     "1871"
2  "GO:0004872" "receptor activity"            "0"     "1122"
3  "GO:0030246" "carbohydrate binding"         "0"     "185"
4  "GO:0008009" "chemokine activity"           "0"     "41"
5  "GO:0042379" "chemokine receptor bindin..." "0"     "41"
6  "GO:0001664" "G-protein-coupled recepto..." "0"     "43"
7  "GO:0019838" "growth factor binding"        "0"     "47"
8  "GO:0005520" "insulin-like growth facto..." "0"     "17"
9  "GO:0004888" "transmembrane receptor ac..." "0"     "724"
10 "GO:0005515" "protein binding"              "0"     "1915"
11 "GO:0005529" "sugar binding"                "0"     "113"
12 "GO:0005102" "receptor binding"             "0"     "481"
13 "GO:0004866" "endopeptidase inhibitor a..." "0"     "99"
14 "GO:0030414" "protease inhibitor activi..." "0"     "100"
15 "GO:0001871" "pattern binding"              "0"     "82"
16 "GO:0042607" "exogenous peptide antigen..." "0"     "2"
17 "GO:0005201" "extracellular matrix stru..." "0.001" "72"
18 "GO:0005537" "mannose binding"              "0.001" "9"
19 "GO:0005539" "glycosaminoglycan binding"    "0.001" "75"
20 "GO:0005125" "cytokine activity"            "0.001" "174"
21 "GO:0030247" "polysaccharide binding"       "0.001" "77"
22 "GO:0048029" "monosaccharide binding"       "0.001" "11"
23 "GO:0004867" "serine-type endopeptidase..." "0.001" "61"
24 "GO:0019955" "cytokine binding"             "0.002" "67"
25 "GO:0046790" "virion binding"               "0.002" "4"
26 "GO:0004857" "enzyme inhibitor activity"    "0.003" "178"
27 "GO:0019199" "transmembrane receptor pr..." "0.004" "74"
28 "GO:0008430" "selenium binding"             "0.004" "5"
29 "GO:0004465" "lipoprotein lipase activi..." "0.004" "5"
30 "GO:0004896" "hematopoietin/interferon-..." "0.004" "53"
31 "GO:0005159" "insulin-like growth facto..." "0.006" "6"
32 "GO:0003706" "ligand-regulated transcri..." "0.006" "6"
33 "GO:0005509" "calcium ion binding"          "0.006" "503"
34 "GO:0008201" "heparin binding"              "0.007" "59"
35 "GO:0004714" "transmembrane receptor pr..." "0.008" "61"
36 "GO:0043120" "tumor necrosis factor bin..." "0.008" "7"
```

e. Visualize these GO terms in the GO graph. The most interesting nodes are those with ten or more annotated genes.

```
> col <- ifelse(go.pv < 0.01, ifelse(counts >= 10, "blue",
+ "orange"), "white")
> names(col) <- names(go.pv)
> att$fillcolor <- col
> plot(go, nodeAttrs=att)
```

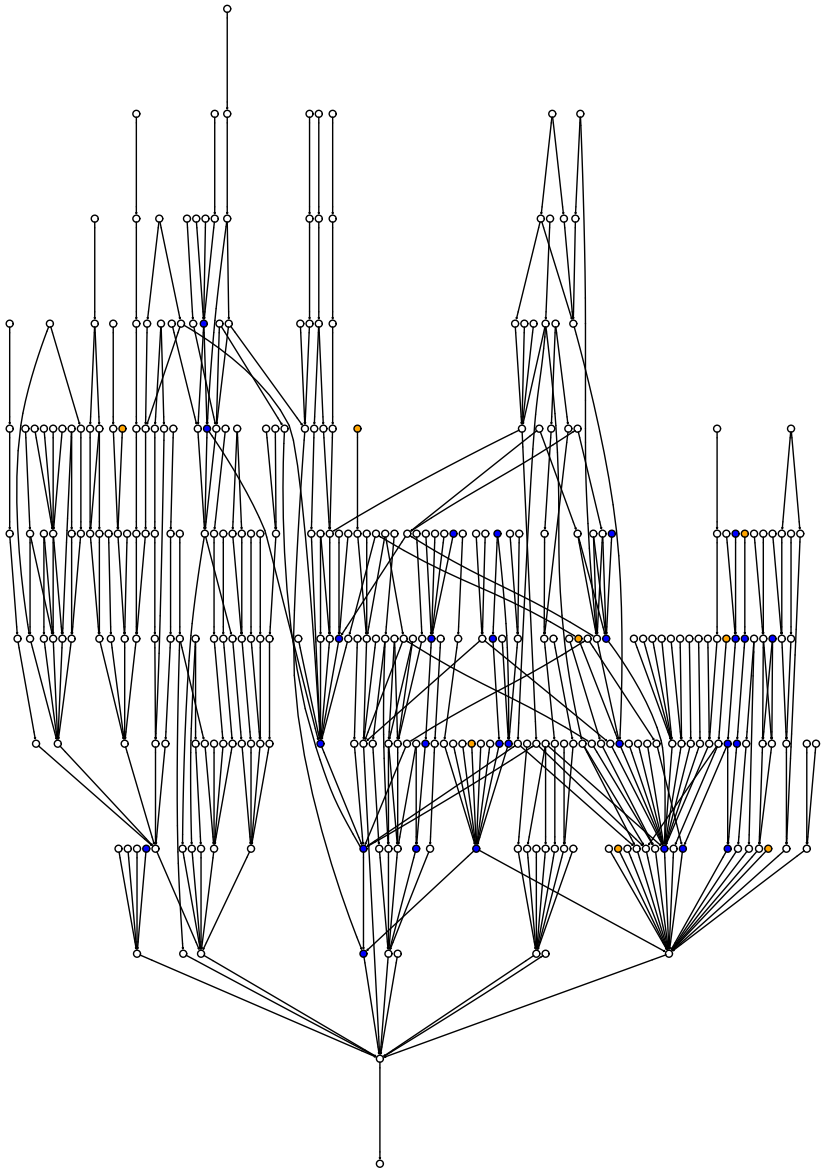

Figure 2: see exercise 3.e.