— Differential Expression and Gene Ontology —

Manuela Hummel and Ulrich Mansmann

Practical DNA Microarray Analysis, Munich, May 2006

1 Preliminaries

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

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

```
> load(url("http://compdiag.molgen.mpg.de/ngfn/data/2006/may/data.rma.RData"))
```

2 Differential expression

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, test = "t.equalvar")
> p <- 2 * pt(-abs(t), df = ncol(exprs(data.rma)) - 2)
> p.adj <- mt.rawp2adjp(p, proc = c("BY"))
```

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, ]
```

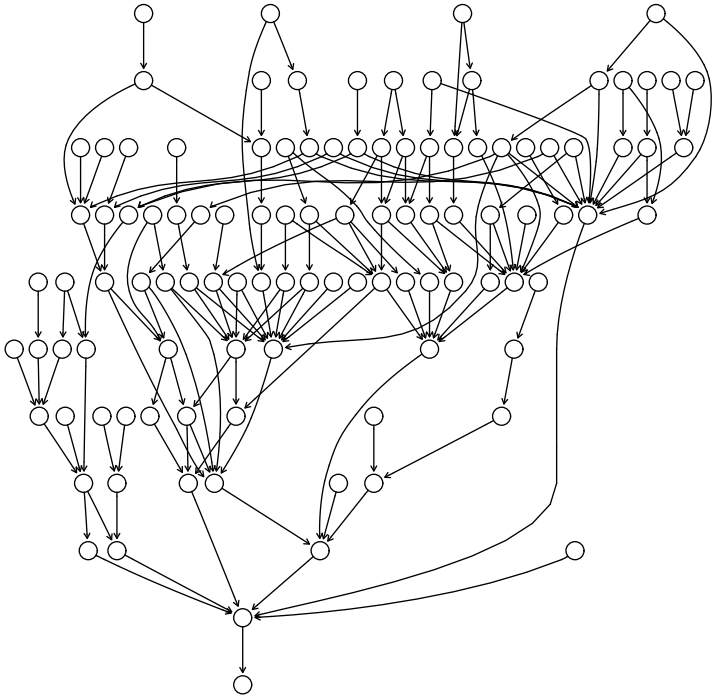
3 Gene Ontology

To explore whether the selected genes belong to a common cellular component we use the Gene Ontology.

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

Plot the GO graph.

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



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

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))[["CC"]]
> 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(si
+     c("GO ID", "Term", "p-value", "# Genes")))
```

```
   GO ID        Term                          p-value # Genes
1  "GO:0005576" "extracellular region"        "0"     "992"
2  "GO:0005615" "extracellular space"         "0"     "448"
```

```
3  "GO:0005578" "extracellular matrix (sen..." "0"     "309"
4  "GO:0031012" "extracellular matrix"         "0"     "312"
5  "GO:0005886" "plasma membrane"              "0"     "1540"
6  "GO:0005887" "integral to plasma membra..." "0"     "1143"
7  "GO:0031226" "intrinsic to plasma membr..." "0"     "1147"
8  "GO:0016021" "integral to membrane"         "0"     "2727"
9  "GO:0031224" "intrinsic to membrane"        "0"     "2731"
10 "GO:0005604" "basement membrane"            "0"     "46"
11 "GO:0005602" "complement component C1q ..." "0"     "2"
12 "GO:0001772" "immunological synapse"        "0.002" "27"
13 "GO:0005581" "collagen"                     "0.004" "31"
14 "GO:0016020" "membrane"                     "0.006" "3520"
15 "GO:0042627" "chylomicron"                  "0.006" "6"
16 "GO:0005587" "collagen type IV"             "0.006" "6"
17 "GO:0030935" "sheet-forming collagen"       "0.009" "7"
```

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)
```