Large Scale Multiple Hypothesis Testing

Structured Problems

Jelle Goeman and Aldo Solari

36th Annual Conference of the International Society for Clinical Biostatistics
August 23, 2015 - Utrecht

Outline

1. **Global test and methods for graph-structured hypotheses**

2. **Removing unwanted variation by using negative controls**

Testing a group of covariates

Group of covariates

$\mathcal{G}$ = group of g covariates (e.g. group of genes)

Null hypothesis $\bigcap \mathcal{G}$ (self-contained)

None of the covariates within $\mathcal{G}$ are associated with the response

Goeman and Buhlmann (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23:980–987.

Alternative hypothesis

At least one covariate within $\mathcal{G}$ is associated with the response

Example

$\boxed{ABC}$: A, B and C are not associated with the outcome

$\xcancel{ABC}$: at least one among $\boxed{A}$ $\boxed{B}$ $\boxed{C}$ is false

Logical relationships

For any $\mathcal{G}' \subset \mathcal{G}$

- the truth of $\bigcap \mathcal{G}$ implies the truth of $\bigcap \mathcal{G}'$
- the falsehood of $\bigcap \mathcal{G}'$ implies the falsehood of $\bigcap \mathcal{G}$

Example

$$\begin{tikzcd}
\boxed{ABC} \arrow[d] & \xcancel{ABC} \\
\boxed{AB} & \xcancel{AB} \arrow[u]
\end{tikzcd}$$

Models

Generalized linear model

$$\mathrm{E}(\underset{n\times 1}{y}) = \ell^{-1}\left(\underset{n\times g}{X}\,\underset{g\times 1}{\beta} + \underset{n\times q}{Z}\,\underset{q\times 1}{\gamma}\right)$$

where

- ℓ is a monotone link function (e.g. logit)
- Y contains the response (e.g. treatment vs control)
- X contains the group of covariates (e.g. genes)
- Z contains the nuisance covariates (e.g. intercept, age)

Cox proportional hazards model

$$h(\underset{n\times 1}{t}) = h_0(\underset{n\times 1}{t}) \exp\left(\underset{n\times g}{X}\,\underset{g\times 1}{\beta} + \underset{n\times q}{Z}\,\underset{q\times 1}{\gamma}\right)$$

where h is the hazard function

Global test

Null hypothesis $\bigcap \mathcal{G}$

$$\bigcap_{i=1}^{g} \{\beta_i = 0\}$$

Global test

Score test statistic, tailored for high-dimensional $g \gg n$

- Goeman et al. (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20:93–99
- Goeman et al. (2005). Testing association of a pathway with survival using gene expression data. *Bioinformatics*, 21:1950–1957.
- Goeman et al. (2006). Testing against a high-dimensional alternative. *JRSS-B*, 68:477–493.
- Goeman et al. (2011). Testing against a high-dimensional alternative in the generalized linear model: asymptotic type I error control. *Biometrika*, 98:381–390.

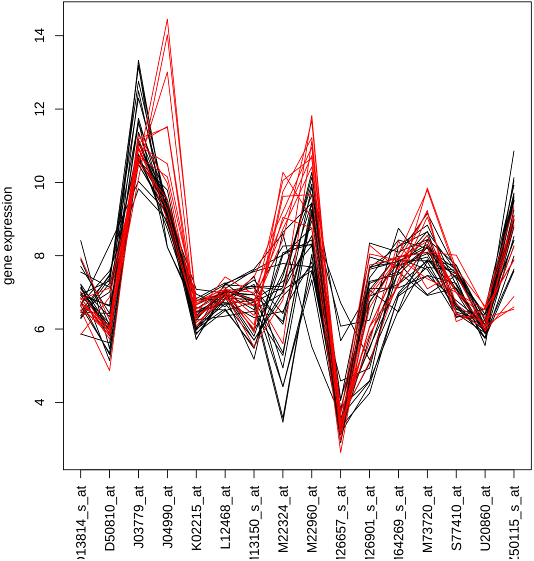
Application: Golub data

Leukemia ALL/AML study with 38 patients with gene expression over 7129 genes

- $\mathcal{G}$: renin-angiotensin system pathway
- ℓ: logit function
- y: AML vs ALL
- X: 16 genes belonging to $\mathcal{G}$ ($g = 16$)
- Z: intercept

```
gtKEGG(ALL.AML, Golub_Train, id = "04614")
                         alias  p-value Statistic Expected Std.dev #Cov
04614 Renin-angiotensin system 7.94e-11      23.5      2.7    1.47   16
```

Application: Golub data



Reducing the number of test: selection

B ? C ? A_1 A_2

Fewer hypotheses

More powerful FWER control and FDP confidence
FDR not always true

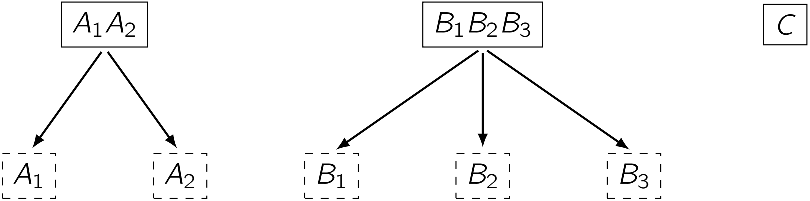
Uninteresting hypotheses

Discard uninteresting hypotheses (e.g. non annotated probes) before testing

Low power hypotheses

Discard hypotheses with low power (e.g. probes with low mean or variance of expression) before testing
Selection must be independent of the p-value if true hypothesis

Reducing the number of test: aggregation



To the level of interest

Aggregate to a lower level of resolution of interest (e.g. from probe to gene level)

Several levels simultaneously

If several levels of resolutions are of interest, use hierarchical multiple testing methods that test more than one level of resolution simultaneously

Graph-structured hypotheses

FWER control and FDP confidence

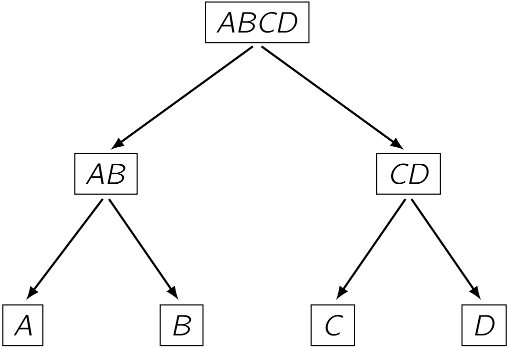
Structure	Method	`R function` (*package*)
Tree	Meinshausen (2005)	-
	Goeman & Finos (2012)	`inheritance` (*globaltest*)
DAG	Goeman & Mansmann (2008)	`focusLevel` (*globaltest*)
	Meijer & Goeman (2015a)	`DAGmethod` (*cherry*)
	Meijer & Goeman (2015b)	`structuredHolm` (*cherry*)
Region	Meijer, Krebs & Goeman (2015)	`regionmethod` (*cherry*)

FDR

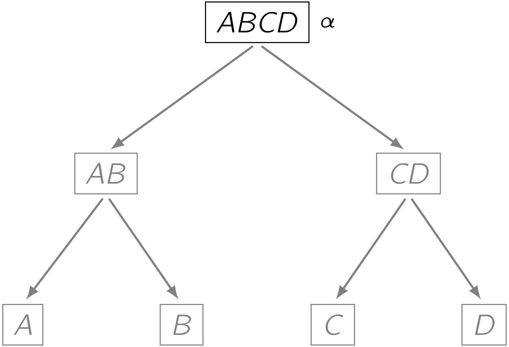
Yekutieli (2008) Hierarchical false discovery rate-controlling methodology. *JASA*, 103:309–316

Benjamini & Bogomolov (2014) Selective inference on multiple families of hypotheses. *JRSS-B*

Trees

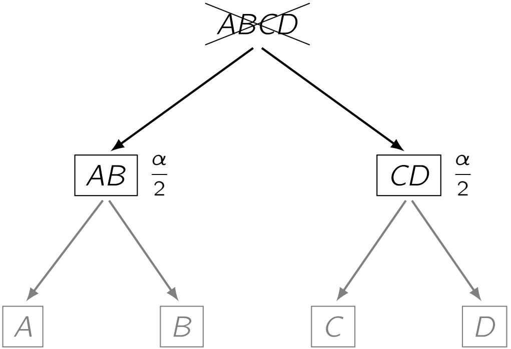


Symmetric binary tree



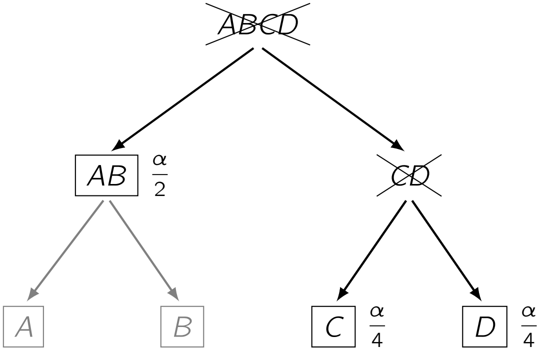
Top-down testing of $\mathcal{N}$ at $\alpha \cdot \frac{\#\mathcal{N}}{\#\mathcal{L}}$: test $ABCD$ at $\alpha \cdot \dfrac{4}{4}$

Symmetric binary tree



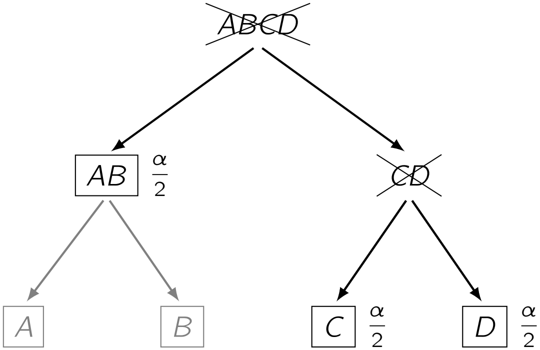
ABCD rejected: test *AB* and *CD* at $\alpha \cdot \dfrac{2}{4}$

Symmetric binary tree



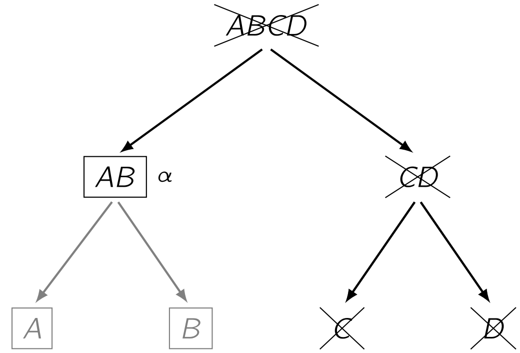
CD rejected: test *C* and *D* at $\alpha \cdot \frac{1}{4}$

Symmetric binary tree

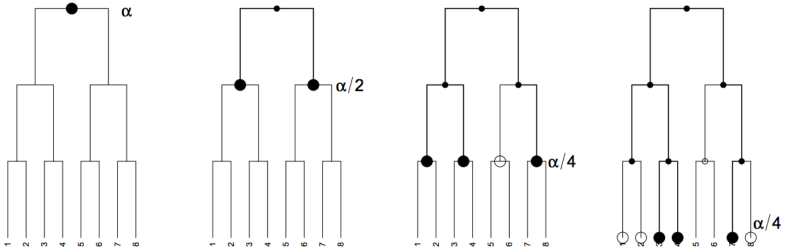


At least C or D must be false: test C and D at $2\alpha \cdot \dfrac{1}{4}$ (Shaffer)

Symmetric binary tree



C and D rejected: test AB at α



Toy example with 8 hypotheses that form a binary tree.

leftmost panel: the global null hypothesis is tested at level α *and it is rejected;*

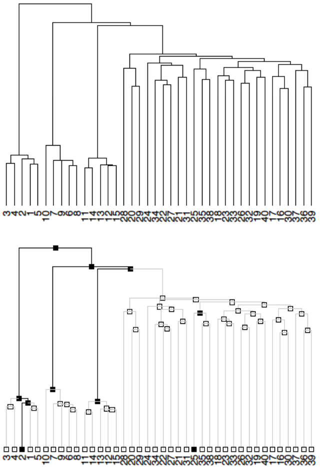
second panel: it is examined if the effect can be attributed to one or both of the sub-clusters that follow in the hierarchy. Each of these two sub-clusters is tested at level $\alpha/2$*. They are again both rejected;*

third panel: the procedure turns to the next four clusters, which are tested at level $\alpha/4$*. Of these four hypotheses, one cluster made up of variables 5 and 6 is not rejected;*

rightmost panel: variables 5 and 6 are not tested anymore at the individual level in the last step. Note that the remaining 6 hypotheses can be tested at level $\alpha/4$ *and not* $\alpha/8$ *(Shaffer' improvement)*

Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95:265–278.

Data-driven tree

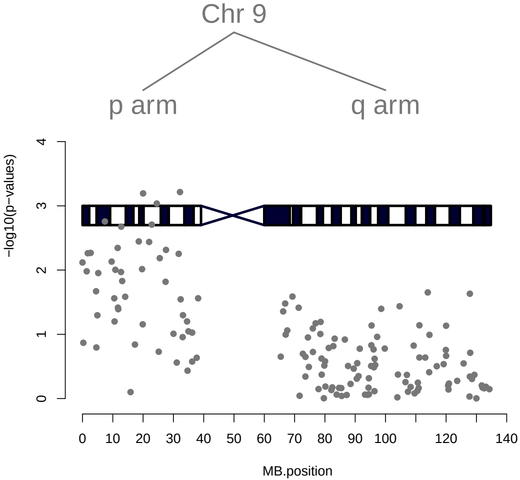


Example (d) of Zou and Hastie (2005)

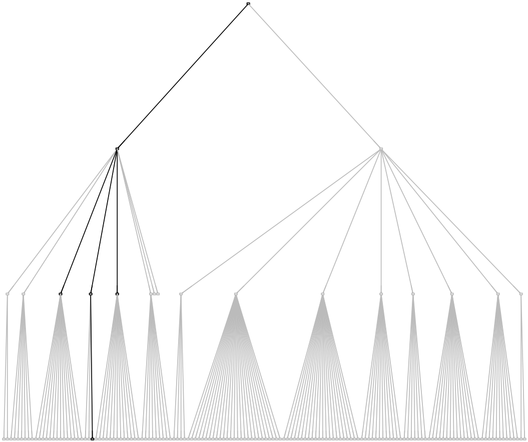
top: hierarchical clustering structure (complete linkage, with Spearman correlation as distance) that enters the testing procedure;

bottom: testing result, where rejected nodes are indicated by darker edges;

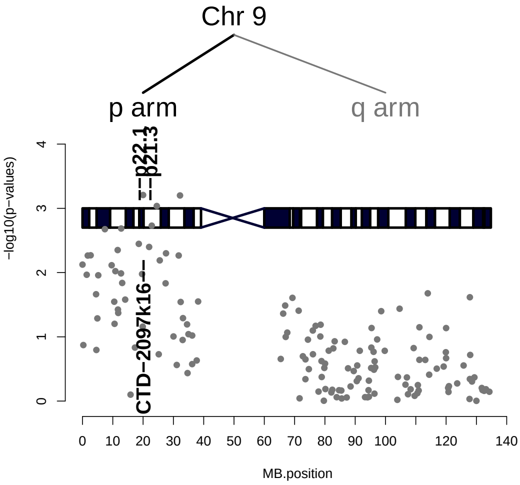
Meinshausen, N. (2008). Hierarchical testing of variable importance. *Biometrika*, 95:265–278.



Goeman and Finos (2012) The Inheritance Procedure: Multiple Testing of Tree-structured Hypotheses, *Statistical Applications in Genetics and Molecular Biology*, Vol. 11, Iss. 1, Article 11



Goeman and Finos (2012) The Inheritance Procedure: Multiple Testing of Tree-structured Hypotheses, *Statistical Applications in Genetics and Molecular Biology*, Vol. 11, Iss. 1, Article 11



Goeman and Finos (2012) The Inheritance Procedure: Multiple Testing of Tree-structured Hypotheses, *Statistical Applications in Genetics and Molecular Biology*, Vol. 11, Iss. 1, Article 11

Alternative use of Shaffer

At the leaves only?

Mehinshausen/Inheritance exploits Shaffer only at the leaves

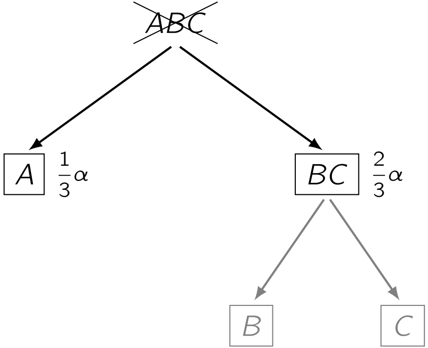
Question

Can we exploit logical relationships elsewhere?

Answer

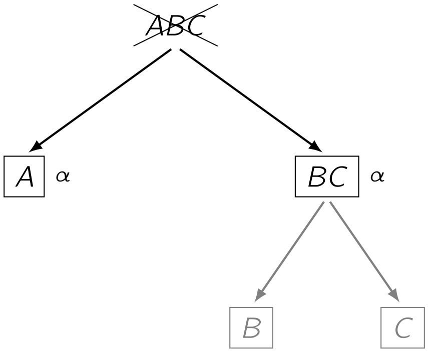
Yes. But not simultaneously in two places

Alternative use of Shaffer



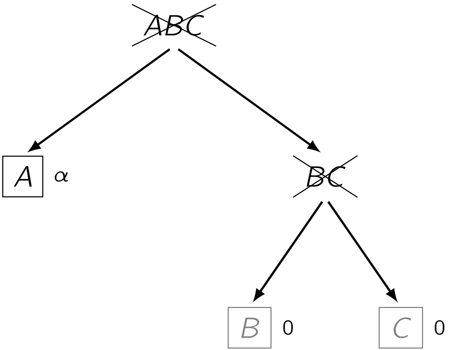
Rejection of ABC implies that at least A or BC must be false

Alternative use of Shaffer



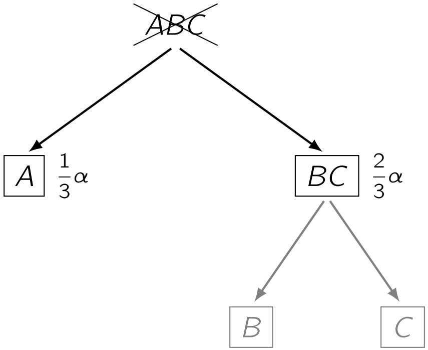
Why not testing both at α?

Alternative use of Shaffer



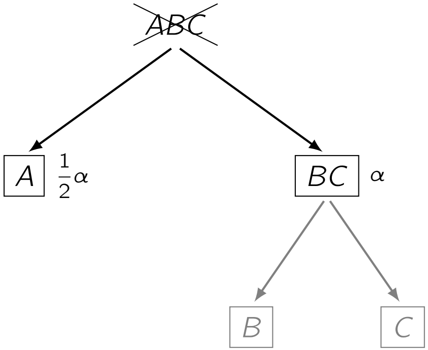
BC rejected, but we cannot increase the α of B and C

Alternative use of Shaffer



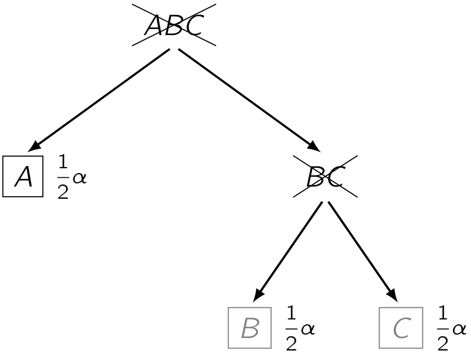
Rejection of ABC implies that at least A, B or C must be false

Alternative use of Shaffer



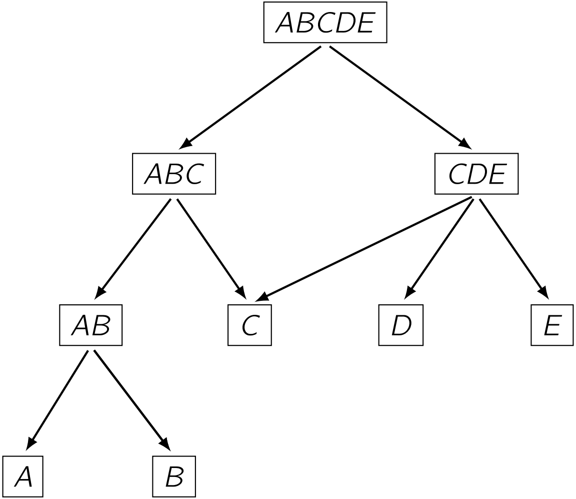
Test A at $\alpha/2$ and BC at α

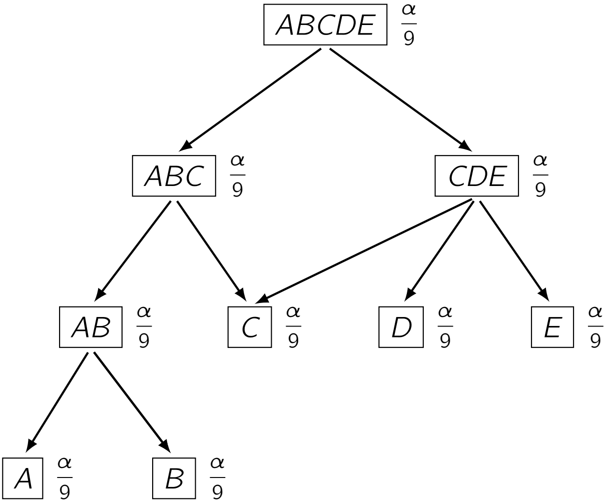
Alternative use of Shaffer



Test B and C at $\alpha/2$

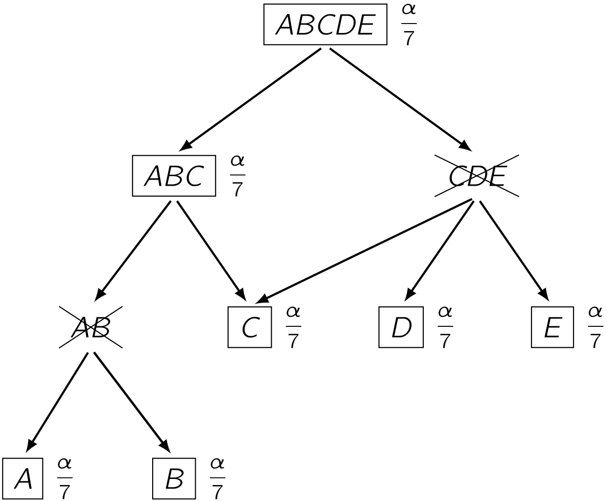
Directed Acyclic Graphs



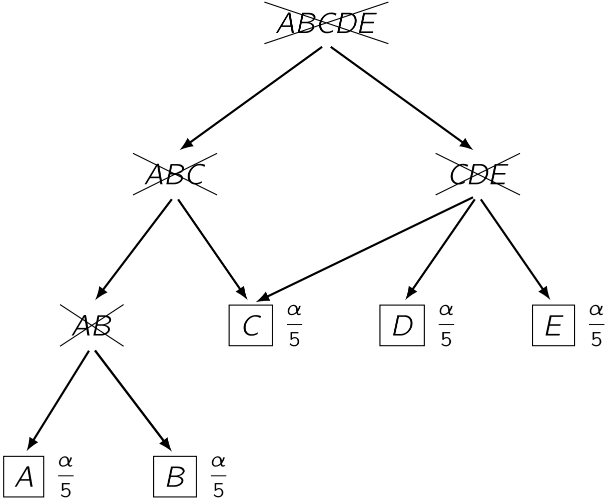


Bonferroni correction: each hypothesis is tested on $\alpha/9$

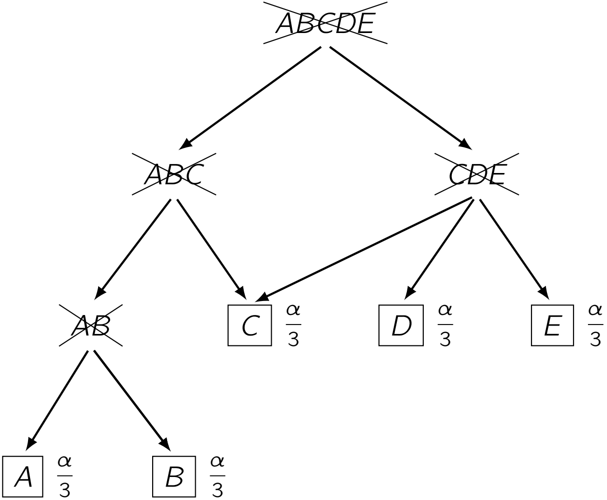
Meijer, R. and Goeman, J. (2015). Multiple testing of gene sets from gene ontology: possibilities and pitfalls. *Submitted for publication*



Two hypotheses have been rejected (denoted by the crosses). Holms procedure would test the remaining hypotheses on $\alpha/7$

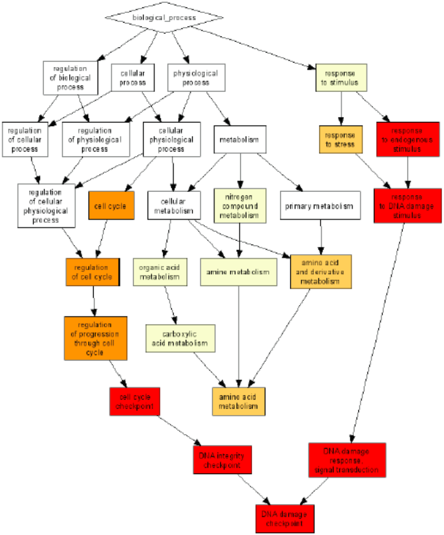


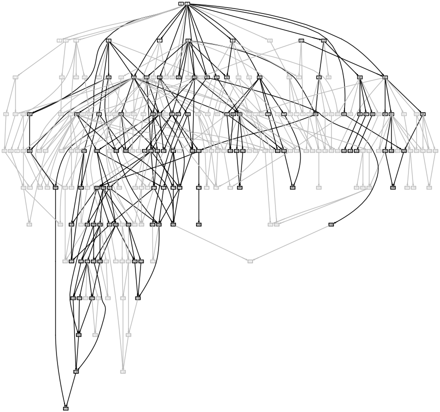
However, using the one-way relations (the falsehood of an hypothesis implies the falsehood of all ancestor hypotheses) shows that two hypotheses can no longer be true. The remaining hypotheses can be tested on $\alpha/5$



Using the two-way relations (the falsehood of an hypothesis implies also the falsehood of at least one of its corresponding leaf nodes), we furthermore know that one of the hypotheses corresponding to gene C, D or E, and one of the hypotheses corresponding to gene A or B have to be false as well. Maximally 3 hypotheses can be simultaneously true.

Gene Ontology

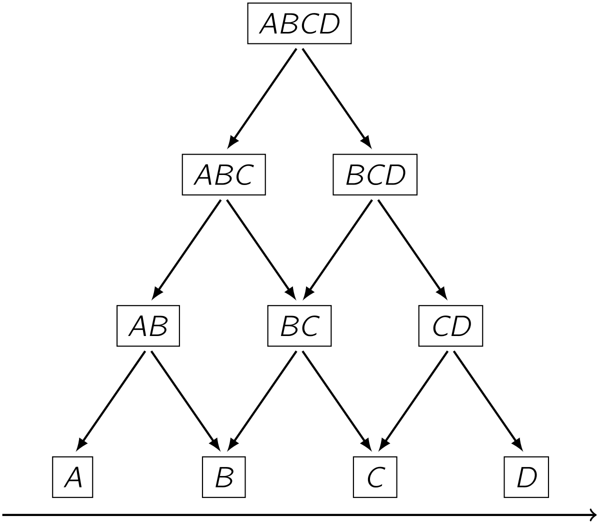

biological_process
regulation of biological process
cellular process
physiological process
response to stimulus
regulation of cellular process
regulation of physiological process
cellular physiological process
metabolism
response to stress
response to endogenous stimulus
regulation of cellular physiological process
cell cycle
cellular metabolism
nitrogen compound metabolism
primary metabolism
response to DNA damage stimulus
regulation of cell cycle
organic acid metabolism
amine metabolism
amino acid and derivative metabolism
regulation of progression through cell cycle
carboxylic acid metabolism
cell cycle checkpoint
amino acid metabolism
DNA integrity checkpoint
DNA damage response, signal transduction
DNA damage checkpoint



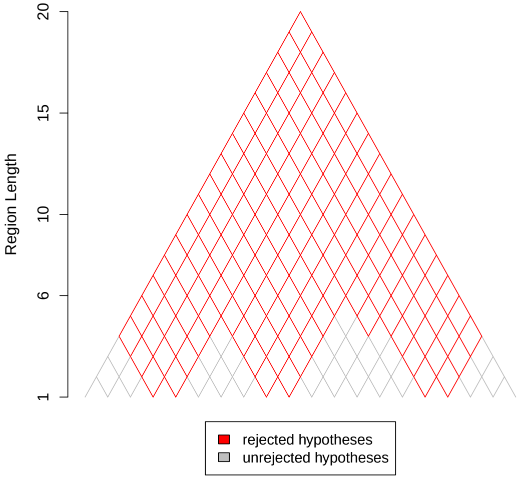
Goeman and Mansmann (2008) Multiple testing on the directed acyclic graph of gene ontology. *Bioinformatics*, 24:537–544.

Meijer and Goeman (2015). A multiple testing method for hypotheses structured in a directed acyclic graph. *Biometrical Journal*, 57:123–143.

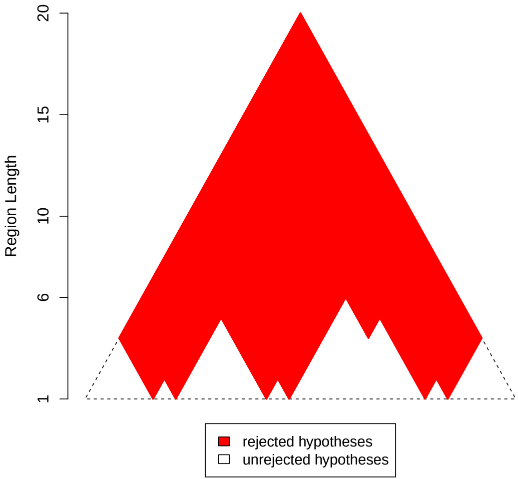
Regions



Hypotheses ordered in space or time



Meijer, Krebs and Goeman (2015) A region-based multiple testing method for hypotheses ordered in space or time. *Statistical Applications in Genetics and Molecular Biology*, 14:1–19



Meijer, Krebs and Goeman (2015) A region-based multiple testing method for hypotheses ordered in space or time. *Statistical Applications in Genetics and Molecular Biology*, 14:1–19

R lab: graphstruct

Outline

1. Global test and methods for graph-structured hypotheses

2. Removing unwanted variation by using negative controls

Unwanted variation

Unwanted variation

High-dimensional data suffer from unwanted variation (e.g. batch effects in microarray data)

Consequences

Unwanted variation may lead to high rates of false discoveries, high rates of missed discoveries, or both

Negative controls

Negative controls are covariates that are known a priori to be truly unassociated with the factor of interest (e.g. housekeeping genes)

Negative controls can be used for identifying the unwanted variation!

Two-step method

Leek and Storey (2007, 2008) and Gagnon-Bartsch and Speed (2012) proposed methods (SVA and RUV-2) to adjust for unwanted variation using negative controls.

RUV-2 is a simple, two-step method:

1. perform SVD on negative controls to estimate the unwanted factors
2. regress the response on covariates of interest and (estimated) unwanted factors

Leek and Storey (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 3, e161.

Gagnon-Bartsch and Speed (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, 13(3):539–552.

Multivariate linear model

$$\underset{n\times m}{Y} = \underset{n\times p}{X}\ \underset{p\times m}{\beta} + \underset{n\times q}{Z}\ \underset{q\times m}{\gamma} + \underset{n\times k}{W}\ \underset{k\times m}{\theta} + \underset{n\times m}{\varepsilon}$$

where

- Y contains the response (e.g. gene expression of m genes)
- X contains the covariate of interest (e.g. treatment vs control)
- Z contains the nuisance covariates (e.g. intercept, age)
- W contains the unobserved covariates (e.g. sample quality)

In what follows, we will consider for simplicity that there are no nuisance covariates

Step 1: estimate of W by negative controls

- From the submatrix $\underset{n\times c}{Y}$ containing c negative controls:

$$\underset{n\times c}{Y} = \underset{n\times p}{X}\ \underbrace{\underset{p\times c}{\beta}}_{=0} + \underset{n\times k}{W}\ \underset{k\times c}{\theta} + \underset{n\times c}{\varepsilon}$$

- Perform the singular value decomposition (SVD) of $\underset{n\times c}{Y}$:

$$\underset{n\times c}{Y} = \underset{n\times n}{U}\ \underset{n\times c}{\Lambda}\ \underset{c\times c}{V}^{\top}$$

- Estimate W

$$\widehat{\underset{n\times k}{W}\ \underset{k\times c}{\theta}} = \underset{n\times n}{U}\ \underset{n\times c}{\Lambda^k}\ \underset{c\times c}{V}^{\top}$$

$$\underset{n\times k}{\hat{W}} = \underset{n\times n}{U}\ \underset{n\times c}{\Lambda^k}$$

where Λ^k contains only the k largest singular values (setting others to zero)

Step 2: estimate β by regressing Y on X and $\hat{W}$

$$\underset{p\times m}{\hat{\beta}} = (\underset{n\times p}{X}^\top \underset{n\times n}{R_{\hat{W}}} \underset{n\times p}{X})^{-1} \underset{n\times p}{X}^\top \underset{n\times n}{R_{\hat{W}}} \underset{n\times m}{Y}$$

where $\underset{n\times n}{R_{\hat{W}}} = \underset{n\times n}{I} - \underset{n\times k}{\hat{W}}(\underset{n\times k}{\hat{W}}^\top \underset{n\times k}{\hat{W}})^{-1} \underset{n\times k}{\hat{W}}^\top$ projects onto the orthogonal complement of the column space of $\hat{W}$

If $\hat{W} = W$ and $\hat{W} \perp X$

$$\underset{n\times n}{R_{\hat{W}}} \underset{n\times m}{Y} = \underset{n\times p}{X} \underset{p\times m}{\beta} + \underset{n\times n}{R_{\hat{W}}} \underset{n\times m}{\varepsilon}$$

R lab: ruv