Group testing: global tests

Ulrich Mansmann
Department of Medical Biometrics and Informatics
University of Heidelberg

Overview

- Gene set enrichment
  Lamb J et al. (2003) *A mechanism of Cyclin D1 Action Encoded in the Patterns of Gene Expression in Human Cancer*,
  Cell, 114: 323-334
- Global test:
  Goeman JJ. Et al. (2003) *A global test for groups of genes: Testing association with a clinical outcome*, Bioinformatics, 20:93-99
  Bioconductor package: *globaltest*
- Example: Differential gene expression between UICC stages II / III colon cancer patients (Groene / Mansmann).

Two questions about group of genes

Question 1: Two groups of genes have to be compared with respect to gene expression: Is the gene expression in gene group A different from the expression in gene group B.

Genes of group A	Genes of group B

Question 2: Is there differential gene expression between different biological entities not in terms of single genes but with respect to a defined group of genes.

	Entity I	Entity II
Well defined group of genes		

Example: Colon Cancer

Study: 18 patients with UICC II colon cancer, 18 patients with UICC III colon cancer, HG-U133A, 22.283 probesets representing ~18.000 genes. Snap-frozen material, laser microdisection.

Question 1: Is the differential gene expression between UICC II /III patients more distinct for genes in cancer related pathways compared to genes in other pathways?

Question 2: Is there differential gene expression in the p53 signalling pathway?

Gene set enrichment

Problem:

Two groups of genes have to be compared with respect to gene expression: Is the gene expression in gene group A different from the expression in gene group B.

Basic idea:

n_A genes in group A, n_B genes in group B

Order the genes with respect to the expression value. If there is a difference between both groups, the expression values will be separated. **The position of a value in group A will have the tendency to be high or low**. In case of no difference, the values will be nicely mixed.

Gene set enrichment

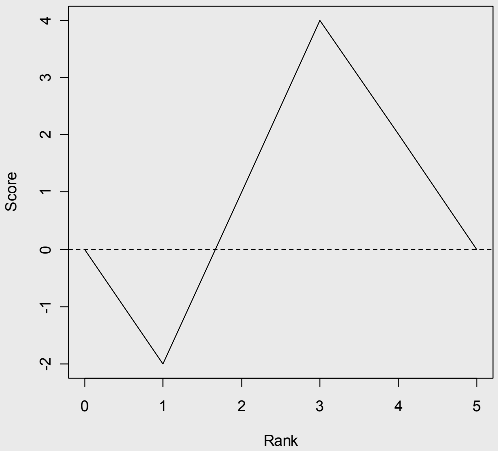
Basic idea:

- n_A genes in group A, n_B genes in group B.
- Order the genes with respect to expression values.
- Create a vector vv of (n_A+n_B) components with value $-n_B$ at each position where a value from group A is sitting and with value n_A at each position where a value from group B is sitting.
- Calculate yy = cumsum(vv).
- Draw a line starting at (0,0) through points (i, yy[i]). The line will end in $(n_A+n_B,0)$ because $(-n_B)\cdot n_A + n_A\cdot n_B = 0$.
- Look at $M_{vv} = \max\{|\min(yy)|, \max(yy)\}$ which will be large in case of a good separation between both groups.
- Permute the vector vv to get vv*, calculate yy* and M_{vv^*}. Use permutation to calculate the distribution of M_{vv} under the Null hypothesis, determine the permutation based p-value: $p_{perm} = \#\{M_{vv^*} \geq M_{vv}\} / \#$ permutations.

Gene set enrichment

Simple Example

- Gene expression in group A: {2, 3}, $n_A = 2$
  Gene expression in group B: {1, 4, 6}, $n_B = 3$
- Order the genes with respect to expression values.
  {**1**, 2, 3, **4**, **6**}
- vv = {-2, 3, 3, -2, -2}
- yy = {-2, 1, 4, 2, 0}
- $M_{vv} = 4$
- Distribution of M_{vv} under the Null hypothesis
  2 ~ 0.1; 3 ~ 0.3, 4 ~ 0.4, 6 ~ 0.2
  10000 permutations
- $p_{perm} = \#\{M_{vv^*} \geq M_{vv}\} / \#$ permutations = 0.4+0.2 = 0.6



Gene set enrichment – Colon cancer

1407 probe sets are studied which belong to 9 cancer specific pathways.

androgen_receptor_signalling	122
apoptosis	245
cell_cycle_control	51
notch_delta_signalling	50
p53_signalling	45
ras_signalling	316
tgf_beta_signalling	100
tight_junction_signalling	425
wnt_signalling	214

Gene set enrichment – Colon cancer

	group.A	group.B	M_{yy}	p.value
androgen_receptor_signaling	118	1289	6983	0.0568
Apoptosis	238	1169	17801	0.7438
cell_cycle_control	51	1356	10413	0.3616
notch_delta_signalling	50	1357	9010	0.6492
p53_signalling	45	1362	12390	0.0924
ras_signalling	311	1096	15486	0.6252
tgf_beta_signaling	100	1307	22615	0.0128
tight_junction_signaling	406	1001	15456	0.4414
wnt_signaling	214	1193	16318	0.8432

Goeman's Global Test

- Test if global expression pattern of a group of genes is significantly related to some outcome of interest (groups, continuous phenotype).
- If this relationship exists, then the knowledge of gene expression helps to improve the prediction of the phenotype of interest. If the prediction can not improved by knowing the gene expression then there will not be differential gene expression.
- Test statistic:

$$Q \sim (Y-\mu)'R\,(Y-\mu)$$

$$\sim \Sigma\,[X_i'(Y-\mu)]^2 \quad \text{sum over genes of the pathway}$$

$$\sim \Sigma\,\Sigma\,R_{ij}(Y_i-\mu)\,(Y_j-\mu)$$

sum over subjects

μ: Mean of phenotype, X_{mi} Expression for gene m in subject i
R : X'X: IxI matrix of correlation between gene expression of subjects

Goeman's Global Test - Example

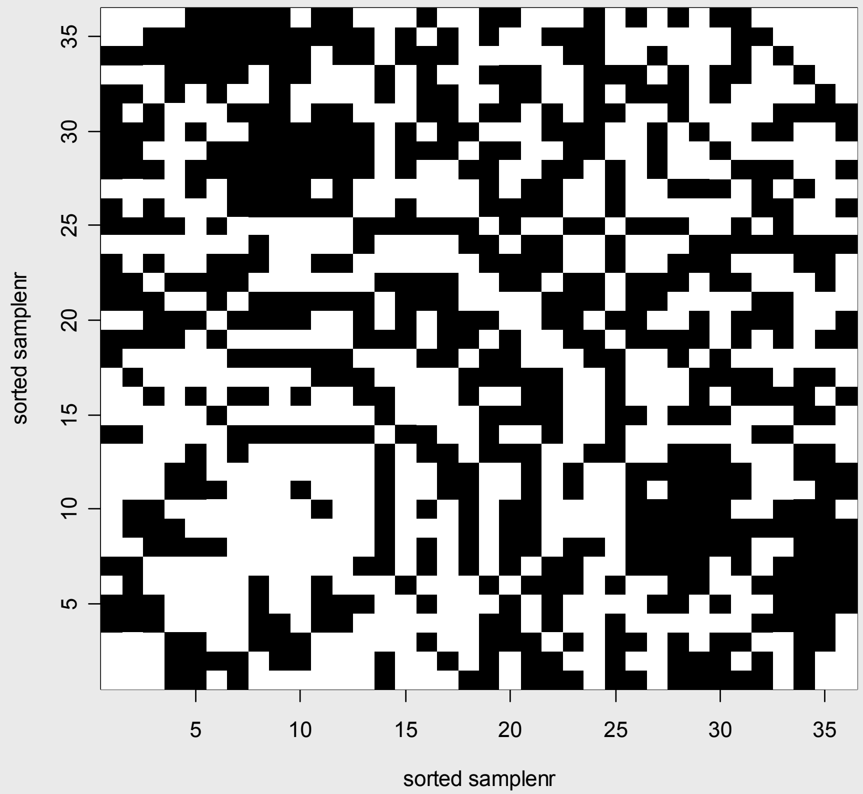
- Test for differential gene expression in *p53 signalling* pathway
  45 probesets

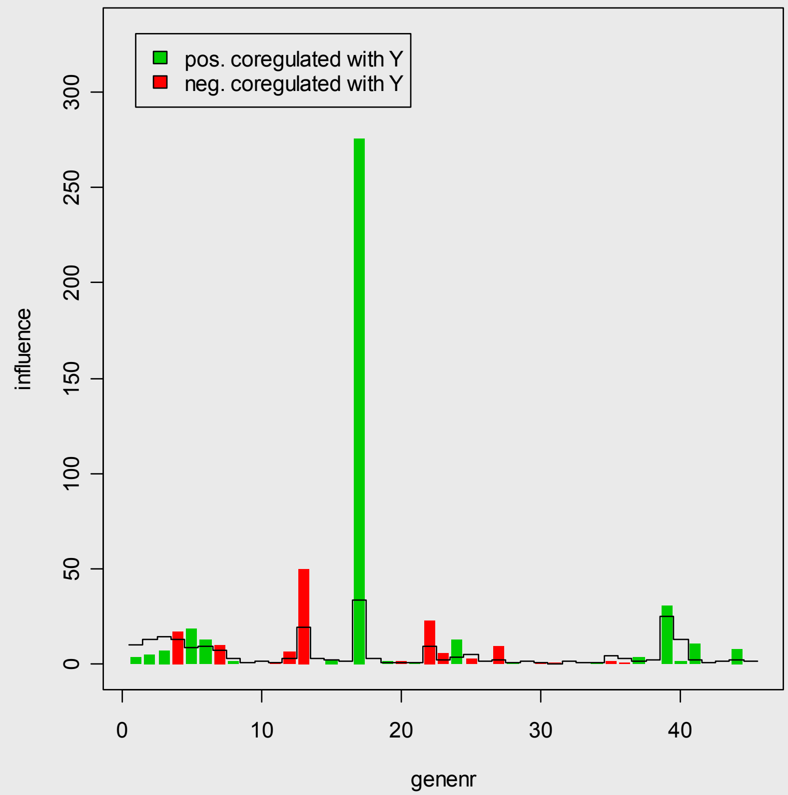
- Global Test result:

```
45 out of 45 genes used; 36 samples

p value = 0.0114
based on 10000 permutations

Test statistic Q = 11.78
with expectation EQ = 5.466
and standard deviation sdQ = 2.152 under the null hypothesis
```

- Informative plots:
  Sample plot: how good fits a sample to its phenotype
  Checkerboard: Correlation between samples
  Gene plot: Influence of single genes to test statistics

Goeman’s Global Test - Example





Goeman’s Global Test - Example