— Group testing —
Gene set enrichments and global tests

Ulrich Mansmann

Practical DNA Microarray Analysis, November 2004, Berlin
http://compdiag.molgen.mpg.de/ngfn/pma2004nov.shtml

Abstract. *This is the tutorial for the exercises on day 4 of the course on Practical DNA Microarray Analysis. Topics are gene set enrichment methods introduced by Lamb et al. (2003) and a global test introduced by Goeman et al. (2004)*

1 Gene set enrichment

Load data and packages. Look at the data in the object eSet.

```
> loadURL("http://compdiag.molgen.mpg.de/ngfn/data/2004/nov/grouptesting.RData")
> library(affy)
> library(multtest)
> str(eSet)
> dim(exprs(eSet))
> dim(pData(eSet))
```

Pathway information and relevant probesets are noted in the dataframe pathways.U133A. Which pathways are considered and how many probesets belong to each pathway?

```
> pw <- pathways.U133A$PATHWAY
> table(pw)
```

In order to perform the gene set enrichment procedure, it is necessary to define two gene groups. Group A consists of all probe sets that are related to cancer specific pathways, group B consists of all other probe sets. Some probe sets belong to multiple pathways. How many belong to 2 or 3 different pathways? How many different probe sets are related to cancer relevant pathways?

```
> ps <- pathways.U133A$AFFYPROBESET
> table(table(ps))
> sum(table(table(ps)))
> length(unique(ps))
```

There are nrow(exprs(eSet)) probesets in total. The number of probesets in group A is given by length(unique(ps)). A group index vector is defined by

```
> group.ind <- rep("B", nrow(exprs(eSet)))
> group.ind <- ifelse(rownames(exprs(eSet)) %in% ps, "A", group.ind)
```

First, the differential gene expression for each single gene will be quantified by using the t-statistic (Welch test when not supposing a common variance). Differential expression between UICC II and UICC III patients is of interest. In the dataframe `pData(eSet)` this information is given in column "UICC". For the function `mt.teststat` from the `multtest` library the group index has to be 0/1 coded. This coding is achieved by

```
> UICC.gr <- pData(eSet)$UICC - 2
> table(UICC.gr)
```

The primary interest is to solve the question if the probe sets in the cancer related pathways show more differential expression than the remaining probesets. Therefore the direction of the differential expression can be ignored by taking the absolute value of the `t` statistics. Up- or down regulation of a gene between UICC II and UICC III patients is not taken into account and emphasis is put on regulation per se. The gene-wise quantification of the differential gene expression is performed by

```
> require(multtest)
> UICC.diff <- abs(mt.teststat(exprs(eSet), UICC.gr))
```

Now the statistics `Myy` for the gene set enrichment procedure can be calculated. The basic idea is to sort the `UICC.diff` values and to analyze at which positions members of group A and group B can be found. In a second step a score vector is created by assigning the score –nB to each group A member and the score nA to each group B member. In a third step the cummulative sum is calculated and the statistic `Myy` can be read off.

```
> order <- order(UICC.diff)
> table(group.ind)
> UICC.score <- ifelse(group.ind == "A", -table(group.ind)["B"],
+     table(group.ind)["A"])
> UICC.score <- UICC.score[order]
> yy <- cumsum(UICC.score)
> M.yy <- max(abs(min(yy)), max(yy))
> M.yy
```

The result can be presented as a figure that shows the course of the cummulatice sum. Additionally the position of elements of group A is spotted by points.

```
> plot(0:length(yy), c(0, yy), type = "l", xlab = "Rank", ylab = "Score")
> loc.A <- UICC.score == as.numeric(names(table(UICC.score)[1]))
> xxx <- which(loc.A)
> yyy <- rep(min(yy), length(xxx))
> points(xxx, yyy, pch = ".")
```

Unfortunately, the figure does not look very convincing. There is an up and down of the line, the spots of group A can be found everywhere. Of course the figure could be improved by adding some information about the density of the A points. In the next step the permutation test will be performed in order to obtain a useful statistical assessment for answering the question of interest. The permutation test is performed by the function `gene.set.enrichment.permut.rfc` that can be found in the workspace.

```
> gene.set.enrichment.permut.rfc(UICC.diff, group.ind, 1000)
```

Interpret the result of the test. Is it possible to state, that there is no difference between the differential expression in probe sets which belong to cancer related pathways and the remaining probe sets on the array? The question can be formulated more specific by looking at a single pathway within the group of cancer related pathways. The names of the nine pathways under consideration are written into the object pathways.

```
> pathways <- unique(pathways.U133A$PATHWAY)
```

The following analysis will be restricted to the 1407 probesets which where extracted. The relevant probesets are listed in `PSS.pathway` and the relevant information on differential gene expression is given in `UICC.pathway`.

```
> UICC.pathway <- UICC.diff[group.ind == "A"]
> PSS.pathway <- rownames(exprs(eSet)) %in% ps
> PSS.pathway <- rownames(exprs(eSet))[PSS.pathway]
```

For analyzing the androgen receptor signaling pathway one needs to define a group variable which indicates the probe sets belonging to this specific pathway. Then the analysis can be performed.

```
> pss.special <- unique(ps[pw == pathways[1]])
> group.special <- ifelse(PSS.pathway %in% pss.special, "A", "B")
> gene.set.enrichment.permut.rfc(UICC.pathway, group.special, 10000)
```

Simply by changing the index number for a pathway helps to check all candidates. The following loop generates a table with the entire relevant information. The number of permutations is set to 10000. This calculation can take some time. Now you have the chance to go for a coffee.

```
> pathway.results <- NULL
> for (i in 1:9) {
+     pss.special <- unique(ps[pw == pathways[i]])
+     group.special <- ifelse(PSS.pathway %in% pss.special, "A", "B")
+     res <- gene.set.enrichment.permut.rfc(UICC.pathway, group.special, 10000)
+     pathway.results <- rbind(pathway.results, unlist(res))
+ }
> rownames(pathway.results) <- pathways
```

2 Global test

This exercise studies differential gene expression in the p53 pathway between UICC II and UICC III patients. The pathway is represented by 45 probe sets. The library `globaltest` is needed.

```
> library(globaltest)
> p53.pss <- ps[pw == "p53_signalling"]
> test.genes <- which(rownames(exprs(eSet)) %in% p53.pss)
> gt.p53 <- globaltest(eSet, "UICC", test.genes = test.genes)
> geneplot(gt.p53)
> checkerboard(gt.p53)
> sampleplot(gt.53)
```

3 References

Lamb J, Ramaswamy S, Ford HL, Contreras B, Martinez RV, Kittrell FS, Zahnow CA, Patterson N, Golub TR, Ewen ME. A mechanism of cyclin D1 action encoded in the patterns of gene expression in human cancer. Cell 2003; 114(3):323-34.

Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics 2004; 20(1):93-9.