Multiple testing on the graph of Gene Ontology

Jelle Goeman[1] Ulrich Mansmann[2]

[1]Medical Statistics
Leiden University Medical Center

[2]IBE, University of Munich

CAMDA Vienna, 2008-12-06

Outline

1. **Introduction**
   - Gene Ontology
   - Testing gene sets
2. **Multiple Testing on the GO graph**
   - Background
   - Bottom-up
   - Top-down
   - The Focus Level Method
3. **Application**
   - Netherlands Cancer Institute data
4. **Conclusion**

Outline

1. **Introduction**
   - Gene Ontology
   - Testing gene sets
2. **Multiple Testing on the GO graph**
   - Background
   - Bottom-up
   - Top-down
   - The Focus Level Method
3. **Application**
   - Netherlands Cancer Institute data
4. **Conclusion**

Gene Ontology (GO)

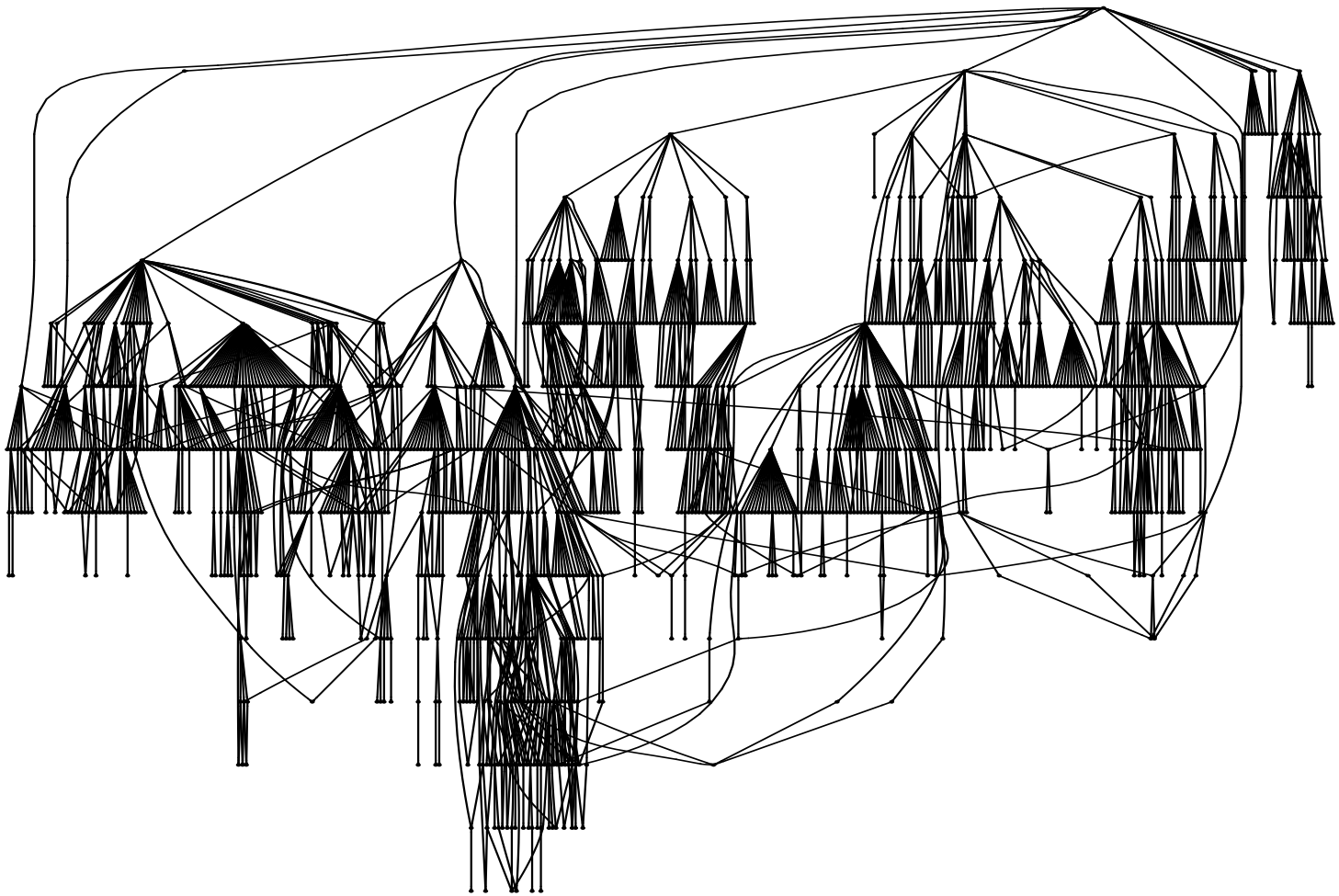
A structured vocabulary of genes

- GO is a collection of terms describing genes and gene products
- Directed relationships between terms: "is a" and "part of"
- Terms + relationships: directed acyclic graph (DAG)

GO as gene sets

- For each GO term there is a set of genes annotated to it
- Genes annotated to a GO node are automatically annotated to all ancestor nodes
- GO is a hierarchy of gene sets
- "is a" and "part of" can be seen as subset relationships

GO Molecular Function Graph



Using GO for finding differential expression?

Methods that use GO post hoc

- First find differentially expressed genes
- Next find patterns in genes called differentially expressed
- e.g. Fisher's exact test, GSEA

Using GO for finding differential expression?

Methods that use GO post hoc

- First find differentially expressed genes
- Next find patterns in genes called differentially expressed
- e.g. Fisher's exact test, GSEA

Methods that use GO directly

- Global Test (Goeman and Van Houwelingen)
- Global ANCOVA (Mansmann and Meister)

Using GO for finding differential expression?

Methods that use GO post hoc

- First find differentially expressed genes
- Next find patterns in genes called differentially expressed
- e.g. Fisher's exact test, GSEA

Methods that use GO directly

- Global Test (Goeman and Van Houwelingen)
- Global ANCOVA (Mansmann and Meister)

Review

Goeman and Bühlmann, *Bioinformatics*, 2007

Using GO for finding differential expression?

Methods that use GO post hoc

- First find differentially expressed genes
- Next find patterns in genes called differentially expressed
- e.g. Fisher's exact test, GSEA

Methods that use GO directly

- Global Test (Goeman and Van Houwelingen)
- Global ANCOVA (Mansmann and Meister)

Review

Goeman and Bühlmann, *Bioinformatics*, 2007

Global testing for GO terms

Global Test:

For which GO terms is the expression profile significantly associated with a response variable?

In a nutshell

- GO term = a set of covariates $(x_1, \ldots, x_p)$
- There is association between $(x_1, \ldots, x_p)$ and response y if part of the variance of y can be predicted using $(x_1, \ldots, x_p)$
- Generalized linear model: $\mathrm{E}(y) = \alpha + \sum_{i=1}^{p} x_i \beta_i$
- Null hypothesis of no association:

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

A locally most powerful test

Testing against a high-dimensional alternative

How to test $H_0 : \beta_1 = \ldots = \beta_p = 0$ if p may be larger than n?

A locally most powerful test

Testing against a high-dimensional alternative

How to test $H_0 : \beta_1 = \ldots = \beta_p = 0$ if p may be larger than n?

Test in a hierarchical model

Take $\beta_1, \ldots, \beta_p$ i.i.d. with $\mathrm{E}(\beta) = 0$ and $\mathrm{var}(\beta) = \tau^2$ and test

$$H_0 : \tau^2 = 0$$

Test H_0 with a score test

Resulting test is Locally Most Powerful

Optimal power on average in a p-ball with $\sum_i \beta_i^2 < \varepsilon$ among all tests of at most the same size

Unstructured testing of GO gene sets

Unstructured testing

- Globaltest type methods used to test association of GO terms with a response
- Tests performed on all (usually around 5,000) terms
- Simple adjustment for multiple testing
  - Bonferroni/Holm
  - False Discovery Rate (Benjamini and Hochberg)

Unstructured testing of GO gene sets

Unstructured testing

- Globaltest type methods used to test association of GO terms with a response
- Tests performed on all (usually around 5,000) terms
- Simple adjustment for multiple testing
  - Bonferroni/Holm
  - False Discovery Rate (Benjamini and Hochberg)

This does not use all of GO

- Part of the attractiveness of GO (structure) left unused
- Unstructured multiple testing correction = inefficient
  —Overlapping subsets → correlated test statistics
- Use logical relationships between tests to gain power

Dutch Cancer Institute—breast cancer data

Breast Cancer data

- From the Dutch Cancer Institute (NKI)
- Van 't Veer *et al.* (*Nature*, 2002)
- Van de Vijver *et al.* (*New England Journal of Medicine*, 2002)
- Rosetta custom arrays
- Normalization: Rosetta

Data summary

- 295 breast cancer patients
- Response: survival. Follow-up of 10+ years
- 4,919 genes preselected on quality criteria
- 2,216 Gene Ontology terms (biological process)

Top 16 pathways: Unstructured

	# genes	Raw p	Holm p
chromosome segregation	14	4e-09	1e-05
cell cycle	230	6e-09	1e-05
cytokinesis	7	7e-09	2e-05
microtubule cytoskeleton organiz. and biogen.	22	8e-09	2e-05
microtubule-based process	47	9e-09	2e-05
mitotic cell cycle	69	9e-09	2e-05
G2/M transition of mitotic cell cycle	4	1e-08	2e-05
DNA replication	49	1e-08	3e-05
mitosis	53	1e-08	3e-05
M phase	66	1e-08	3e-05
M phase of mitotic cell cycle	54	1e-08	3e-05
sister chromatid segregation	9	1e-08	3e-05
mitotic sister chromatid segregation	9	1e-08	3e-05
establishment of organelle localization	3	2e-08	4e-05
cytoskeleton organization and biogenesis	128	2e-08	4e-05

Outline

1. Introduction
   - Gene Ontology
   - Testing gene sets
2. **Multiple Testing on the GO graph**
   - Background
   - Bottom-up
   - Top-down
   - The Focus Level Method
3. Application
   - Netherlands Cancer Institute data
4. Conclusion

Goals of the new procedure

Structured Testing

Make a significant graph instead of a list of terms

Make use of the structure in the GO

Use the logical relationships between null hypotheses to to gain power

More biological input possible

Allow researchers to focus the method at a chosen the level of detail in the GO graph

Make strong statements about the whole graph

Strong control of the FWER

FWER or FDR?

Popular: False Discovery Rate

Controls proportion of false discoveries among discoveries

Implicit assumption

Discoveries are exchangeable:
False discoveries may be compensated by true discoveries elsewhere

FWER or FDR?

Popular: False Discovery Rate

Controls proportion of false discoveries among discoveries

Implicit assumption

Discoveries are exchangeable:
False discoveries may be compensated by true discoveries elsewhere

Subsets

Control of FDR does not imply control on subset of the discoveries

FWER or FDR?

Popular: False Discovery Rate

Controls proportion of false discoveries among discoveries

Implicit assumption

Discoveries are exchangeable:
False discoveries may be compensated by true discoveries elsewhere

Subsets

Control of FDR does not imply control on subset of the discoveries

FDR on a graph?

Difficult: discoveries are not exchangeable

Use logical relationships

Formal null hypothesis of Global Test

No gene in the set is associated with the response

Logical consequences

- If the null hypothesis for a gene set is false, then it is also false for all its supersets
- If the null hypothesis for a gene set is false, then it is also false for at least one of its subsets

Remark

This holds for actual truth falsehood

Not for actual rejection of null hypotheses

Three methods

First two are not useful in themselves, but explain principles

1: Bottom-up method

Based on Holm (1979)

2: Top-down method

Based on Marcus, Peritz, Gabriel (1976)

3: Focus Level method

Generalization of Top-down and Bottom-up

Three methods

First two are not useful in themselves, but explain principles

1: Bottom-up method

Based on Holm (1979)

2: Top-down method

Based on Marcus, Peritz, Gabriel (1976)

3: Focus Level method

Generalization of Top-down and Bottom-up

Strong control of Family-wise error

The probability of any error in the resulting graph is controlled at α
Without assumptions on the joint distribution of the test statistics

Theory: bottom-up

Bottom-up using Holm (1979)

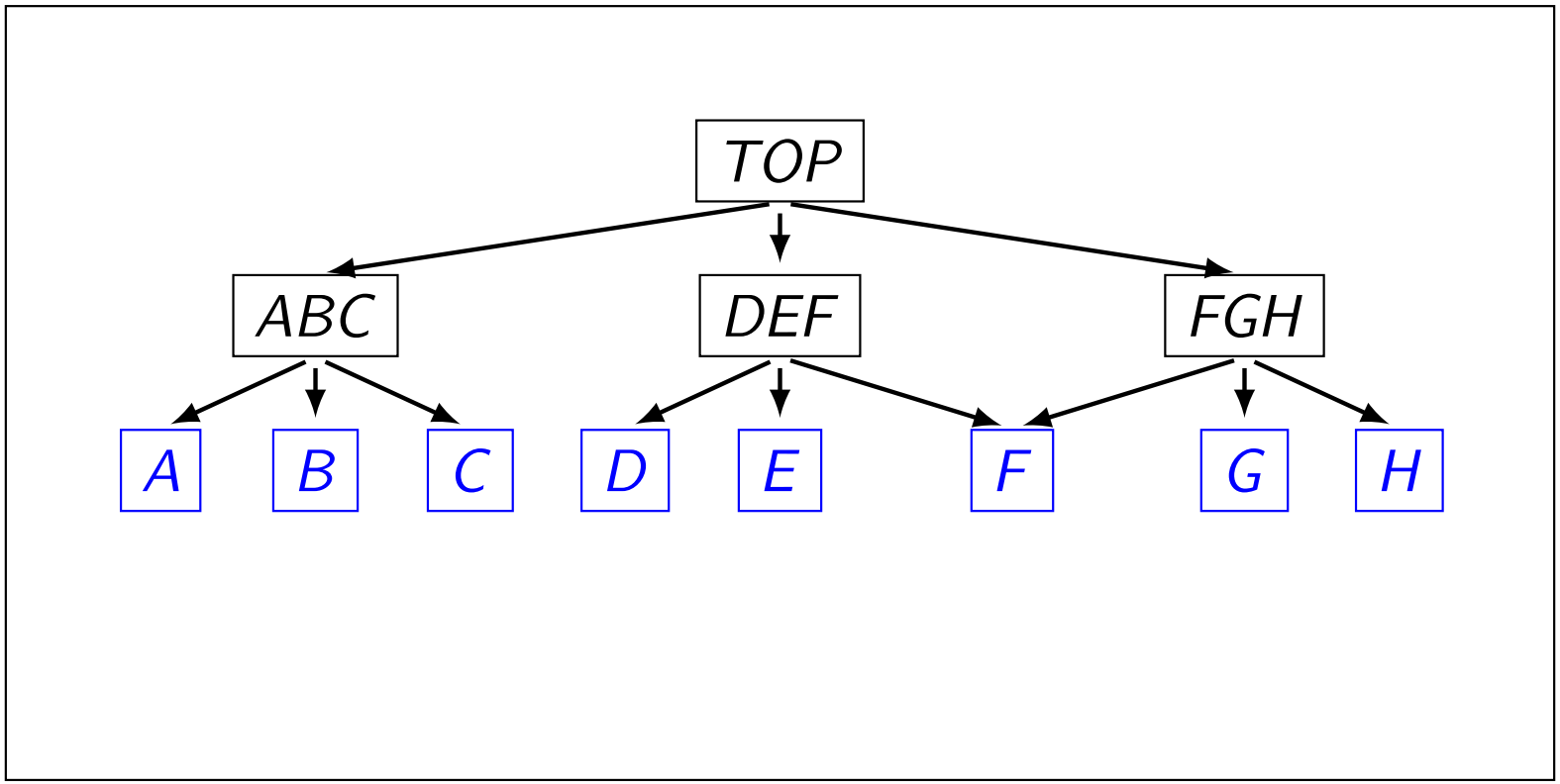
1. Take all m leave nodes and calculate raw p-values
2. Do Holm's procedure on the leave nodes
3. Declare all supersets of significant nodes significant

Property

This procedure keeps the family-wise error at level $\leq \alpha$ (strongly)

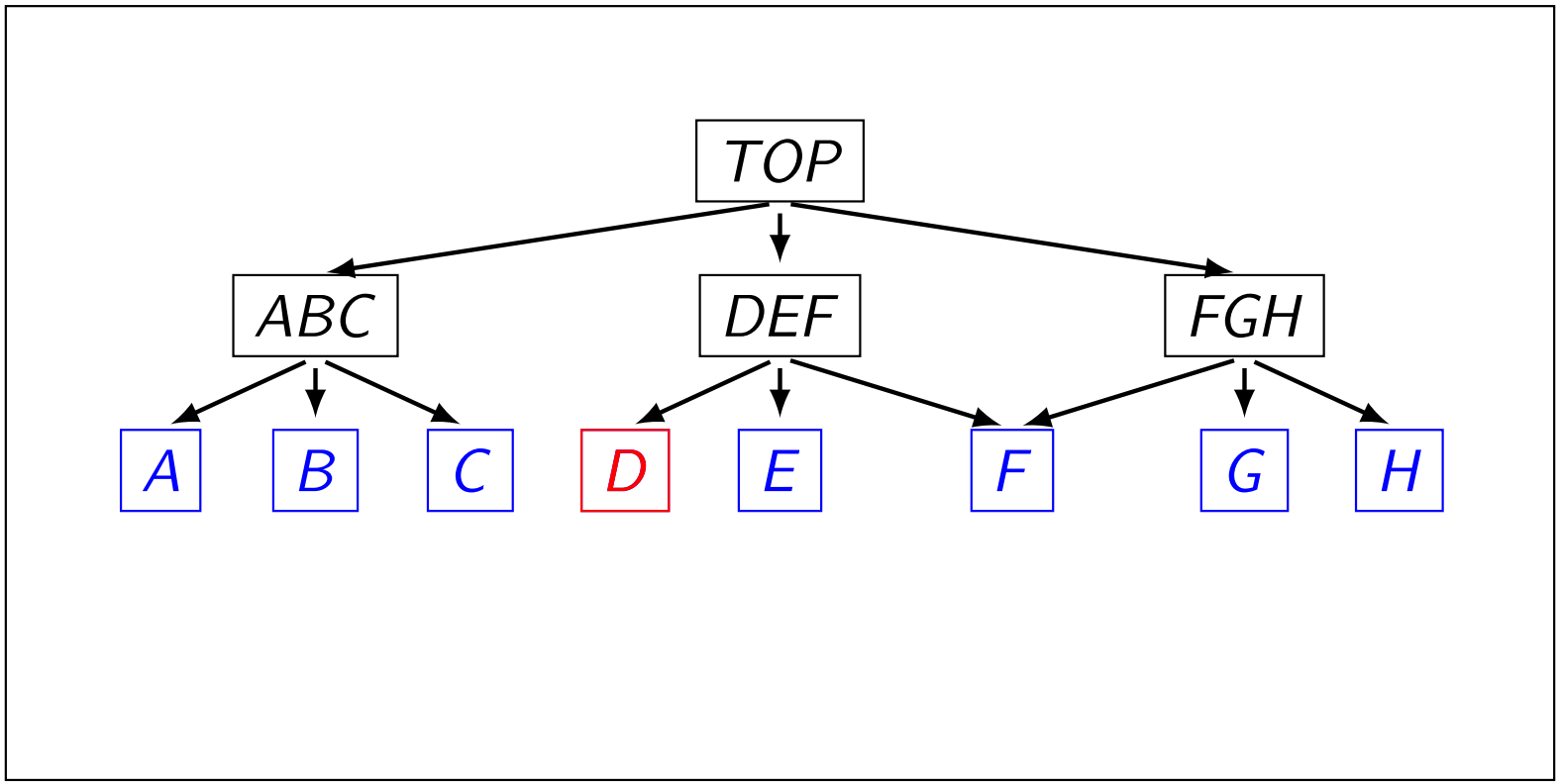
Graphical: bottom-up

Start testing only the bottom nodes



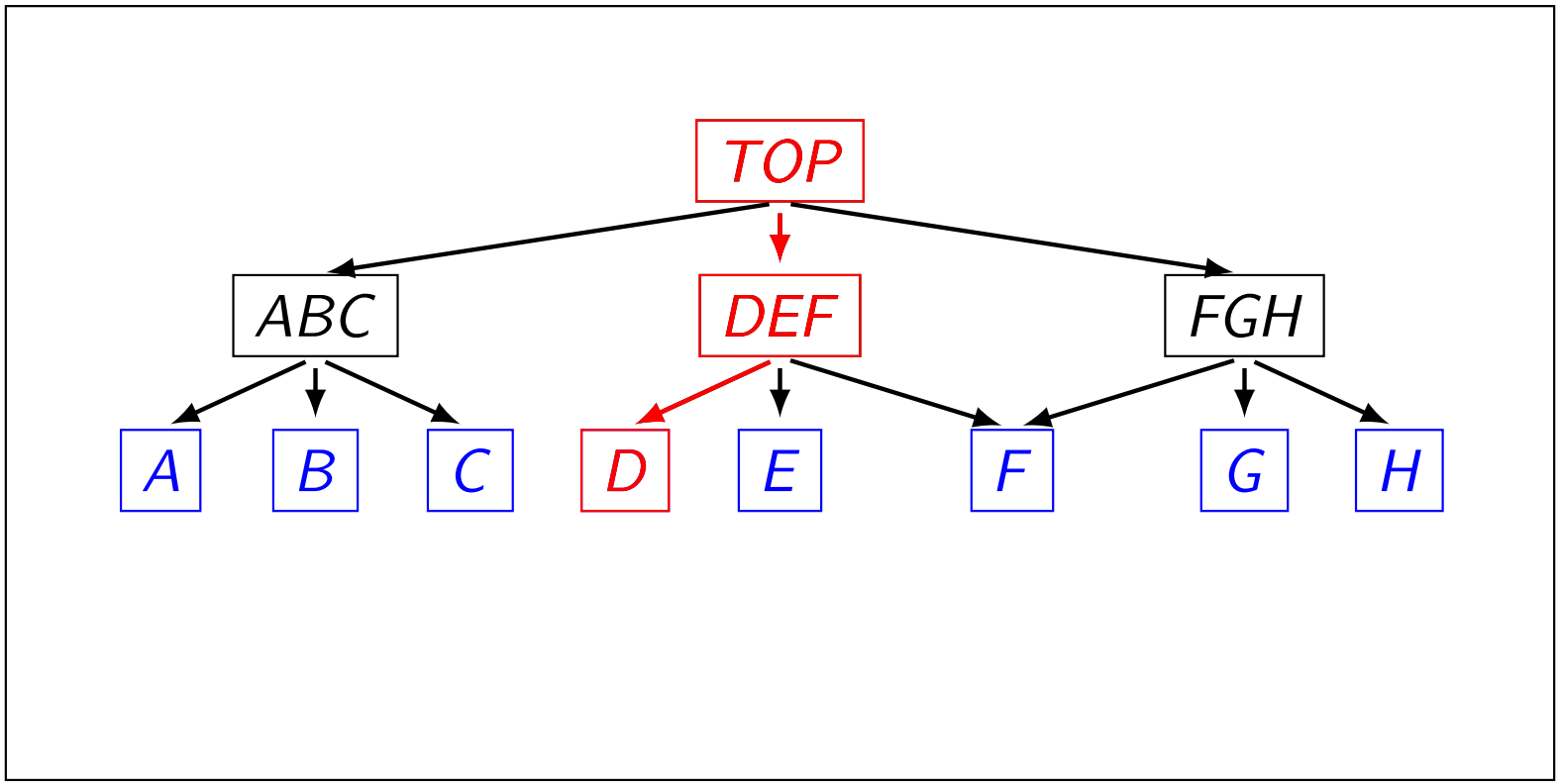
Graphical: bottom-up

Find the bottom node with smallest p-value ($\leq \alpha/8$), say D
Call D significant



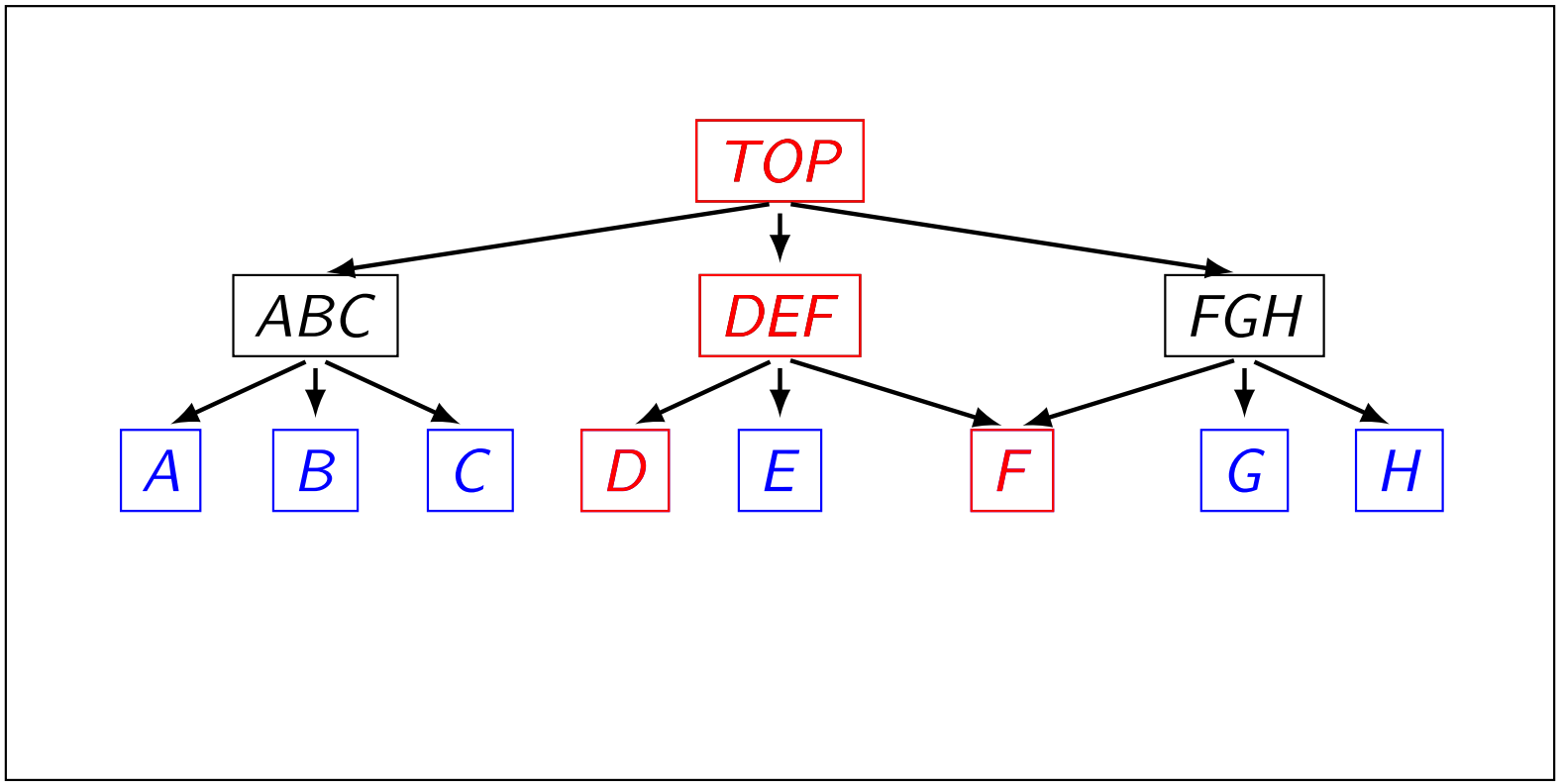
Graphical: bottom-up

Call all ancestors of D significant



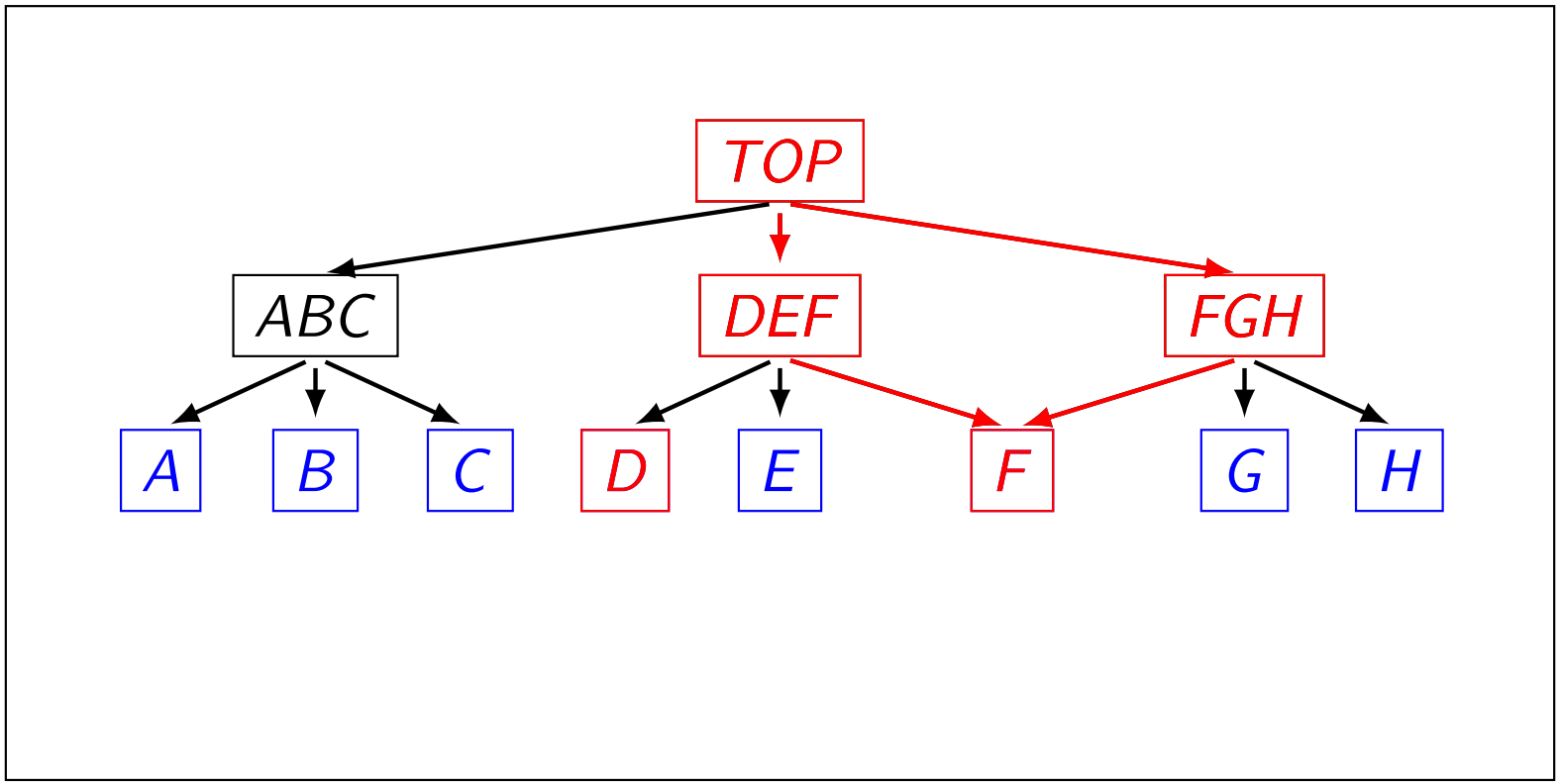
Graphical: bottom-up

Find the next bottom node with smallest p-value ($\leq \alpha/7$), say F
Call F significant



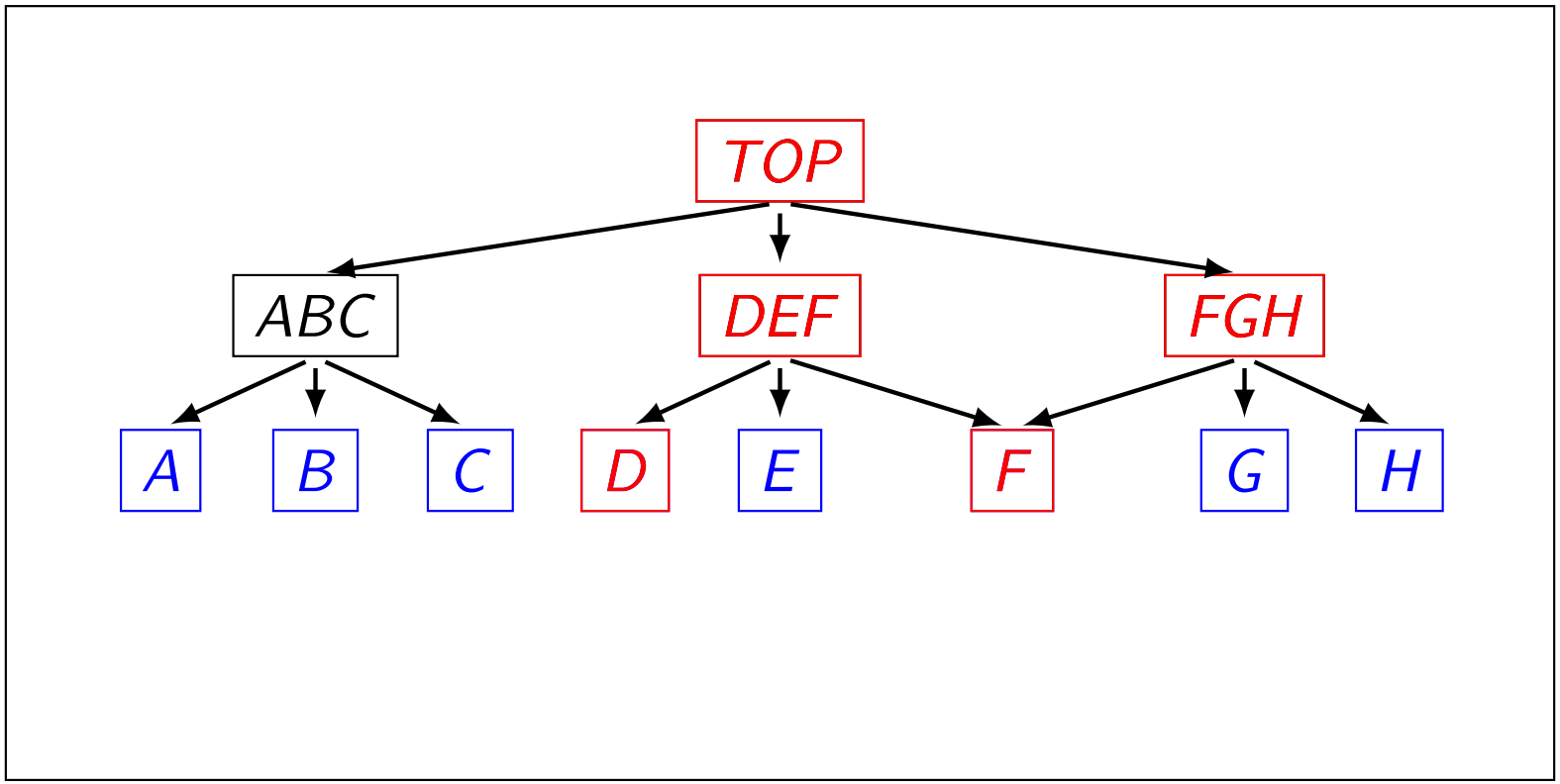
Graphical: bottom-up

Call all ancestors of F significant



Graphical: bottom-up

Find the next bottom node with smallest p-value ($\leq \alpha/6$)...



Bottom-up: Advantages and disadvantages

Advantages

- Very fast: Number of tests to be done is smaller than the number of GO nodes
- Very powerful for finding detailed GO-terms that are very significant

Disadvantages

- Many end nodes $\longrightarrow$ still severe multiple testing correction
- Significance of nodes higher up is governed by the significance of small subsets

Bottom-up: Advantages and disadvantages

Advantages

- Very fast: Number of tests to be done is smaller than the number of GO nodes
- Very powerful for finding detailed GO-terms that are very significant

Disadvantages

- Many end nodes $\longrightarrow$ still severe multiple testing correction
- Significance of nodes higher up is governed by the significance of small subsets

Interesting effect

- FWER adjusted p-values may be smaller than raw p-values

Theory: Top-down

Preliminary: expanding the graph

- Expand the graph so that it is closed w.r.t union
  $\longrightarrow$ for all A, B in the expanded graph, $A \cup B$ is also in
  $\longrightarrow$ include all possible unions of all sets in the graph

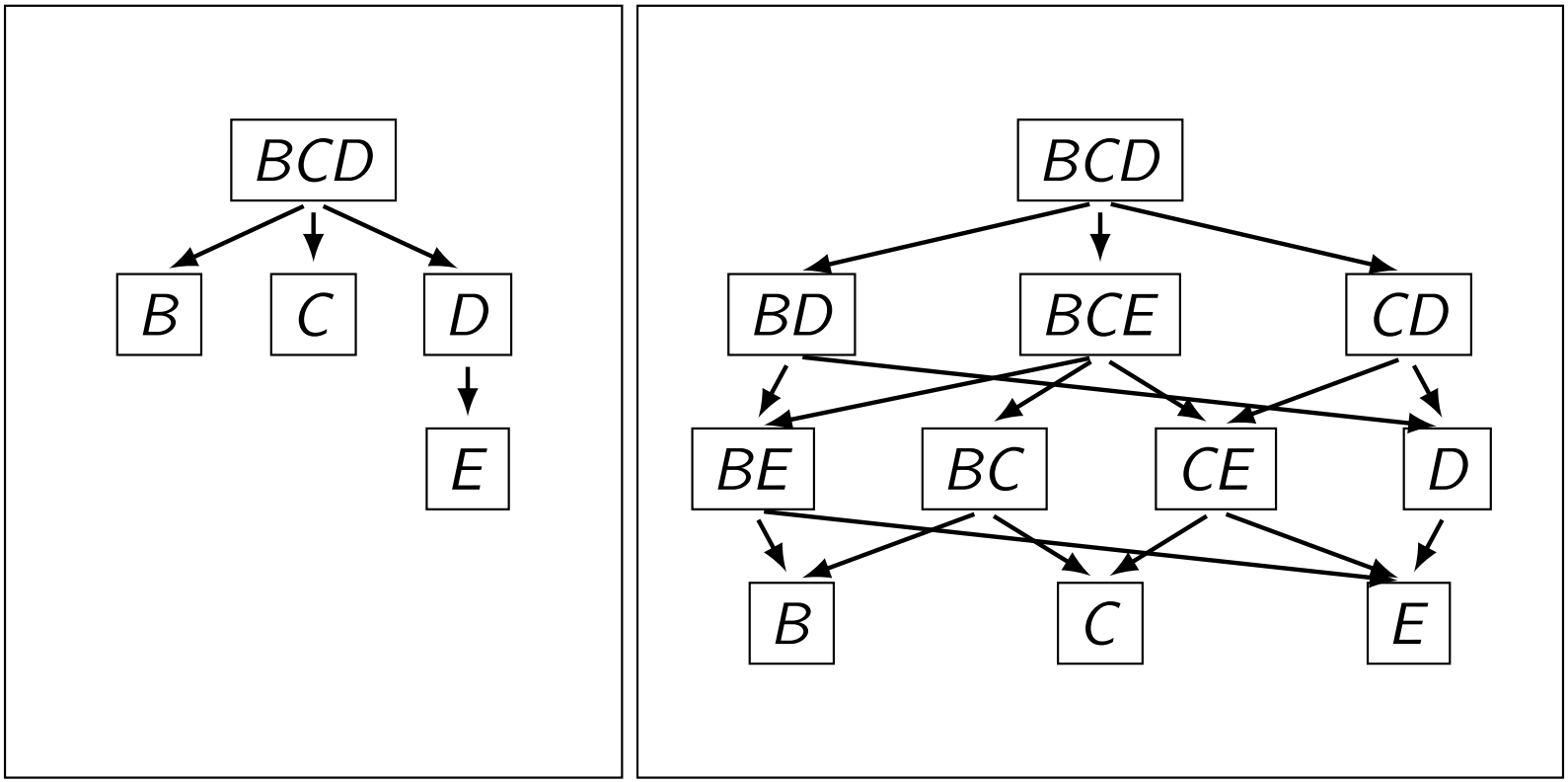
Closed testing procedure (Marcus, Peritz, Gabriel, 1976)

- Start testing the top node
- At each step, test those sets for which all supersets are declared significant
- Do all tests at level α

Property

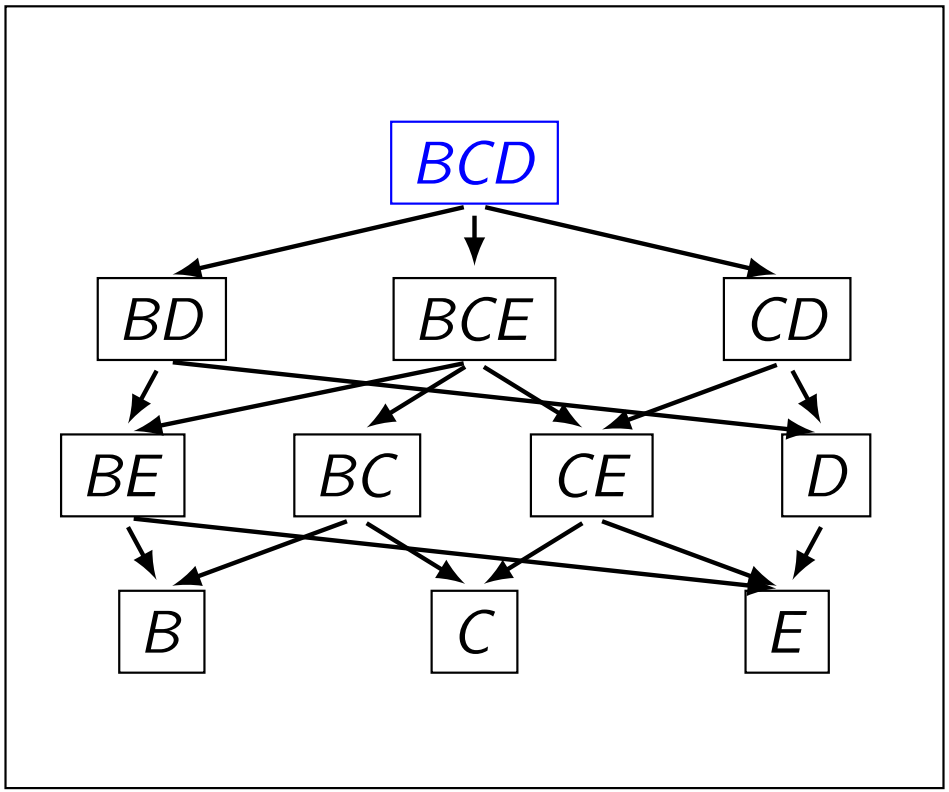
This procedure keeps the family-wise error at level $\leq \alpha$ (strongly)

Top-down procedure
Preliminary: expanding a graph



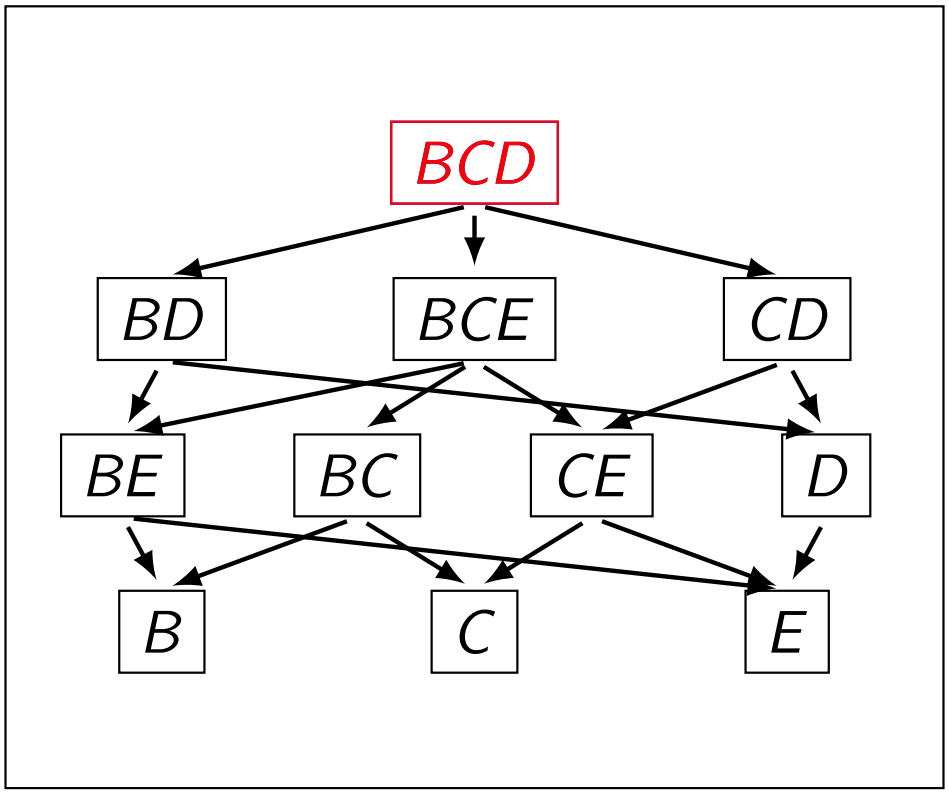
The Top-down procedure

Start testing at the top node at α



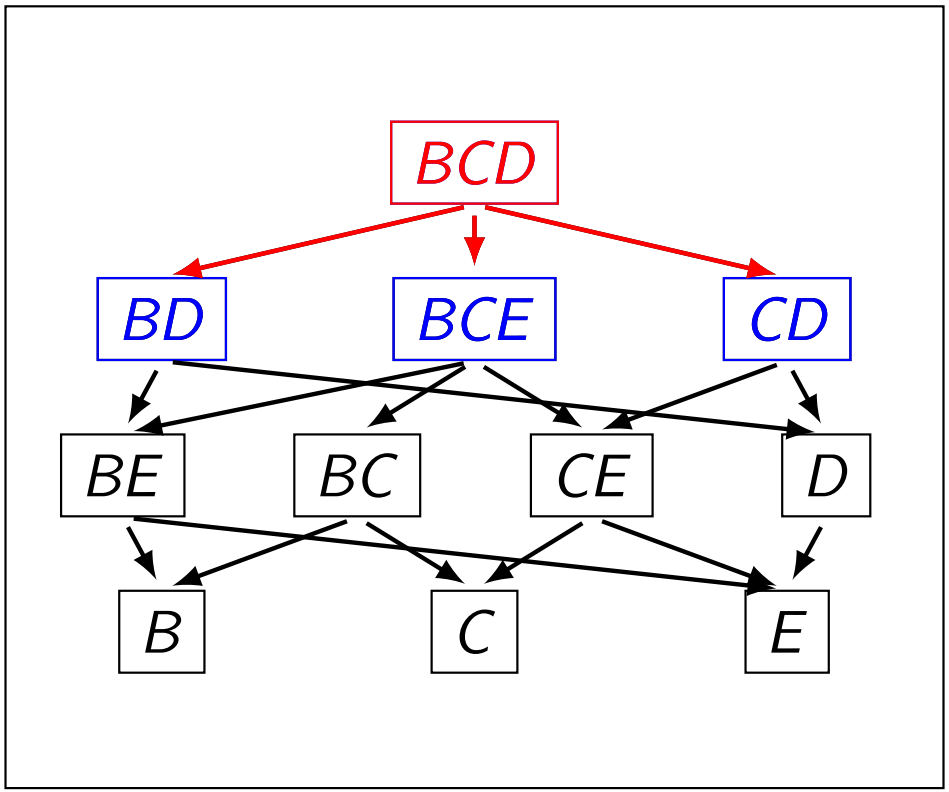
The Top-down procedure

Suppose the top node is significant



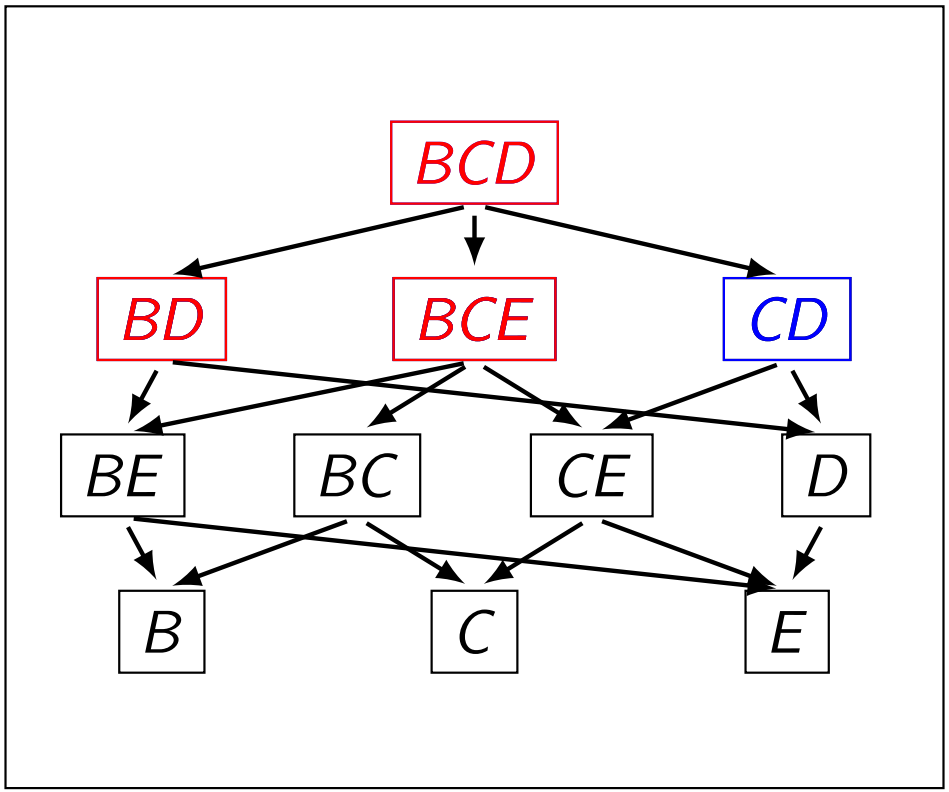
The Top-down procedure

Go on to test all child nodes of which all ancestors are significant



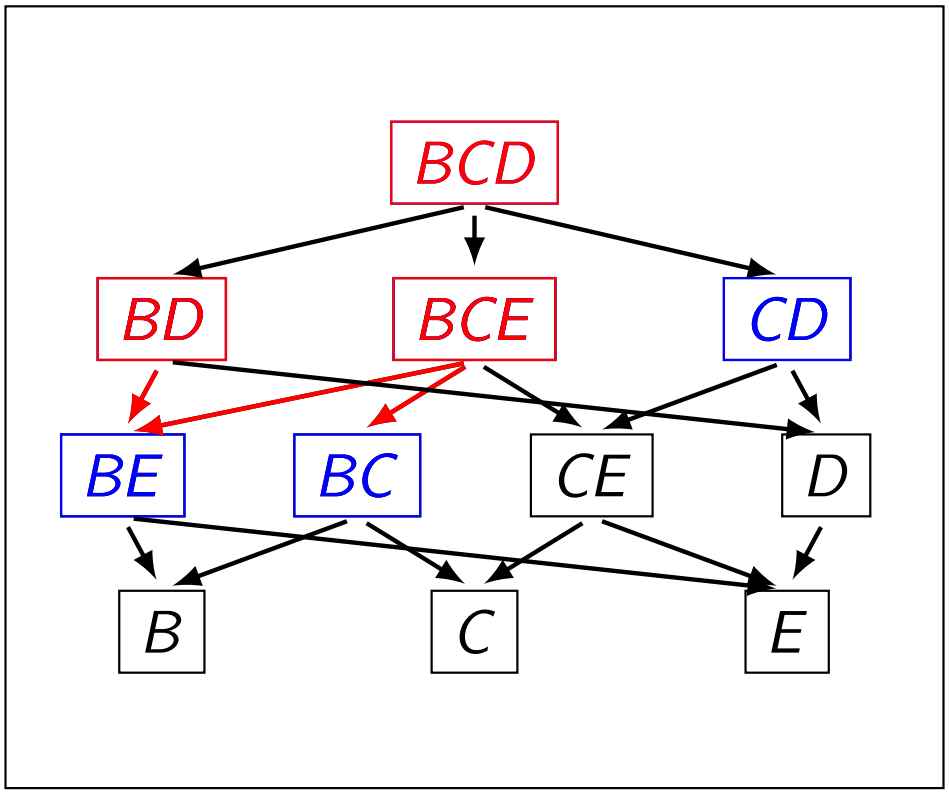
The Top-down procedure

Find those that are significant at level α



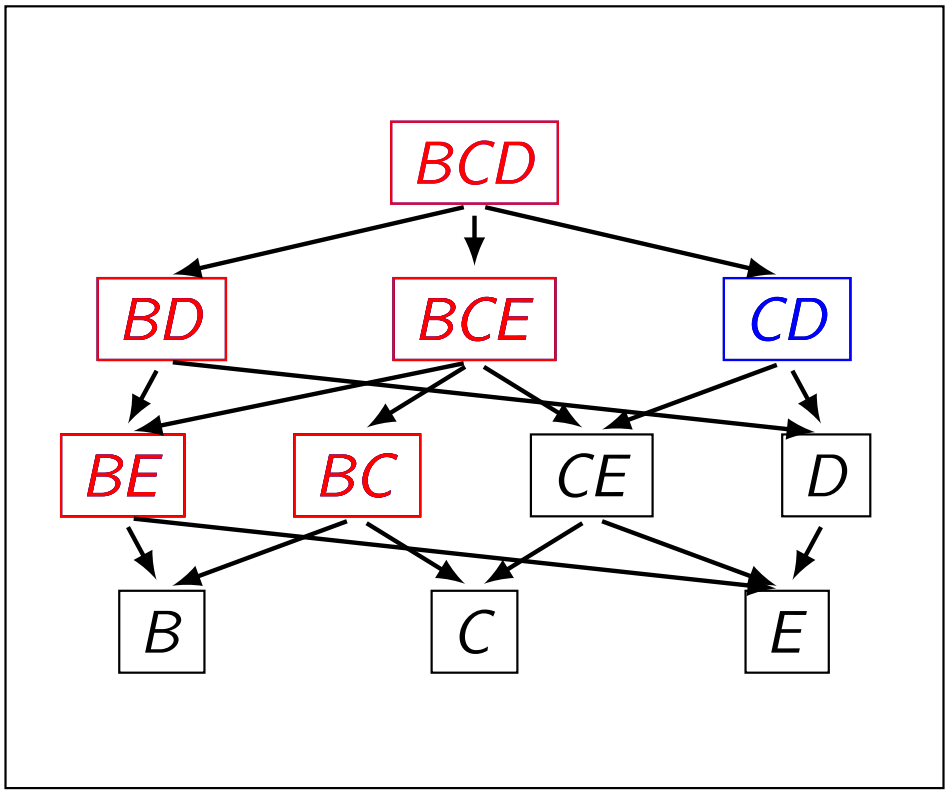
The Top-down procedure

Go on to test all child nodes of which all ancestors are significant



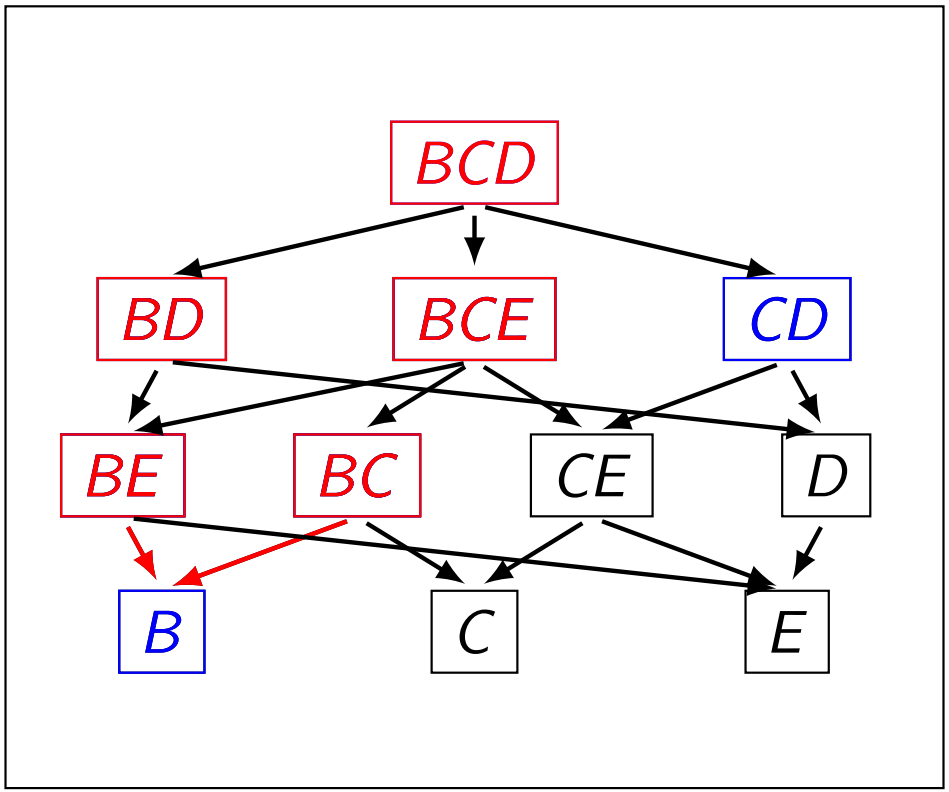
The Top-down procedure

Find those that are significant at level α



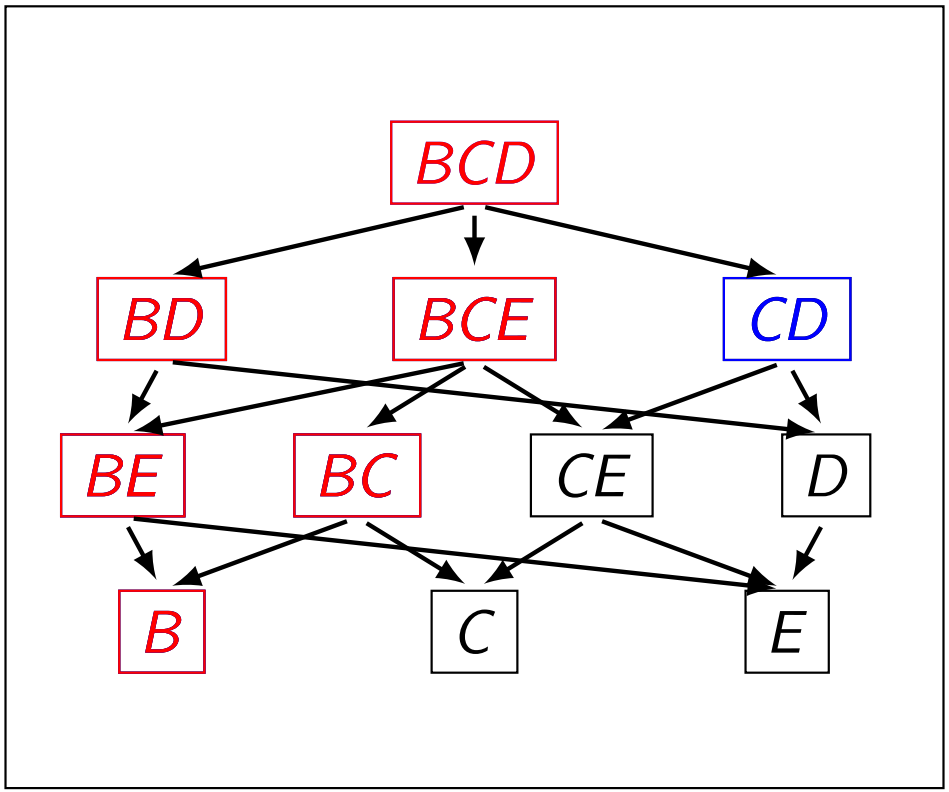
The Top-down procedure

Go on to test all child nodes of which all ancestors are significant



The Top-down procedure

Find those that are significant at level α



Top down: Advantages and disadvantages

Advantages

- Very light multiple testing correction around the top of the graph
- Very good at finding a small effect of many genes in a non-specific GO term

Disadvantages

- Computationally prohibitive: Number of tests to be done is in the order of magnitude of $2^{(\#\text{nodes})}$
- Significance of nodes lower down is governed by the significance of large supersets

Comparing Bottom-up and Top-down

Power focus

- Bottom-up:
  —good at detecting local effects at the bottom of GO
  —bad at detecting global effects at the top of GO
- Top-down:
  —good at detecting global effects at the top of GO
  —bad at detecting local effects at the bottom of GO

Comparing Bottom-up and Top-down

Power focus

- Bottom-up:
  —good at detecting local effects at the bottom of GO
  —bad at detecting global effects at the top of GO
- Top-down:
  —good at detecting global effects at the top of GO
  —bad at detecting local effects at the bottom of GO

Where are the interesting effects?

Somewhere in the middle

Starting in the middle

Choose a focus level

A subset of the nodes

Splitting

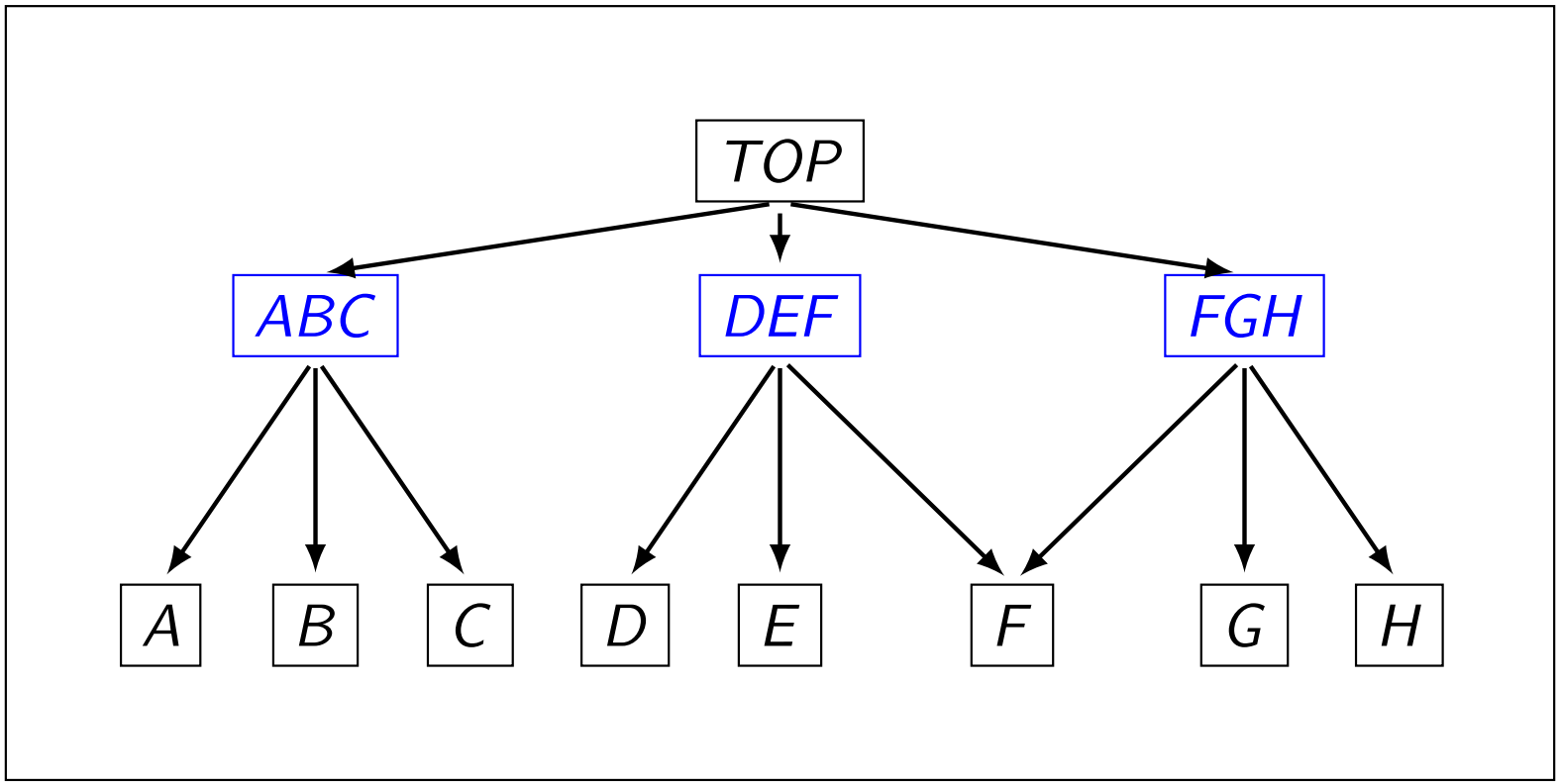
Split the graph in subgraphs at the focus level
Expand each subgraph as in the closed testing procedure

Algorithm

- Start with the focus level set $\mathcal{H}$ of m hypotheses
- Test at level α/m
- If a test is significant, add subsets to $\mathcal{H}$ as in closed testing
- If any subgraph is completely significant, reduce m by 1.
- Always call all supersets of significant hypotheses significant

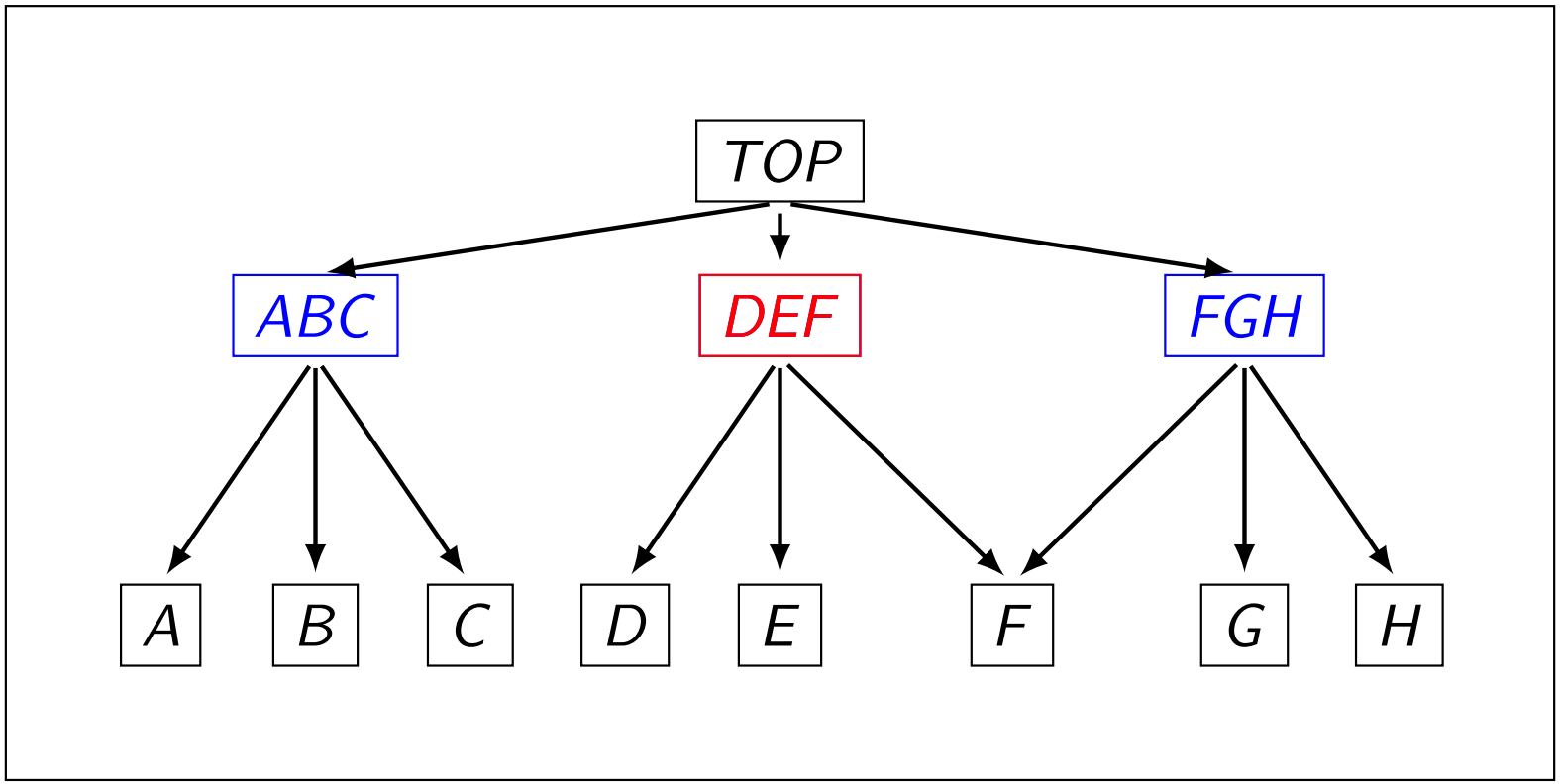
The focus level procedure

Start somewhere in the middle: choose a "focus level"



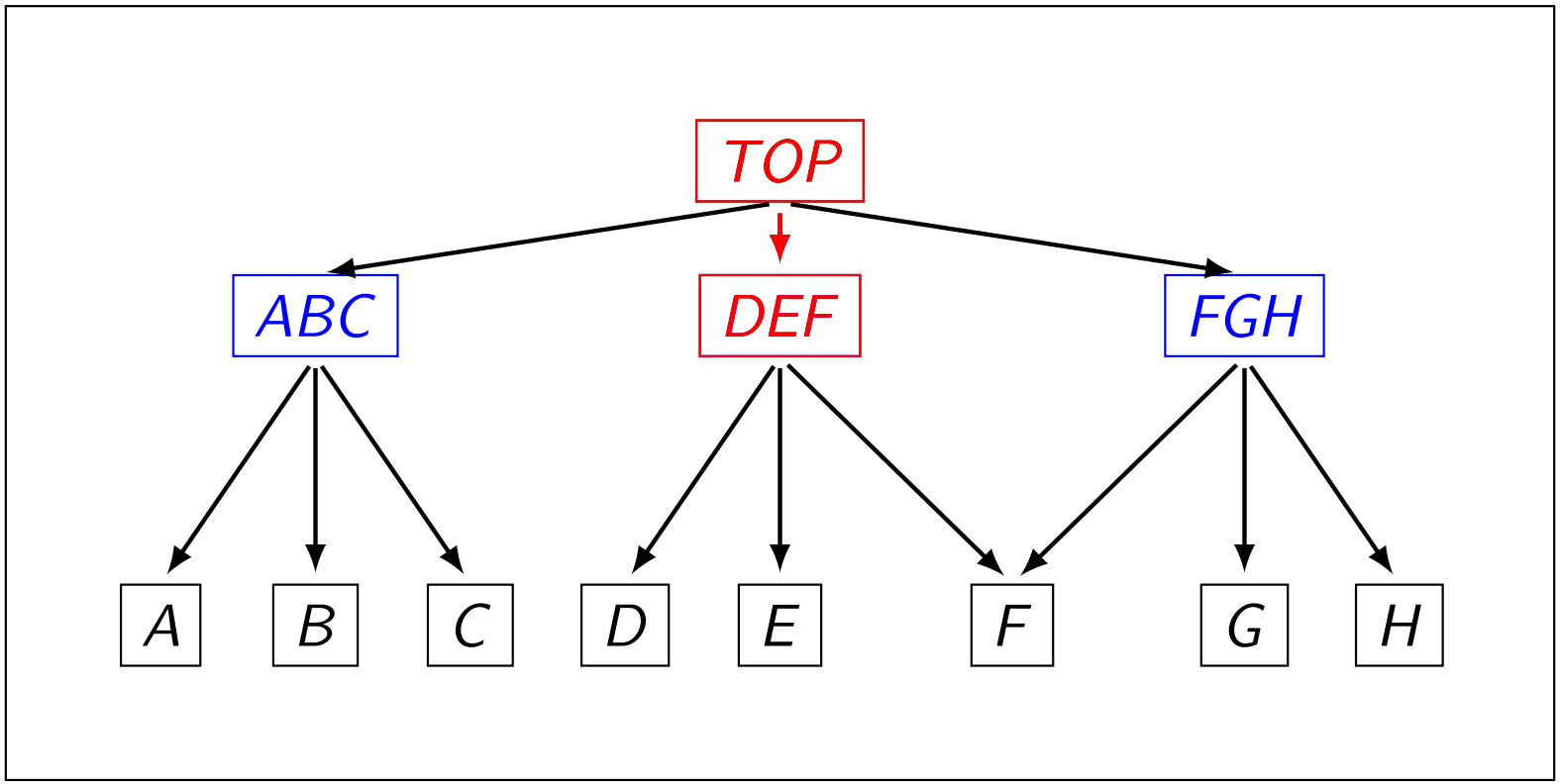
The focus level procedure

Find the focus level node with smallest p-value ($\leq \alpha/3$), say *DEF*
Call *DEF* significant



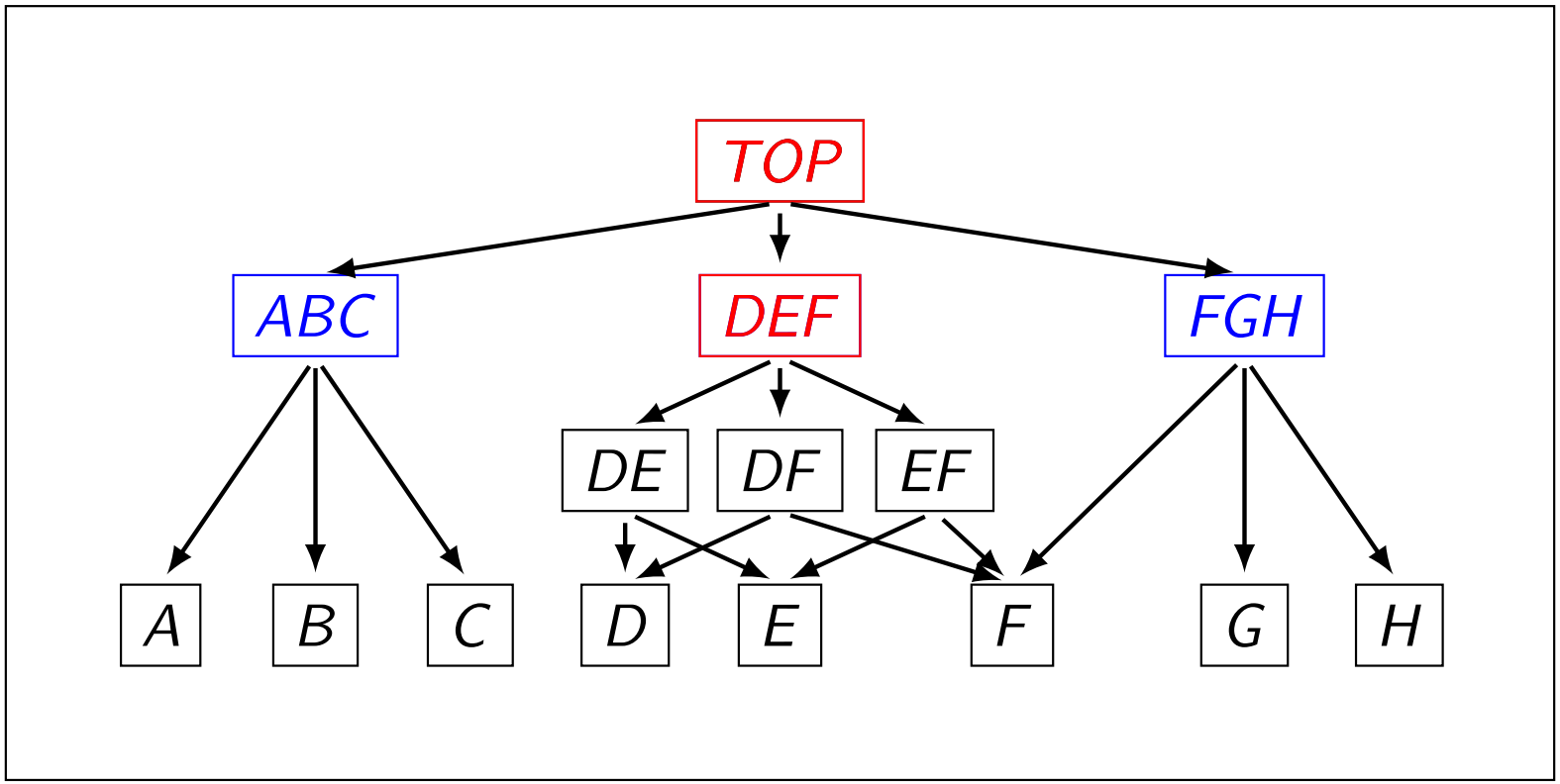
The focus level procedure

Use the bottom-up procedure to propagate significance upwards



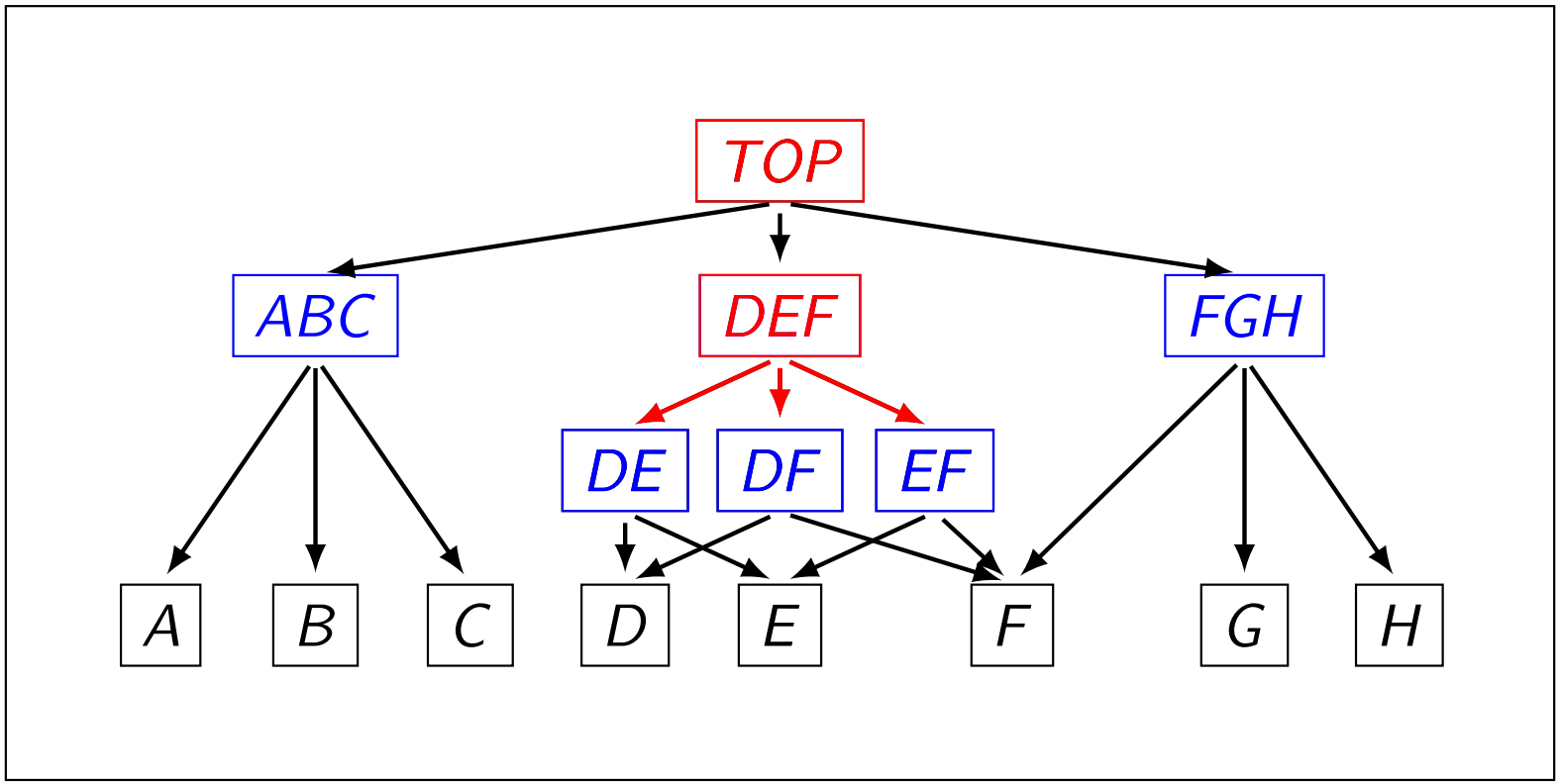
The focus level procedure

Expand the graph below *DEF*



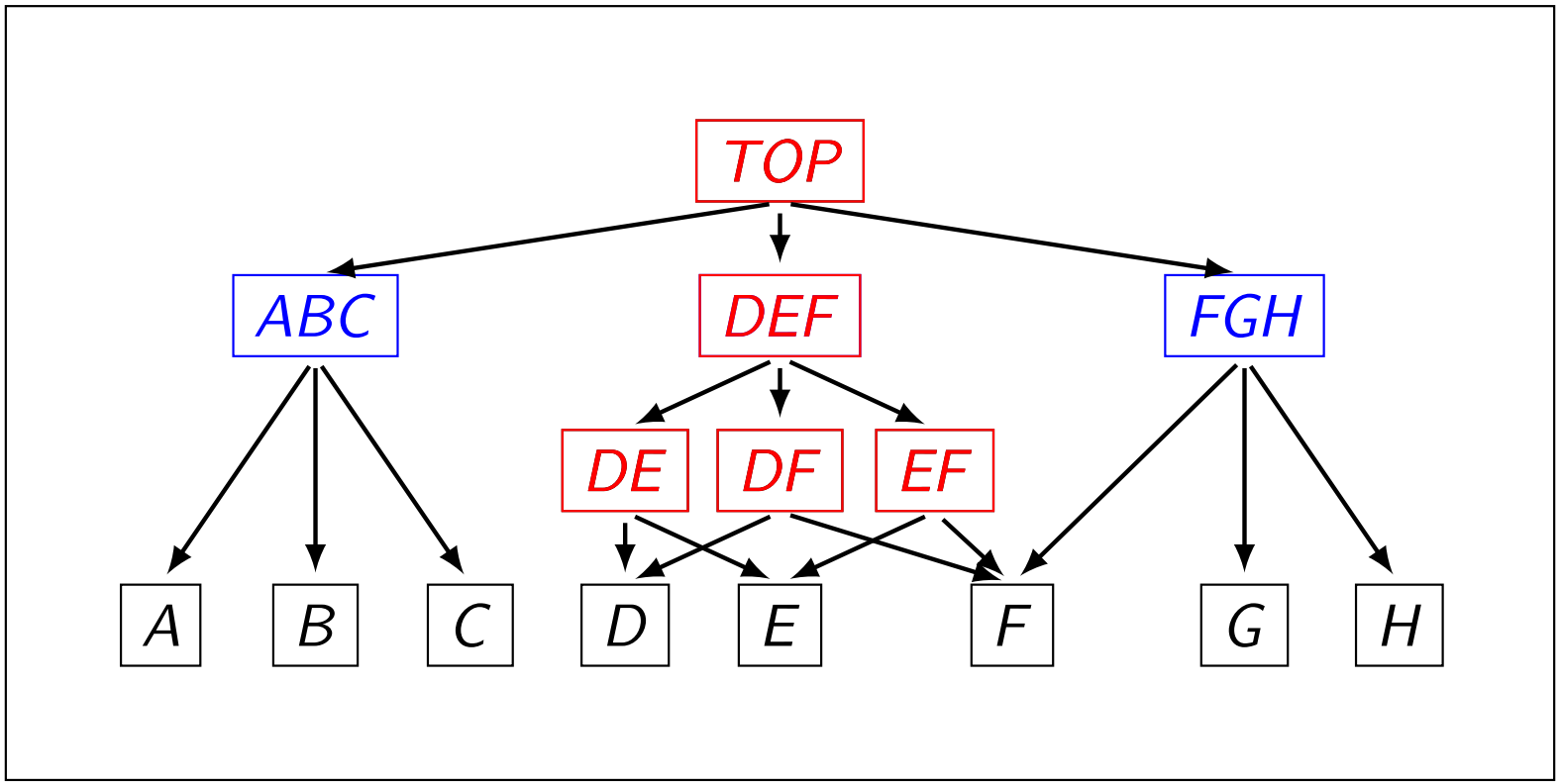
The focus level procedure

Use the top-down procedure at $\alpha/3$ to propagate significance downward



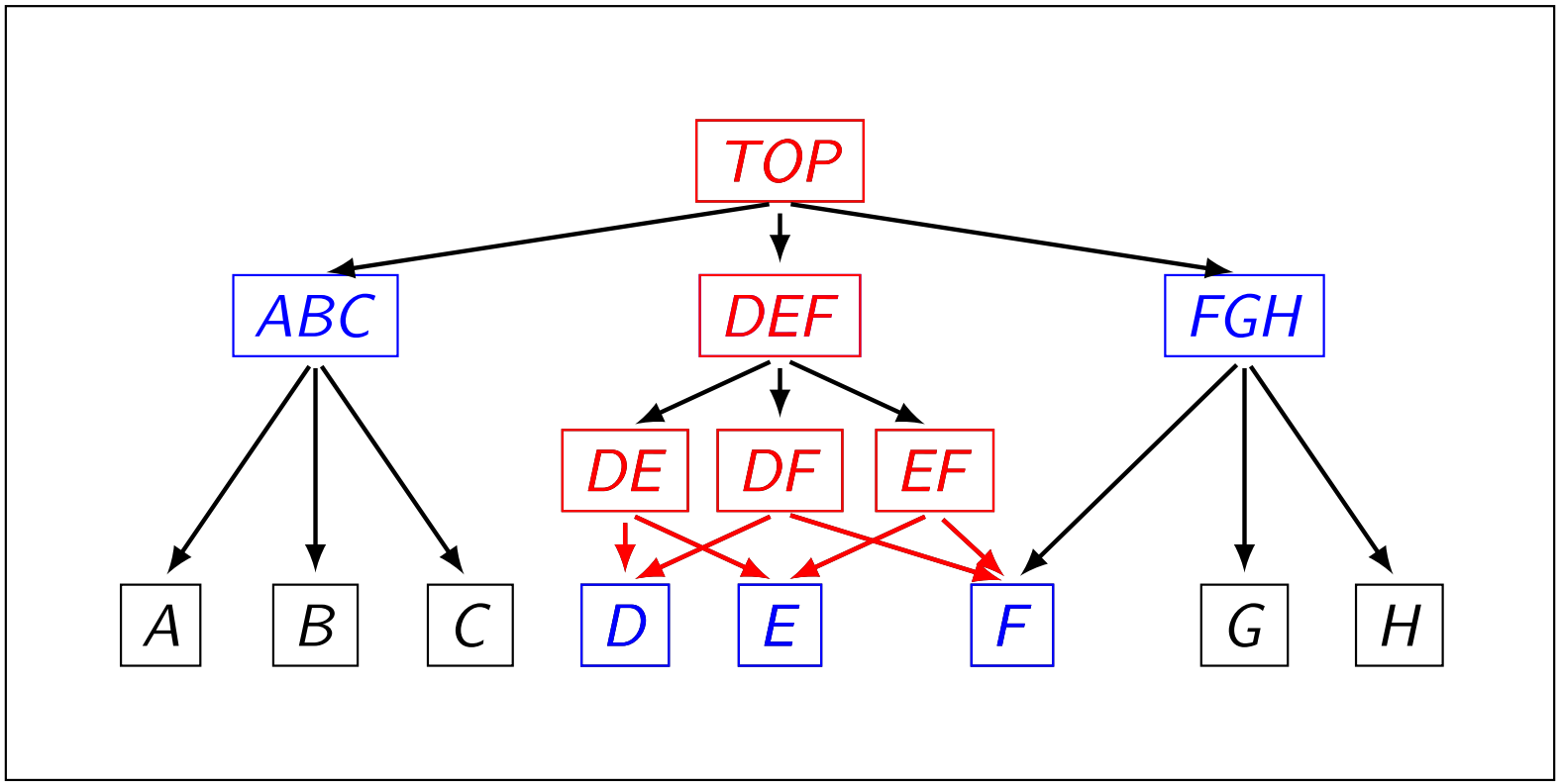
The focus level procedure

Use the top-down procedure at $\alpha/3$ to propagate significance downward



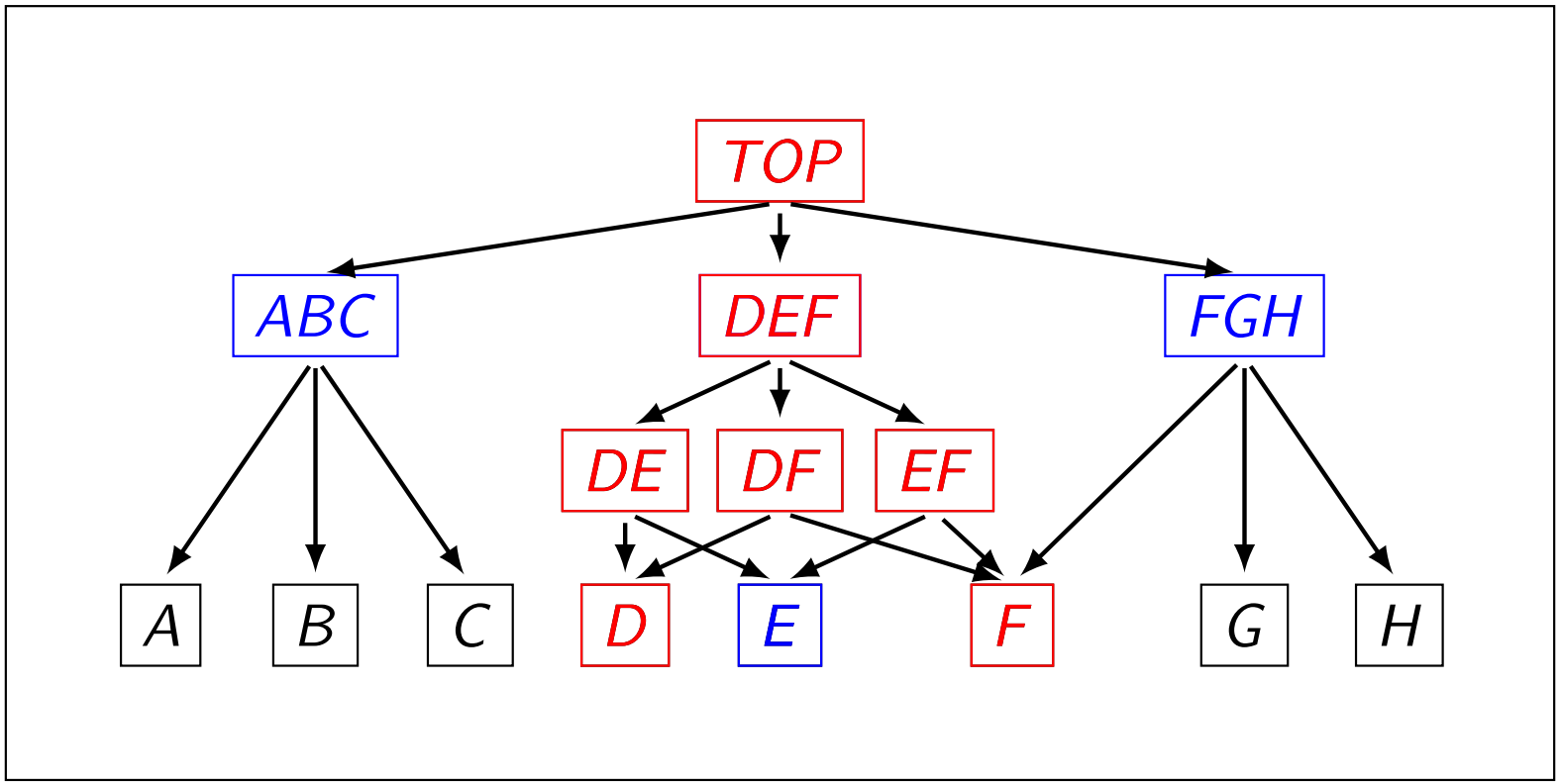
The focus level procedure

Use the top-down procedure at $\alpha/3$ to propagate significance downward



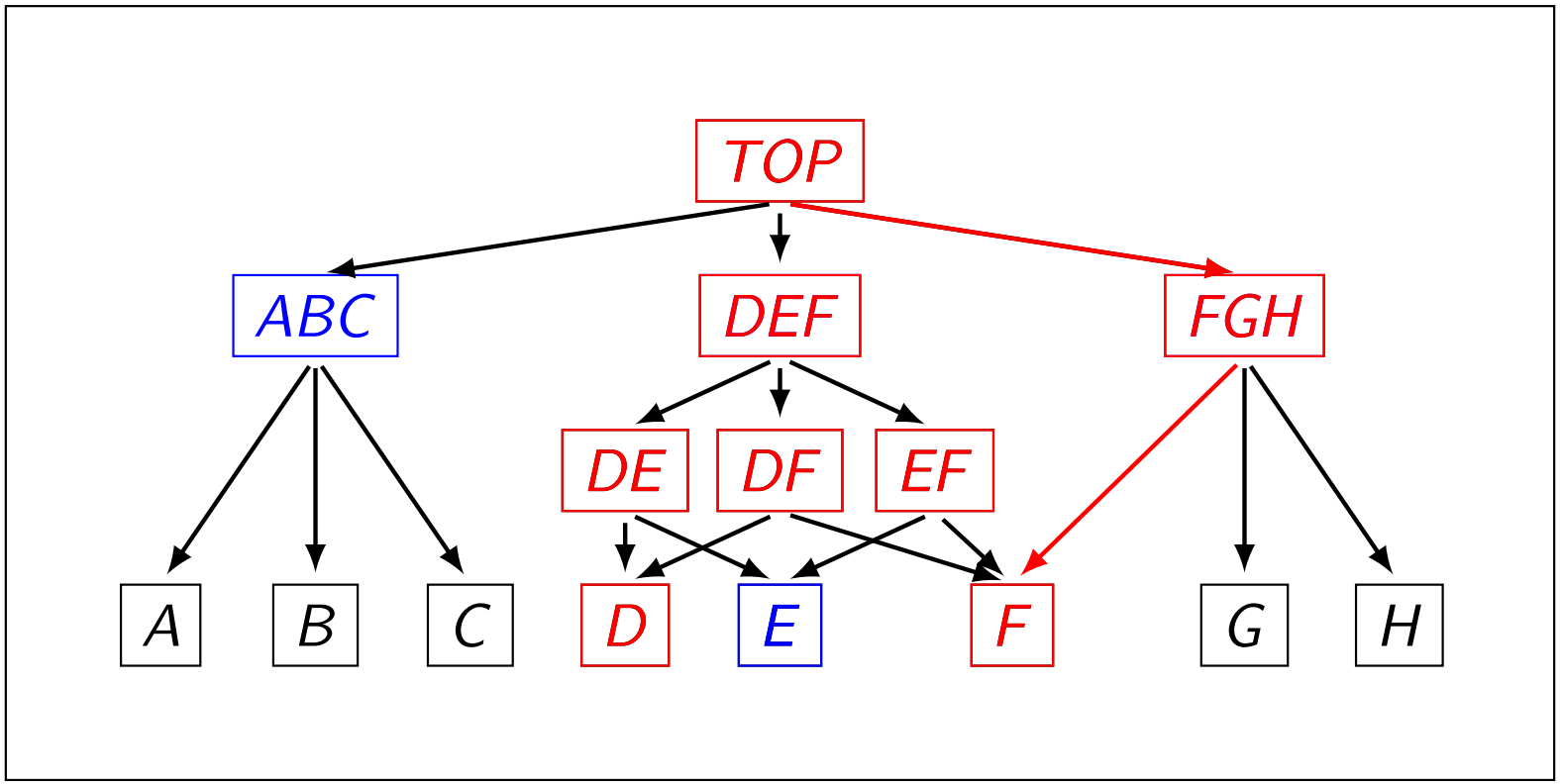
The focus level procedure

Use the top-down procedure at $\alpha/3$ to propagate significance downward



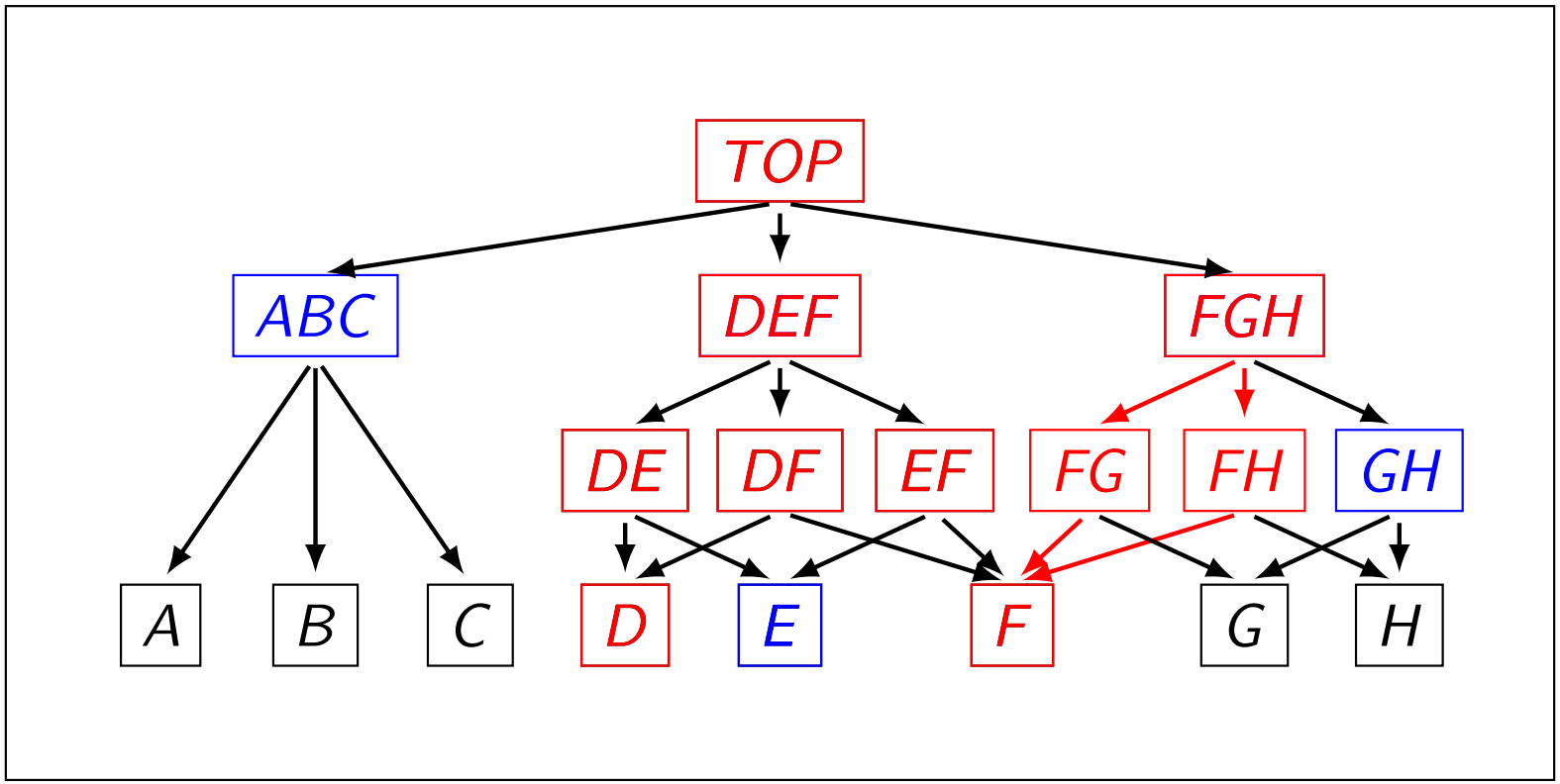
The focus level procedure

Call all ancestors of significant nodes significant



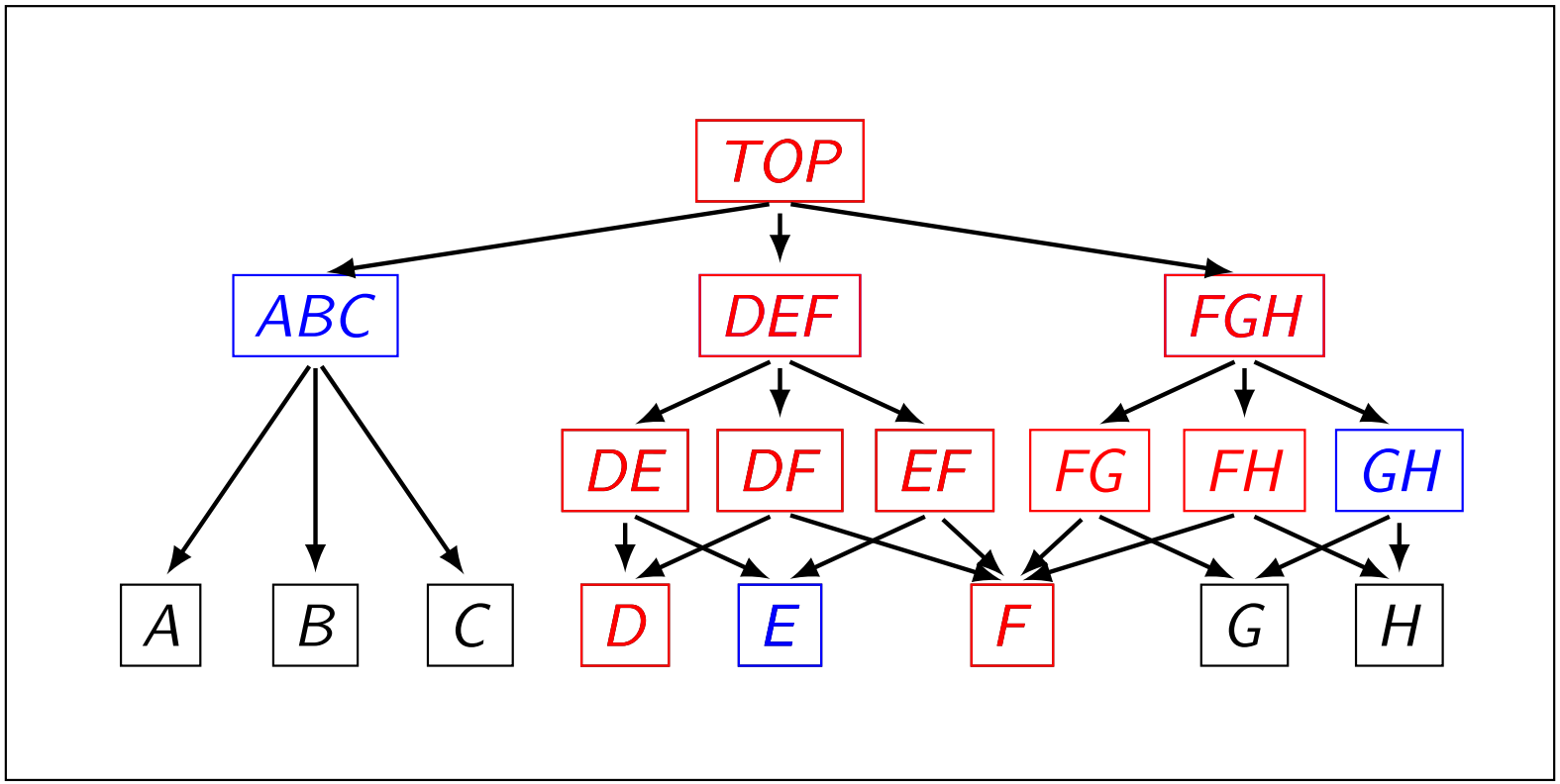
The focus level procedure

Expand subgraphs when necessary



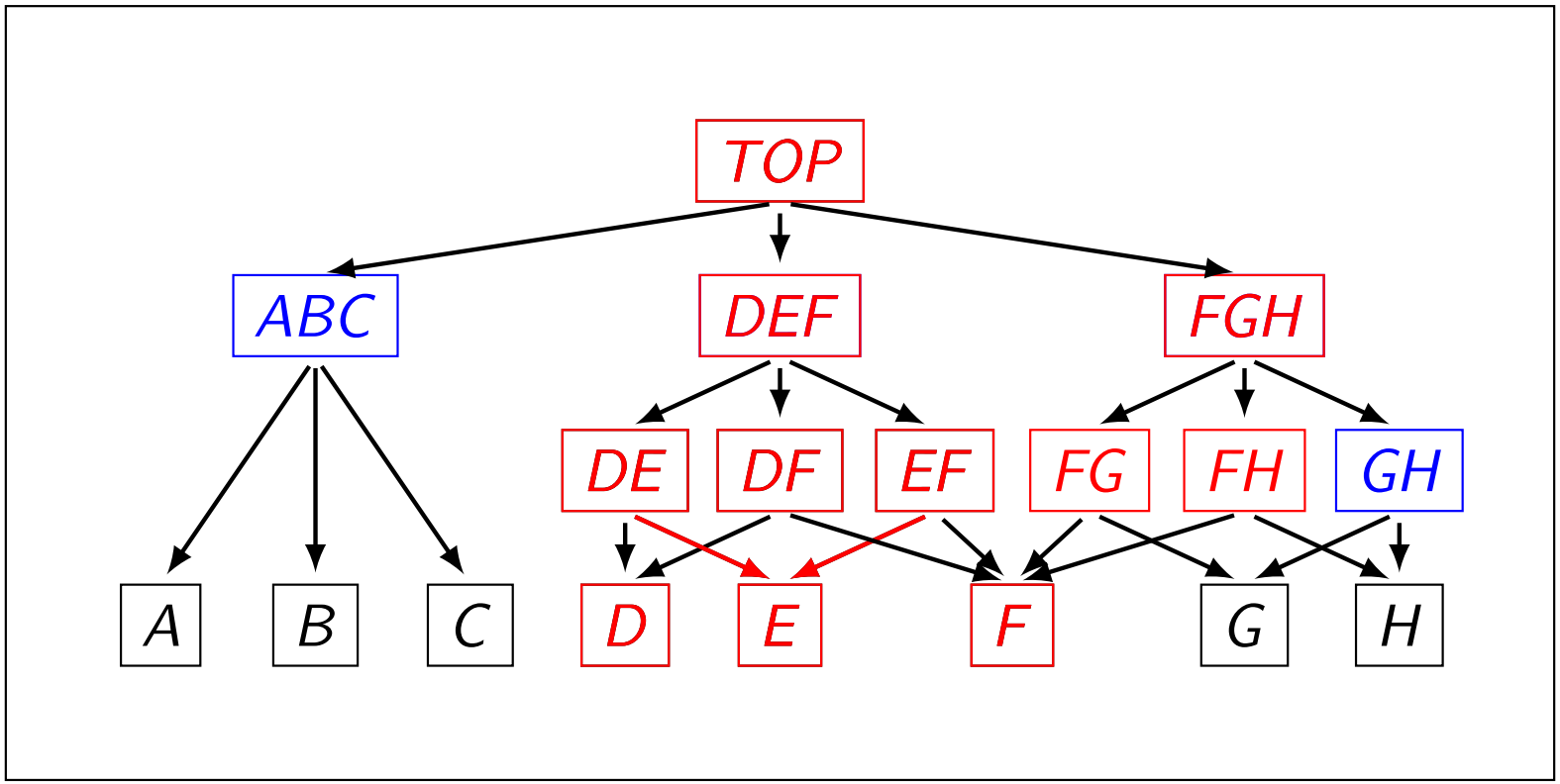
The focus level procedure

Go on until no more significant sets can be found



The focus level procedure

Any subgraph completely significant
$\longrightarrow$ recalibrate significance criterion ($\alpha/2$)



This procedure keeps the FWER

FWER control

This procedure strongly controls the FWER

Proof

Combine the proofs of Holm and Closed Testing

This procedure keeps the FWER

FWER control

This procedure strongly controls the FWER

Proof

Combine the proofs of Holm and Closed Testing

Remark 1

FWER-corrected p-values can be smaller than the raw p-values, but not for leave nodes of the significant subgraph

Remark 2

For focus level = top node this is the top down method
For focus level = leave nodes this is the bottom up method

Starting in the middle

Choose a focus level

- In general: any subset of the nodes
- Good choice: a "level" in the graph:
  - No focus node has focus offspring
  - All other nodes have focus ancestor or offspring

Choose the level of detail that interests you

Depends on the research goal

Procedure has most power to detect effects at the focus level

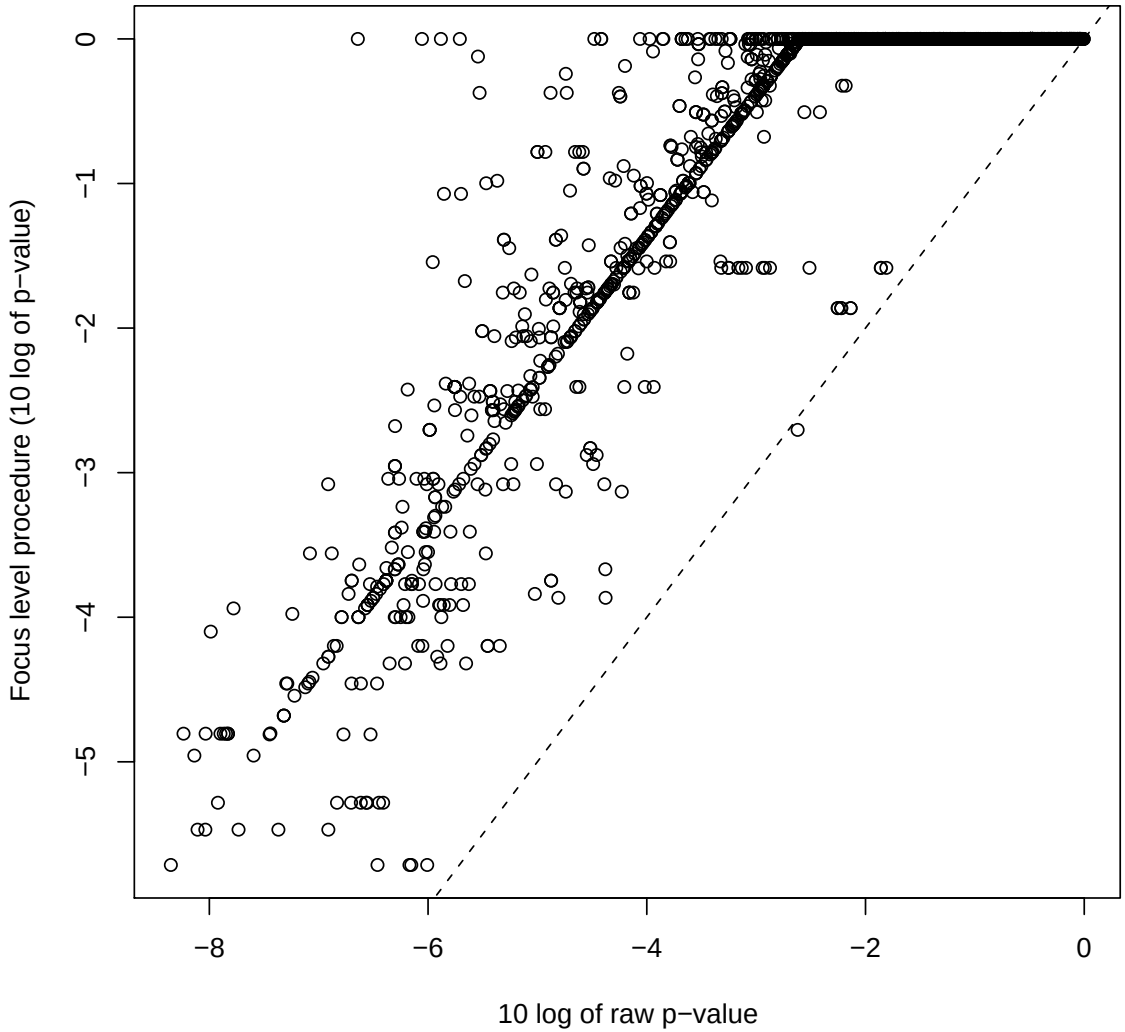
Not too general

Too general focus level terms are computationally too demanding

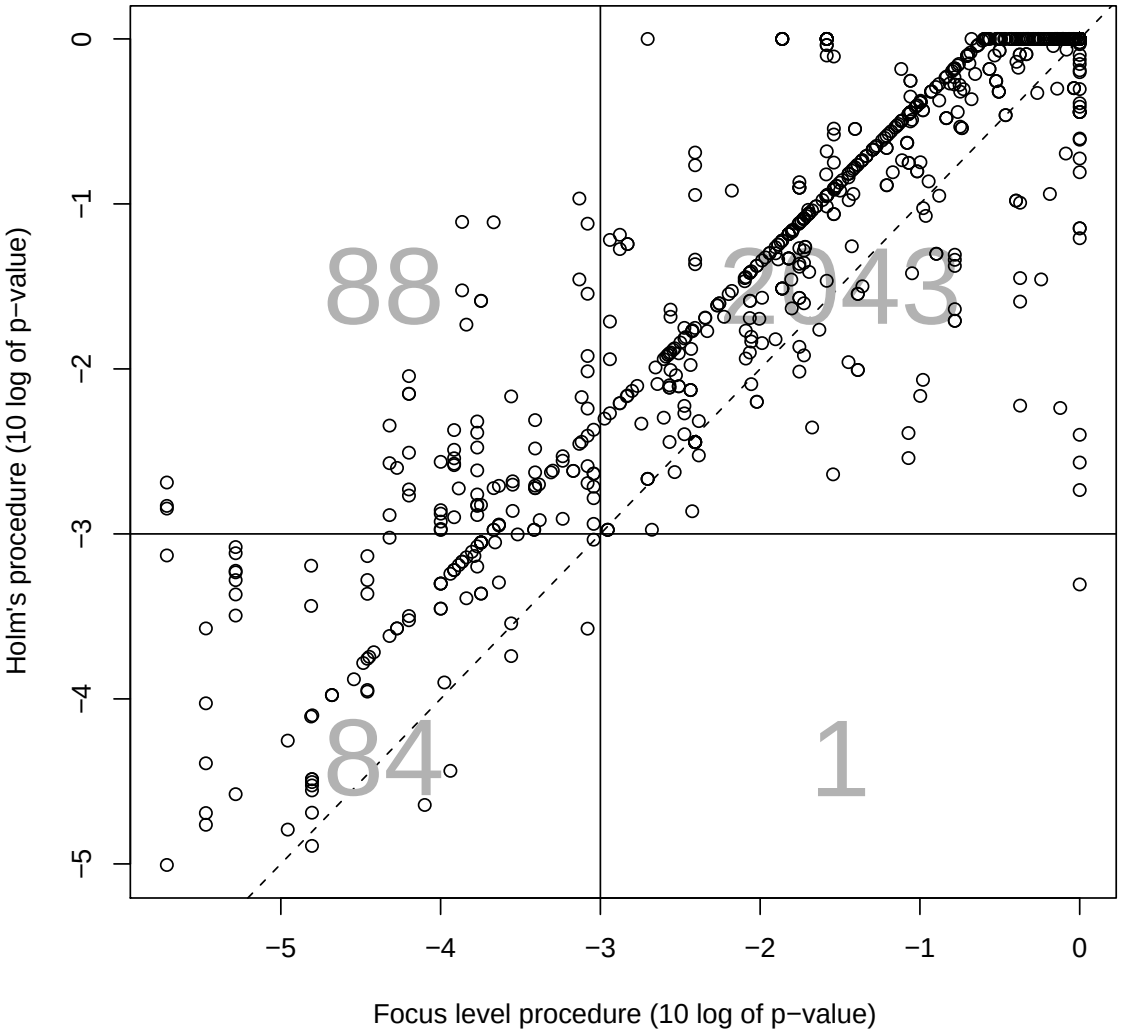
Outline

1. Introduction
   - Gene Ontology
   - Testing gene sets
2. Multiple Testing on the GO graph
   - Background
   - Bottom-up
   - Top-down
   - The Focus Level Method
3. **Application**
   - Netherlands Cancer Institute data
4. Conclusion

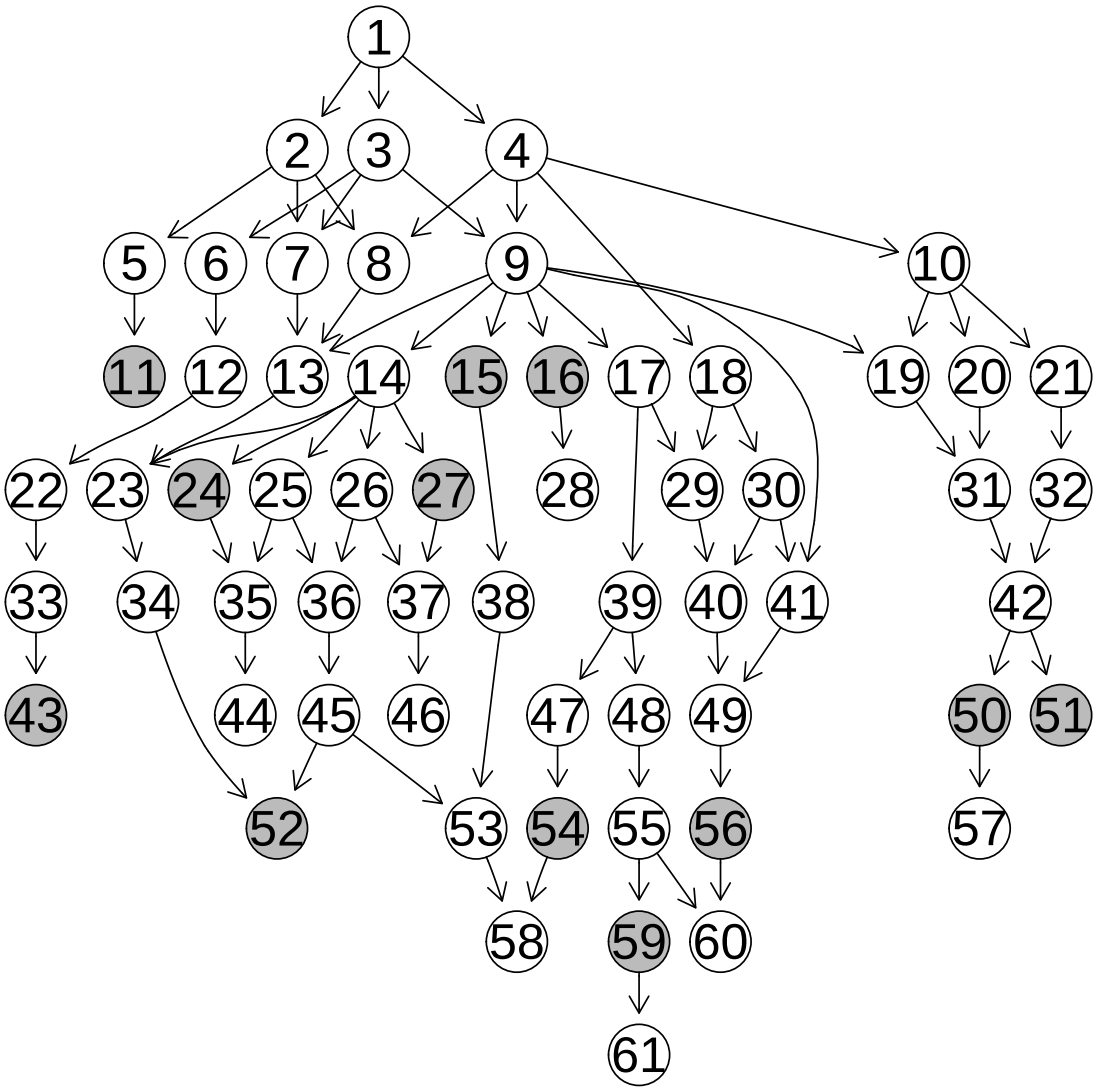
Raw versus adjusted p-values



Focus level adjusted p-values versus Holm



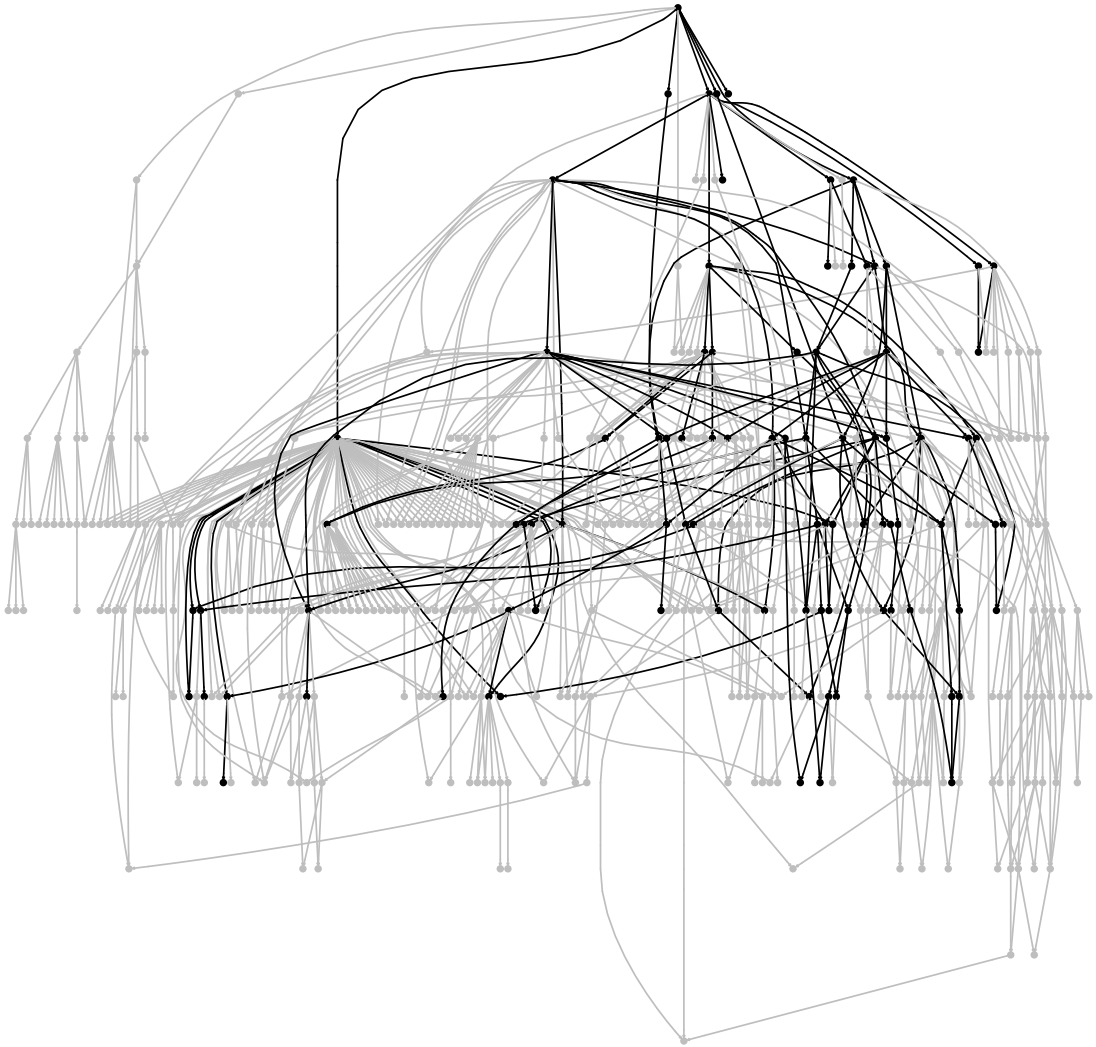
The BP significant subgraph at $\alpha = 10^{-4}$



Legend

1	biological process		and nucleic acid metabolism
2	regulation of biological process	32	biopolymer metabolism
3	cellular process	33	second-messenger-mediated signaling
4	physiological process	34	regulation of progression through cell cycle
5	regulation of enzyme activity	35	interphase of mitotic cell cycle
6	cell communication	36	M phase of mitotic cell cycle
7	regulation of cellular process	37	M phase of meiotic cell cycle
8	regulation of physiological process	38	sister chromatid segregation
9	cellular physiological process	39	organelle organization and biogenesis
10	metabolism	40	establishment of cellular localization
11	regulation of hydrolase activity	41	transport
12	signal transduction	42	DNA metabolism
13	regulation of cellular physiological process	43	phosphoinositide-mediated signaling
14	cell cycle	44	G2/M transition of mitotic cell cycle
15	chromosome segregation	45	mitosis
16	cell division	46	meiosis
17	cell organization and biogenesis	47	chromosome organization and biogenesis
18	localization	48	cytoskeleton organization and biogenesis
19	cellular metabolism	49	intracellular transport
20	primary metabolism	50	DNA replication
21	macromolecule metabolism	51	DNA recombination
22	intracellular signaling cascade	52	regulation of mitosis
23	regulation of cell cycle	53	mitotic sister chromatid segregation
24	interphase	54	chromosome condensation
25	mitotic cell cycle	55	microtubule-based process
26	M phase	56	cytoskeleton-dependent intracellular transport
27	meiotic cell cycle	57	DNA-dependent DNA replication
28	cytokinesis	58	mitotic chromosome condensation
29	cellular localization	59	microtubule cytoskeleton organization and biogenesis
30	establishment of localization	60	microtubule-based movement
31	nucleobase, nucleoside, nucleotide	61	spindle organization and biogenesis

The significant CC-graph within the whole CC-graph



Outline

1. Introduction
   - Gene Ontology
   - Testing gene sets
2. Multiple Testing on the GO graph
   - Background
   - Bottom-up
   - Top-down
   - The Focus Level Method
3. Application
   - Netherlands Cancer Institute data
4. Conclusion

The Focus Level method

Advantages

- More power by using GO structure
- Focus power at the interesting level of detail
- Returns a coherent subtree of GO
- Strong statement due to control of the Family-wise error rate

Important remarks

- FWER statement is a statement about the whole graph
- Interpret individual FWER-adjusted p-values in context
- FWER sometimes more appropriate than FDR

Read more?


Goeman and Mansmann (2006).
Multiple testing on the DAG of Gene Ontology.
Bioinformatics **24** (4) 537544.


Goeman, Van de Geer, De Kort, Van Houwelingen (2004).
A global test for groups of genes.
Bioinformatics, **20** (1) 93–99.


Goeman, Oosting, Cleton-Jansen, Anninga, Van Houwelingen (2005).
Testing association of a pathway with survival.
Bioinformatics, **21** (9) 1950–1957.


Goeman, Van de Geer and Van Houwelingen (2006).
Testing against a high-dimensional alternative.
JRSSB, **68** (3) 477–493.