Multiple testing in gene set test analysis: possibilities and pitfalls

Jelle Goeman

Based on joint work with Rosa Meijer, Aldo Solari, Livio Finos, Ulrich Mansmann

Padua, 2015-10-09

Genetic data: two approaches

Single markers

H0: no association between marker and response

Sets of markers: 'gene sets'

- Sets of markers chosen a priori
- Chromosomic regions, genes, pathways, functional groups
- H0: no association between any marker and response

Self-contained approach

Single marker test = test of gene set of size 1

Multiple testing of gene sets

Structure in gene sets

- Subset relationships
- Large sets, small sets
- Include individual genes in single analysis?

How to use subset relationships

- Interpretation
- Choice of error rate
- Gain power?

Logical relationships

Null hypothesis of gene set A

$H_A : \theta_i = 0$ for all $i \in A$

Logical relationships: 1-way

If $A \subset B$:
H_A false implies H_B false

Logical relationships: 2-way

Additionally, if $A = \bigcup A_i$:
H_A false implies at least one of H_{A_i} false

Note

2-way relationships not relevant in all graphs

Implications: interpretation

Large sets, small sets

If $A \subset B$, rejection of B not informative if A also rejected

Summarizing: defining hypotheses

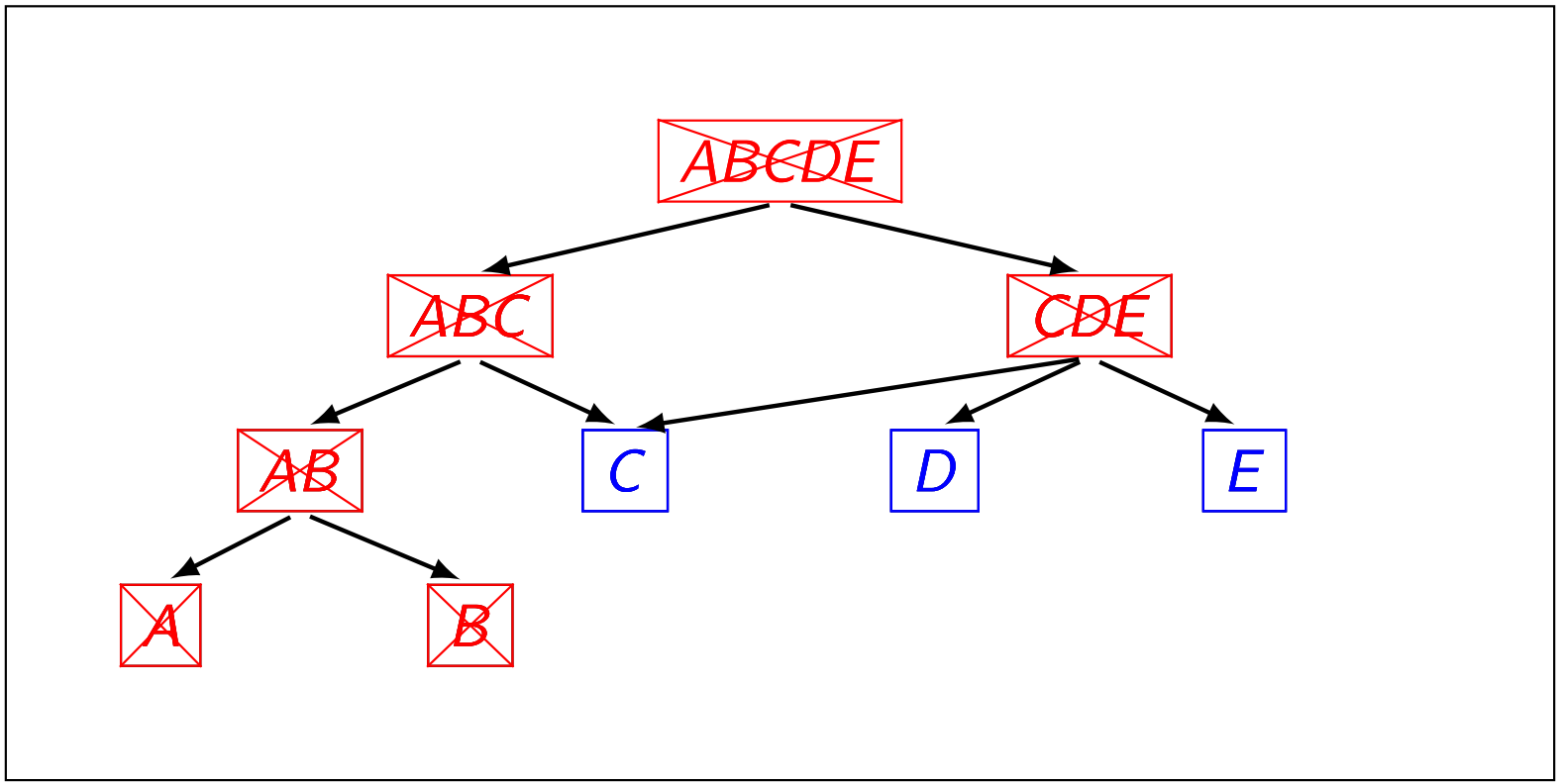
Relevant rejections: only those with no rejected subset hypotheses

Explaining

Most specific statements that can be made

Defining hypotheses

Defining hypotheses: A, B, CDE



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 problem

Control of FDR does not imply FDR control on subset

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

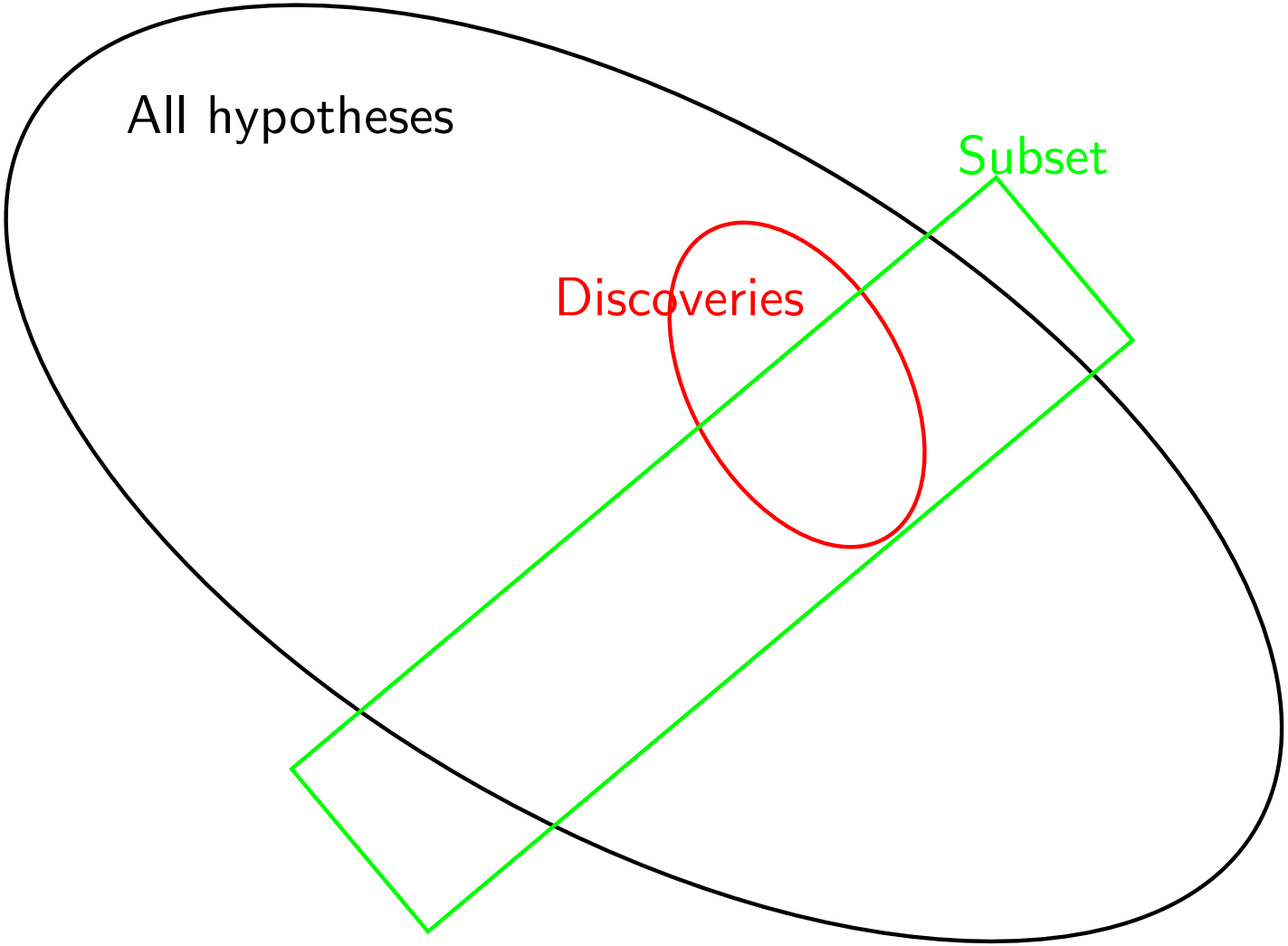
Subsets problem

Control of FDR does not imply FDR control on subset

Finner and Roters

- FDR control on all subsets = FWER control
- FWER control on all subsets = FWER control

Subsets of the discoveries



Sequential rejection

Sequential procedure

Many FWER control procedures follow this format

- Start testing hypotheses at some significance criterion
- If any hypotheses rejected, set new significance criterion for unrejected hypotheses
- Possibly new rejections...
- Stop if no new rejections occur

Strong control

All have strong control of the FWER

Is there a connection?

All special cases of the same procedure

Romano and Wolf

Romano & Wolf (2005)

"an ideal situation would be to proceed at any step without regard to previous rejections, in the sense that once a hypothesis is rejected, the remaining hypotheses are treated as a new family, and testing for this new family proceeds independent of past decisions."

The sequential rejection principle

In fact, past decisions even make the tests in each new family easier

Formal setup

Model

- A statistical model $\mathbb{M}$
- Each $M \in \mathbb{M}$ indexes a probability measure P_M

Null hypotheses

- A collection of null hypotheses $\mathcal{H}$ (possibly infinite)
- Each $H \in \mathcal{H}$ is a submodel $H \subset \mathbb{M}$
- Some hypotheses are true (depends on $M \in \mathbb{M}$)
- True hypotheses: $T(M) = \{H \in \mathcal{H} : M \in H\} \subseteq \mathcal{H}$

Test statistics

- A test statistic S_H for every $H \in \mathcal{H}$
- Reject H for large values of S_H

A general sequentially rejective procedure

The critical value function

- Choose critical value function $\mathbf{c} = \{c_H\}_{H \in \mathcal{H}}$
- Each $c_H : 2^{\mathcal{H}} \to \mathbb{R}$
- $2^{\mathcal{H}}$: collection of all subsets of $\mathcal{H}$

The general procedure

$R_i \subseteq \mathcal{H}$: the rejected hypotheses after step i

$$\begin{aligned} R_0 &= \varnothing \\ R_{i+1} &= R_i \cup \{H \in \mathcal{H} : S_H > c_H(R_i)\}. \end{aligned}$$

The Sequential Rejection Principle

Theorem

If a general sequentially rejective procedure fulfils two conditions

- Monotonicity
- Single step control

Then it strongly controls the FWER

Monotonicity

The monotonicity condition

For every $A \subset B \subset \mathcal{H}$ and for every $H \in \mathcal{H} \setminus B$,

$$c_H(A) \geq c_H(B),$$

In words

Critical values of unrejected null hypotheses never increase with more rejections

Single step control

The single step condition

For every $A \subset \mathcal{H}$ and for every $M \in \mathbb{M}$ for which $T(M) = \mathcal{H} \setminus A$,

$$\mathrm{P}_M\left(\bigcup_{H \in T(M)} \{S_H > c_H(A)\} \right) \leq \alpha.$$

In words

Weak FWER control at each single step

But only partially (weaker than weak control)

It may be assumed that all previous rejections were correct rejections

Three methods for DAGs

Goeman and Mansmann (2009)

Focus level method: starts in the 'middle'

Meijer and Goeman (2015a)

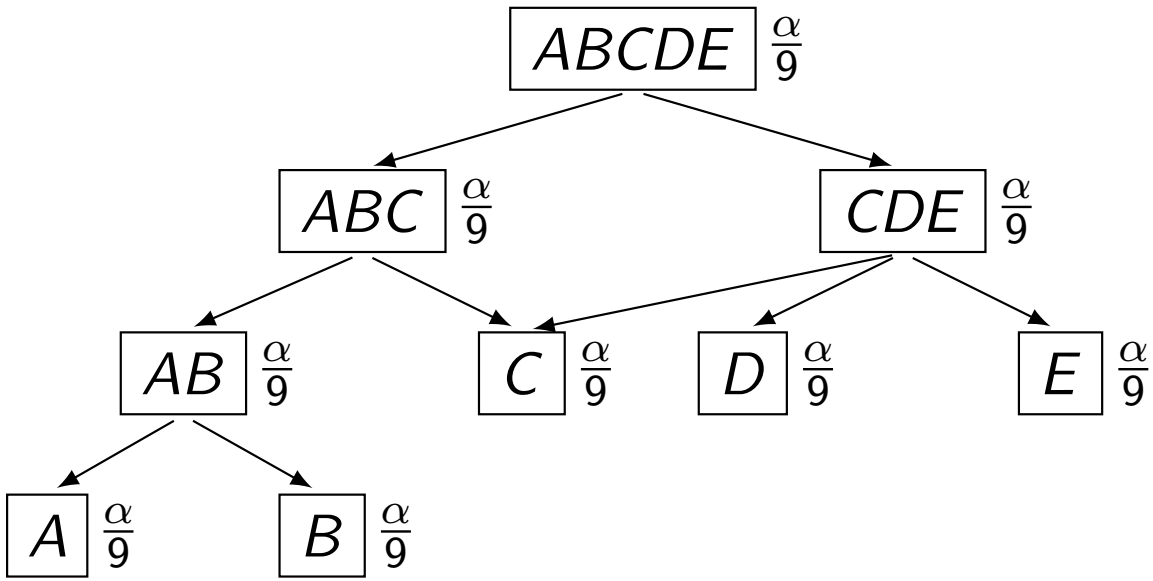
Generalized Meinshausen (2008) to DAG: start at the top

Meijer and Goeman (2015b)

Structured Holm (Shaffer variant): simultaneous method

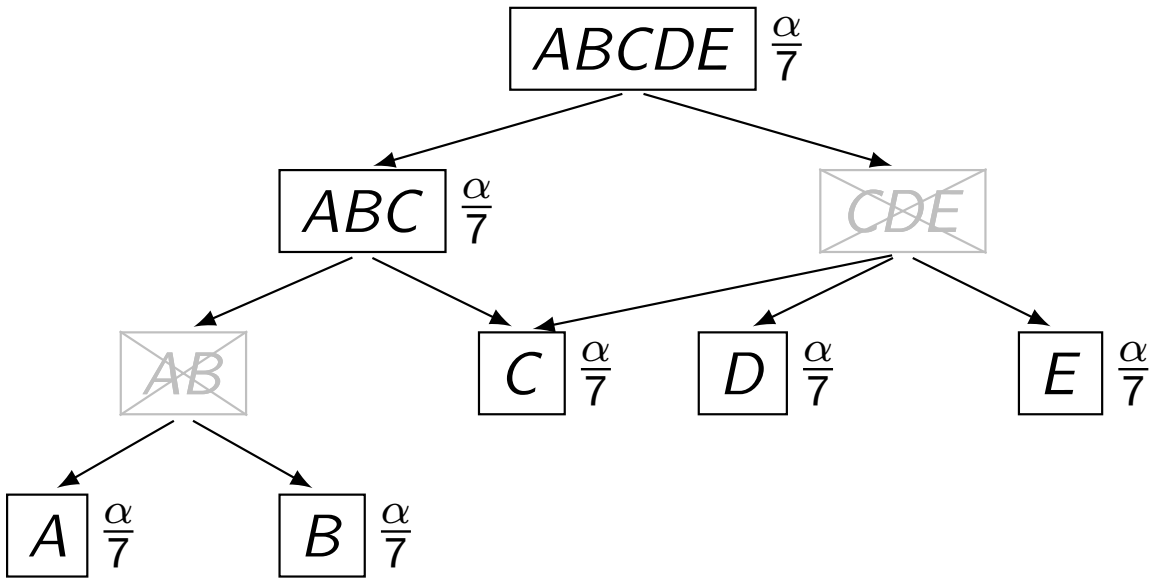
Logical Holm: simultaneous

First, a Bonferroni correction is applied. Each hypothesis is tested on $\alpha/9$.



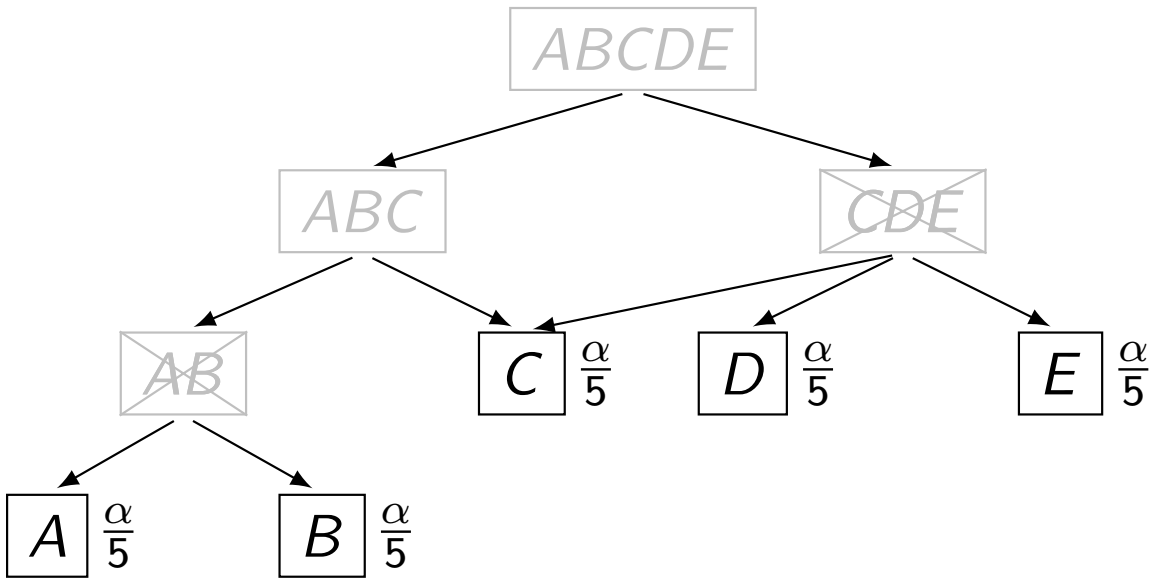
Logical Holm: simultaneous

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



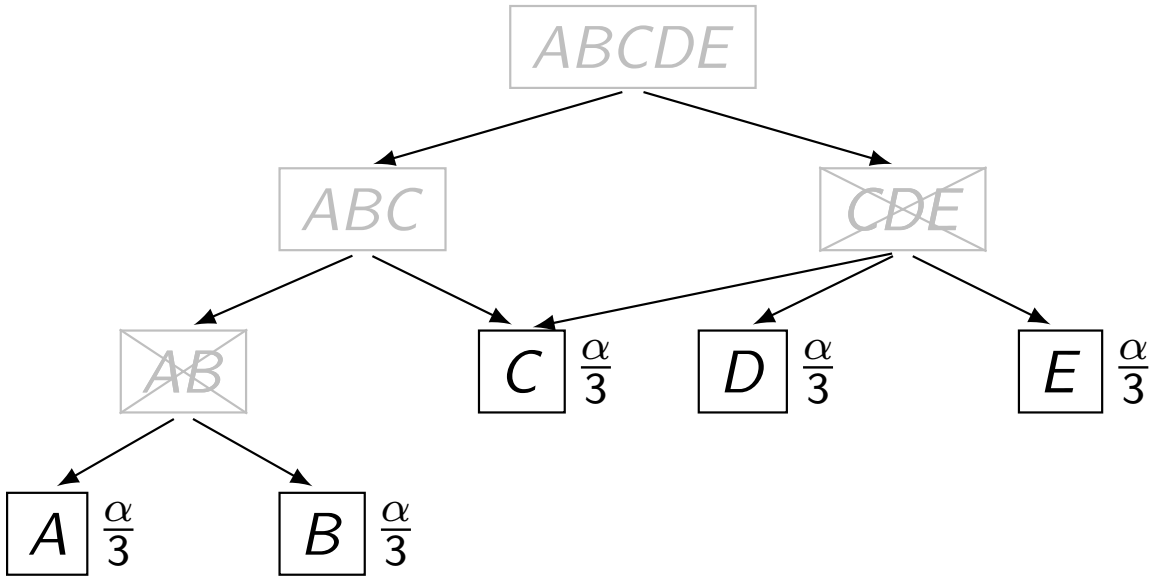
Logical Holm: simultaneous

However, using the one-way relations shows that two hypotheses can no longer be true. The remaining hypotheses can be tested on $\alpha/5$.



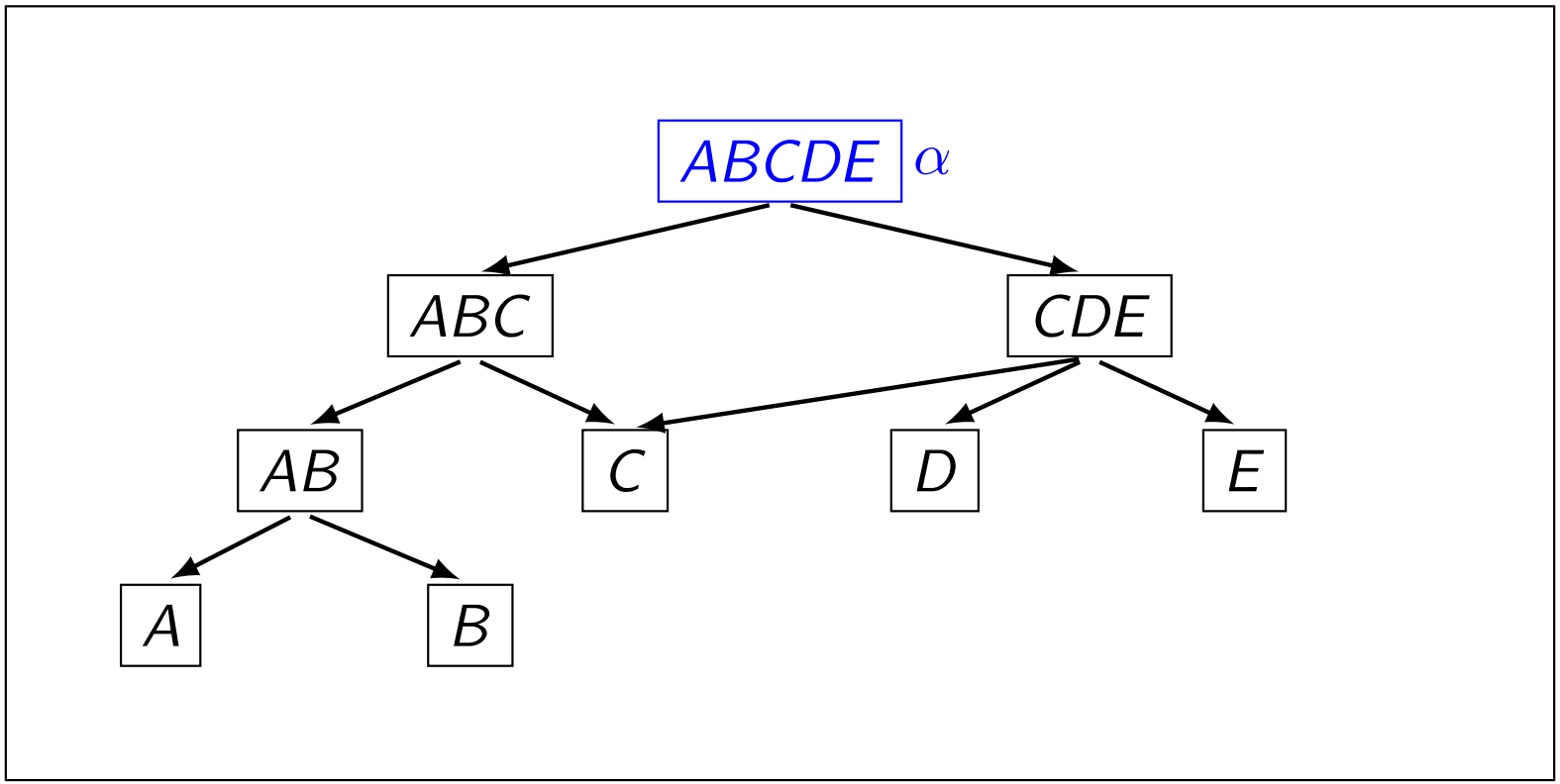
Logical Holm: simultaneous

Using the two-way relations, 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.



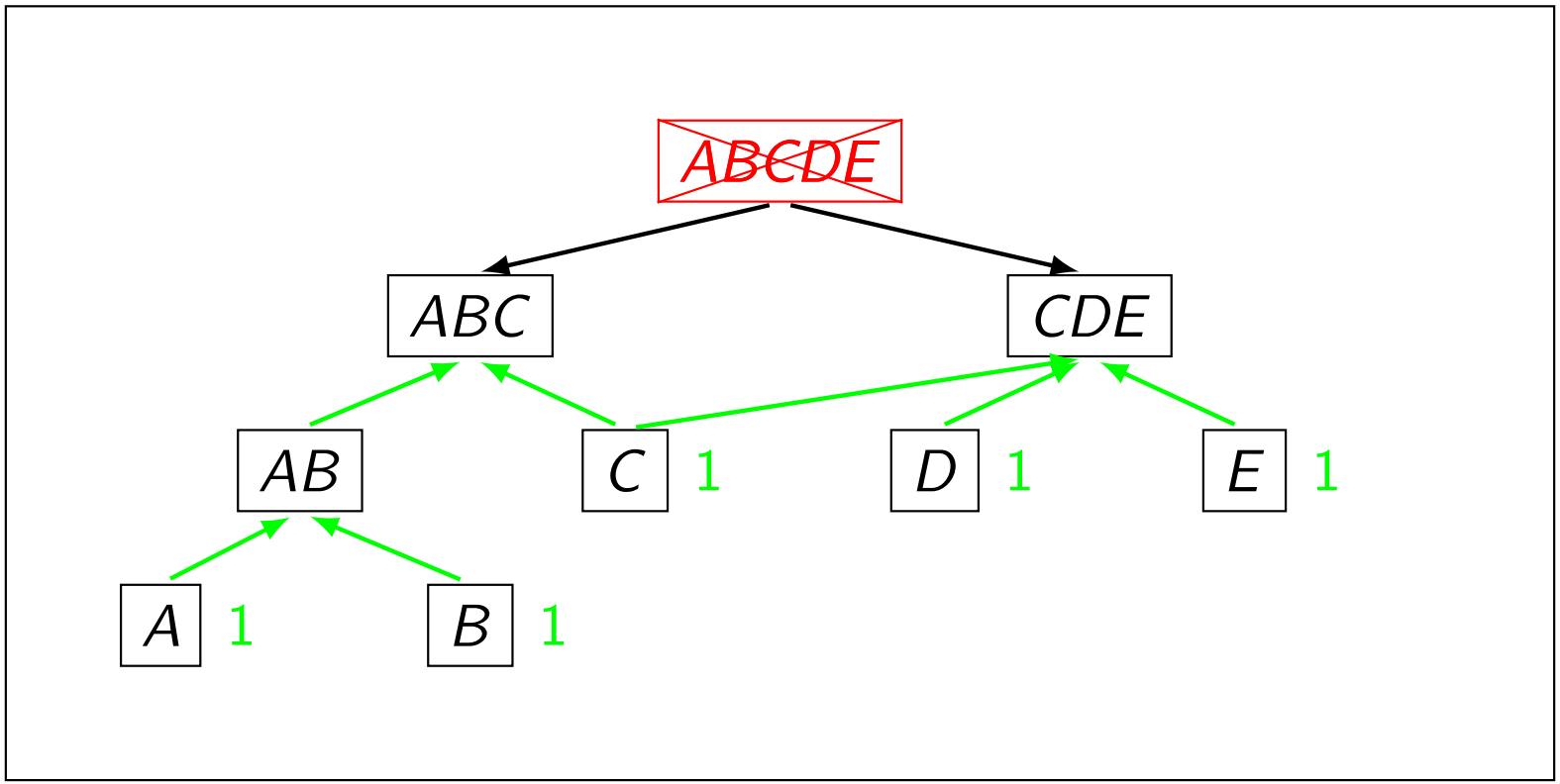
Top down method

Start at the top on level α (all weight goes there)



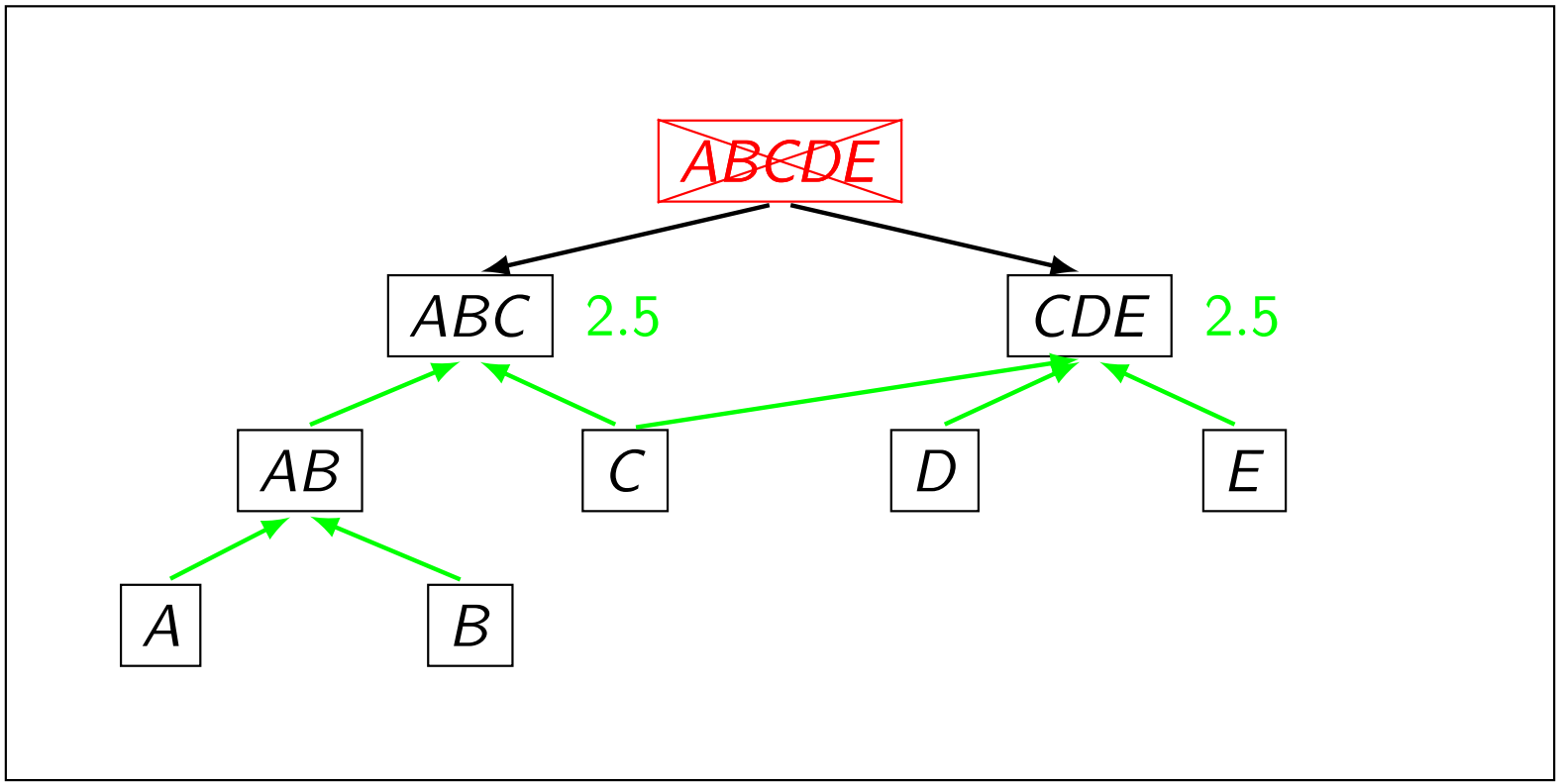
Top down method

If rejected: weight goes to ABC and CDE



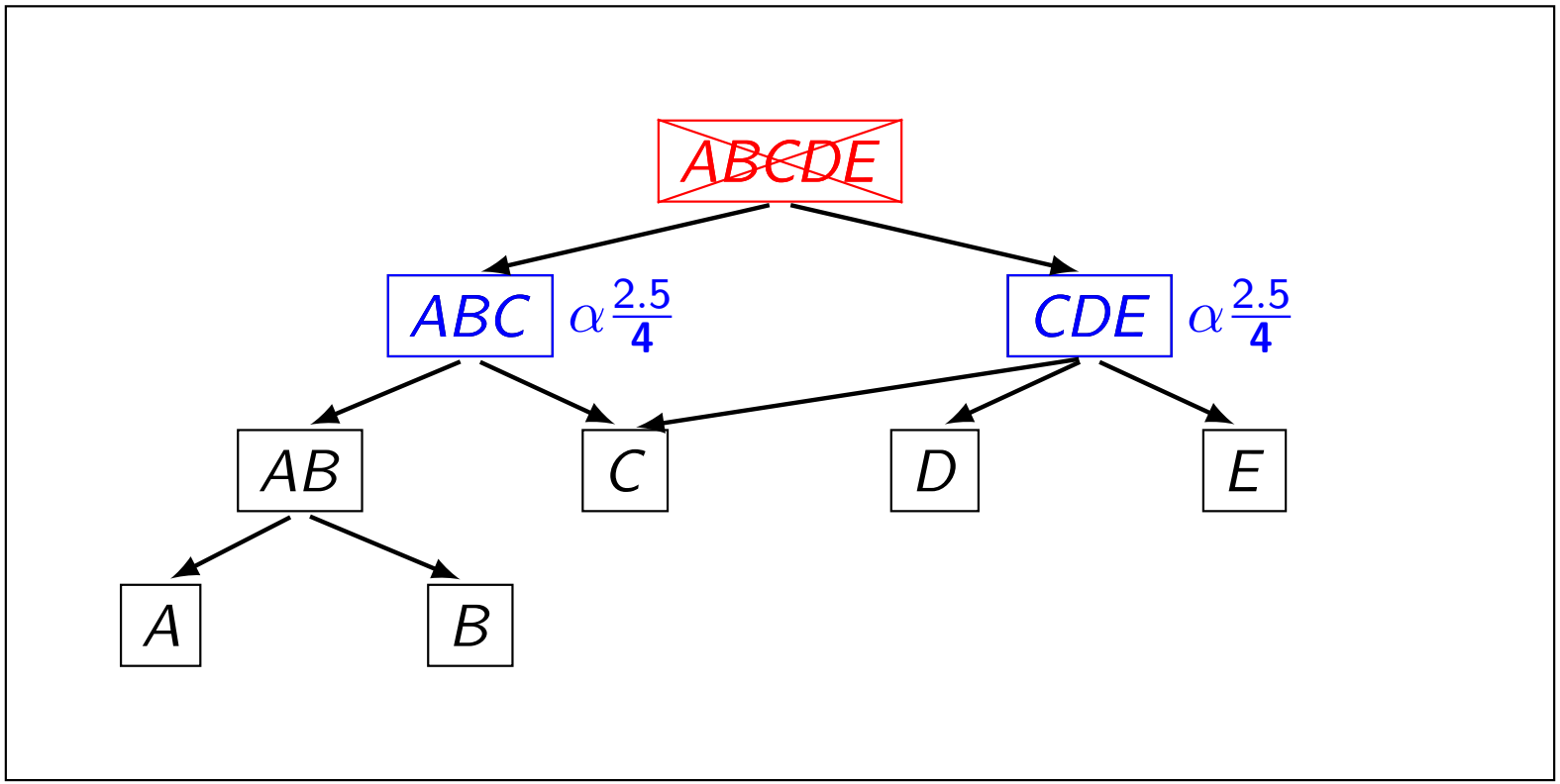
Top down method

weight C equally split gives this



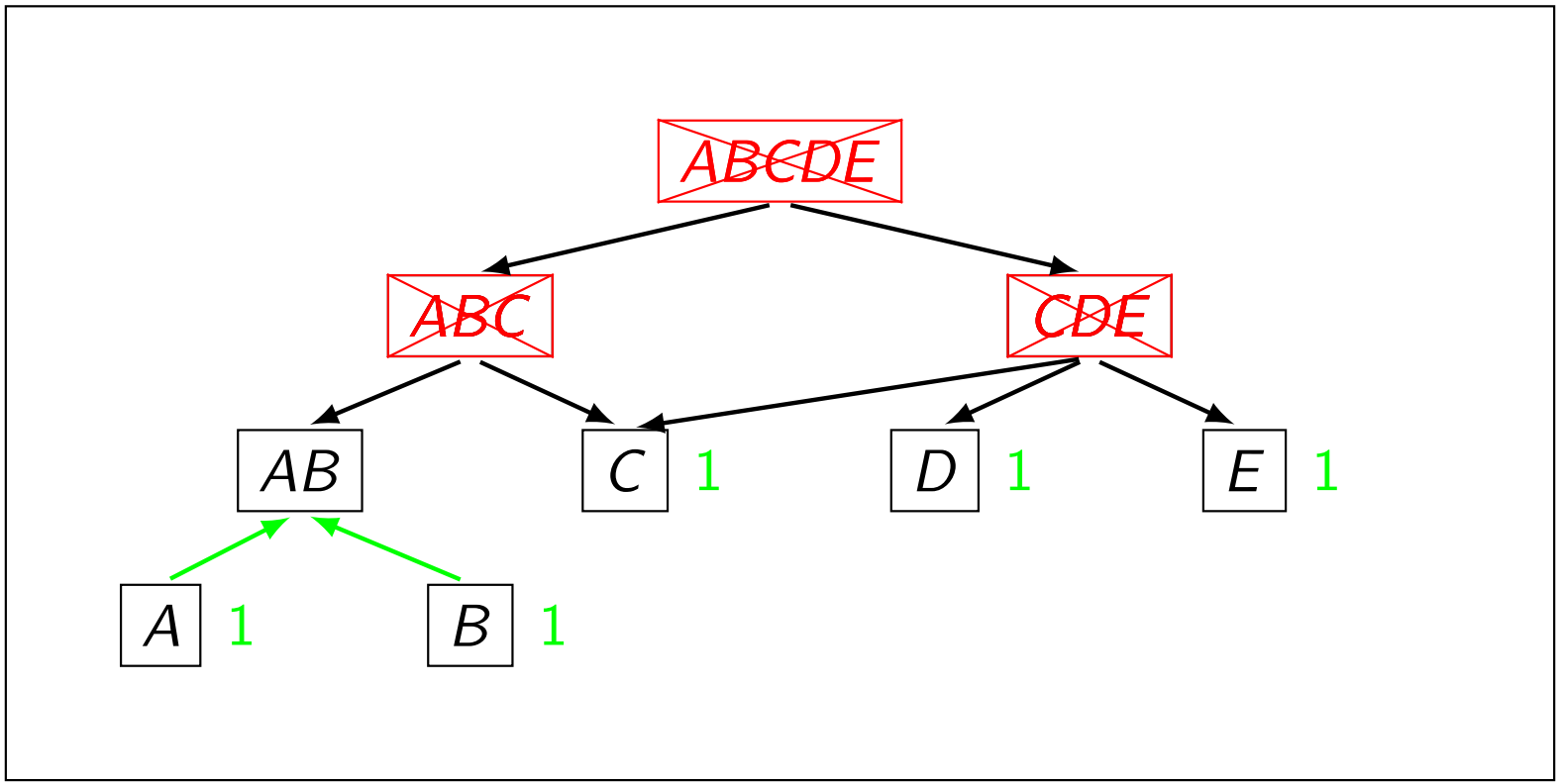
Top down method

implication $ABCDE$: congruent set ≥ 1 unrejected leaf $\Rightarrow$ /(5-1)



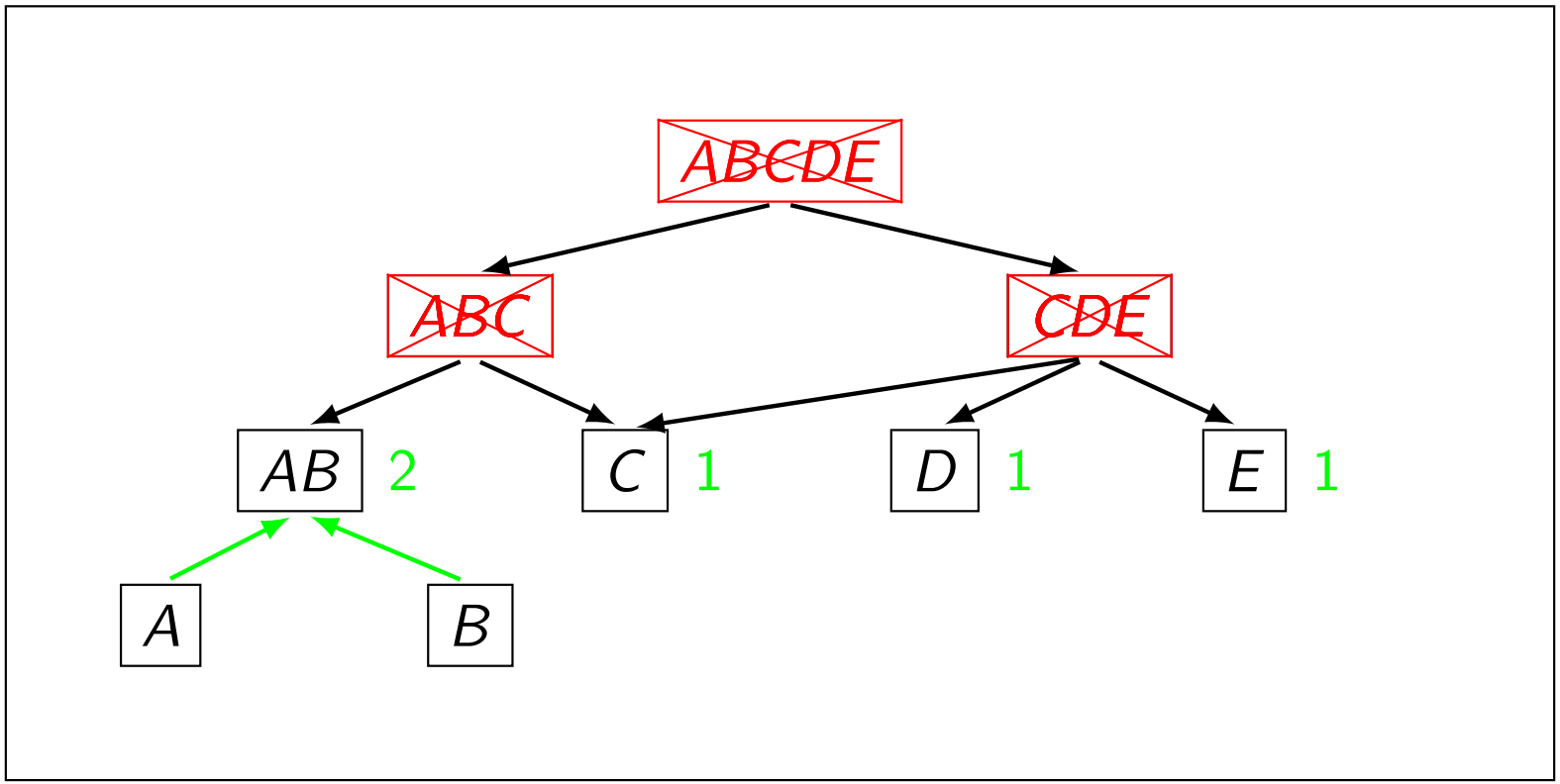
Top down method

If both rejected, weight flows to AB, C, D and E



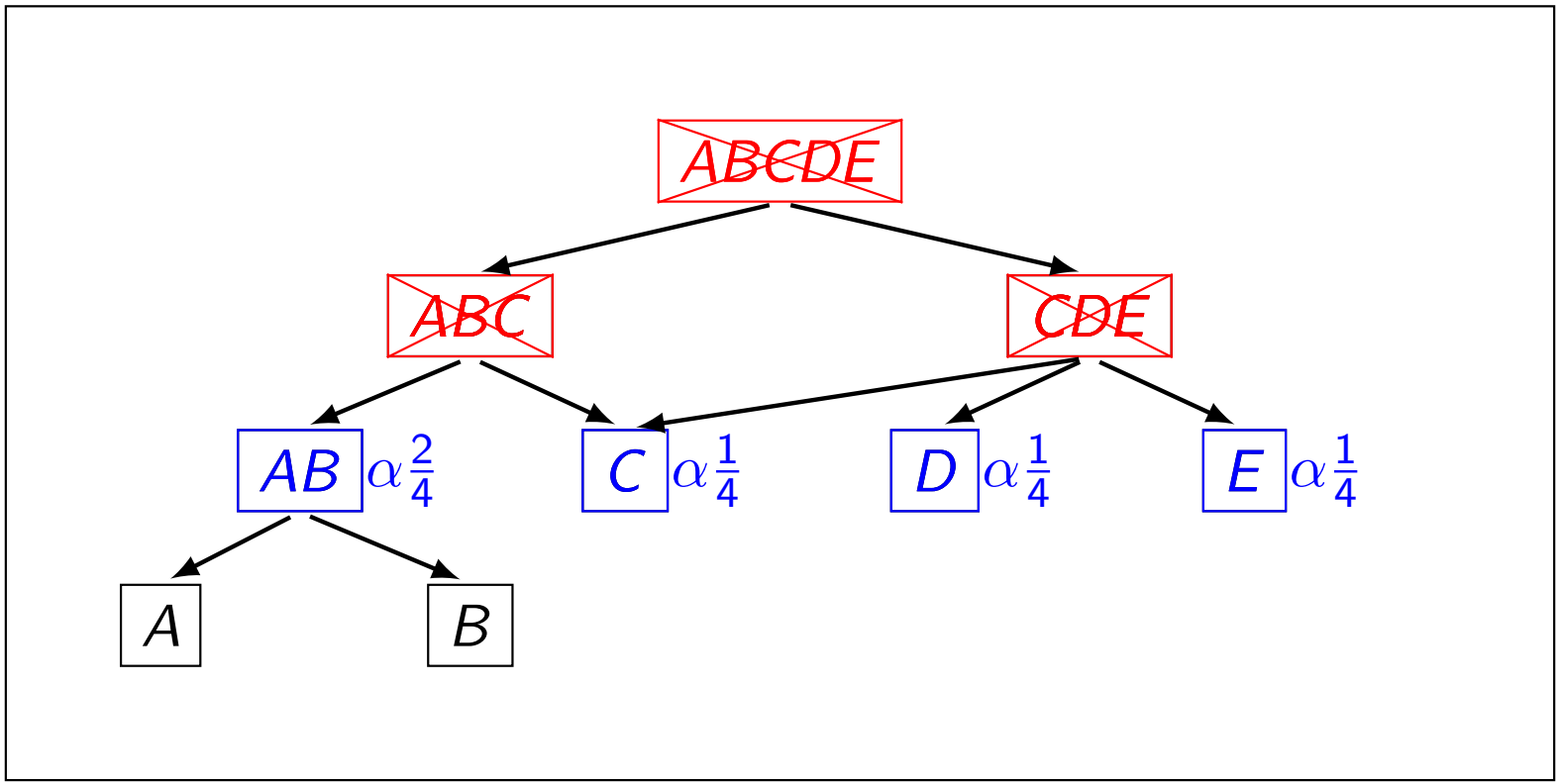
Top down method

Weights are added



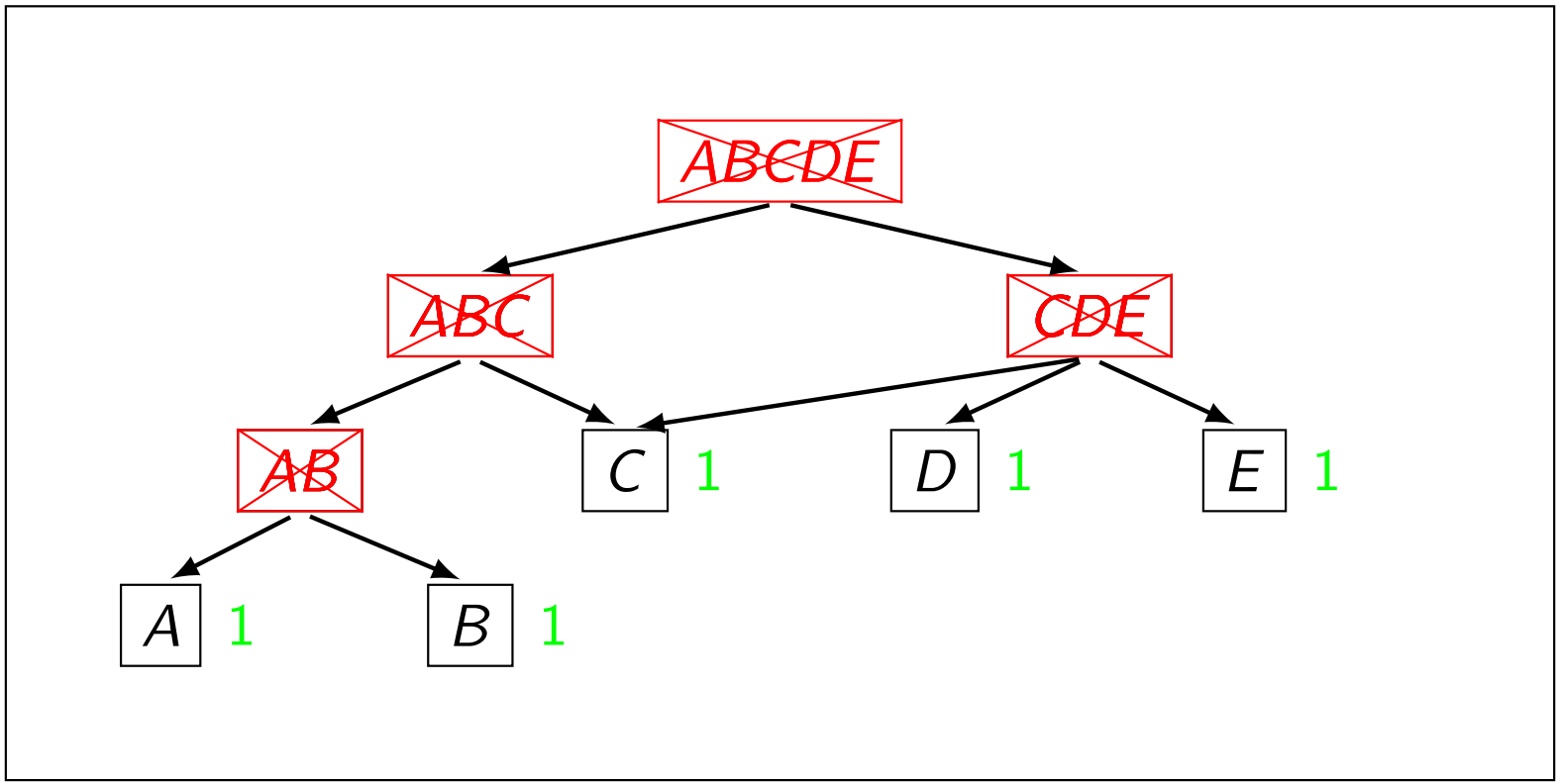
Top down method

Implications ABC and CDE, only implies 1 rejection (C): /(5-1)



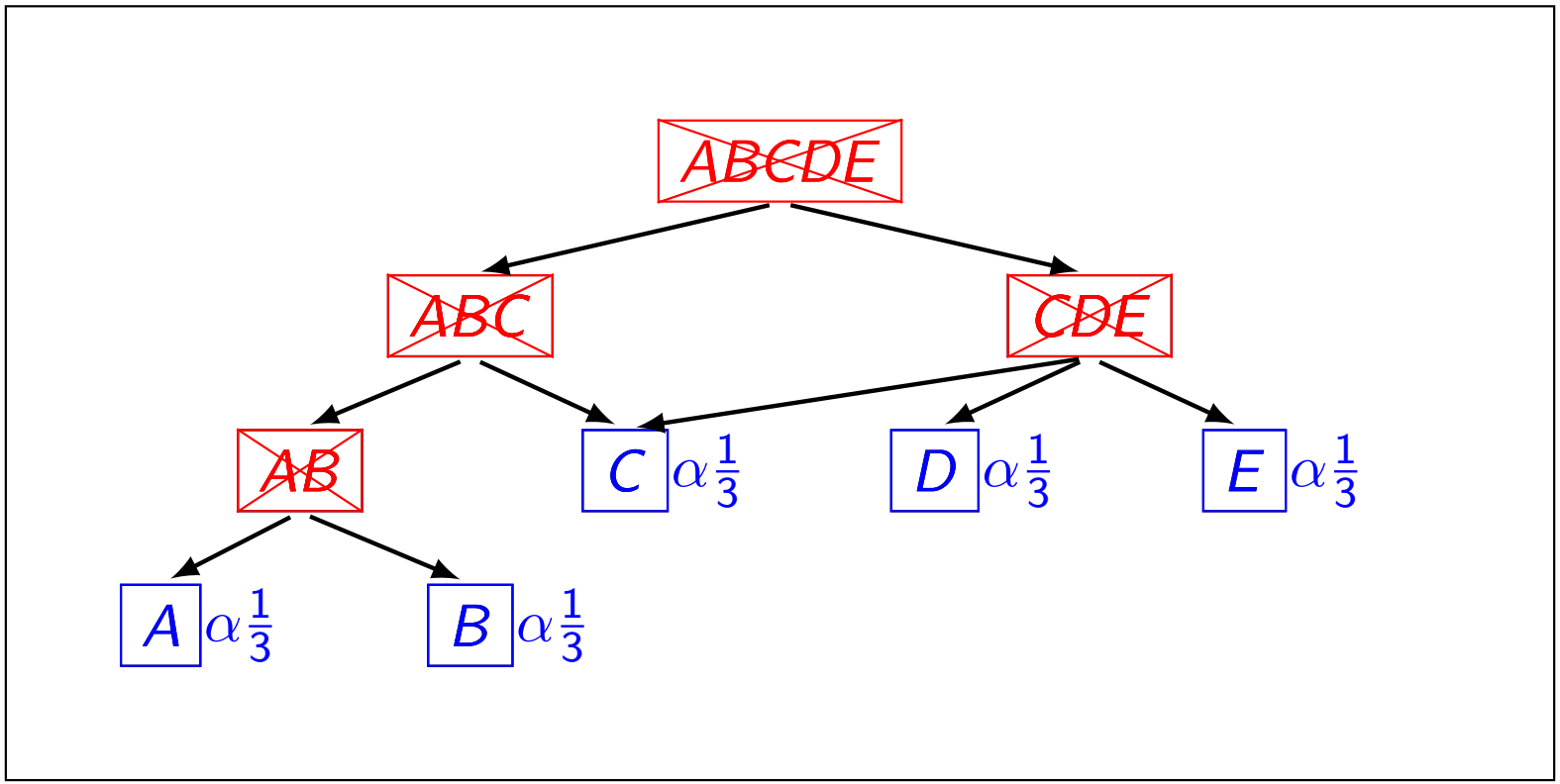
Top down method

If AB rejected: all leafs weight 1



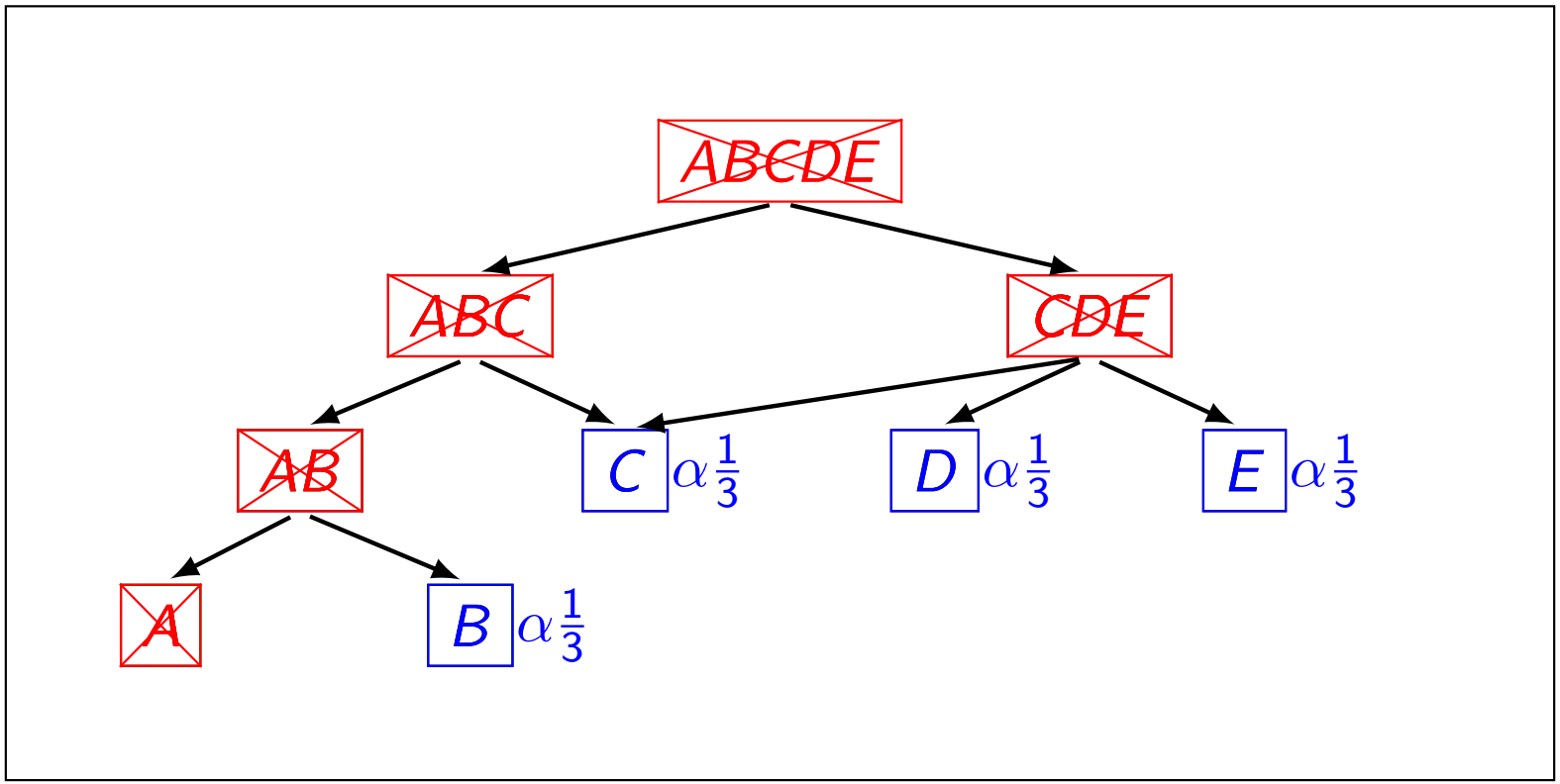
Top down method

Implications AB and CDE, no overlap: 2 rejections: /(5-2)



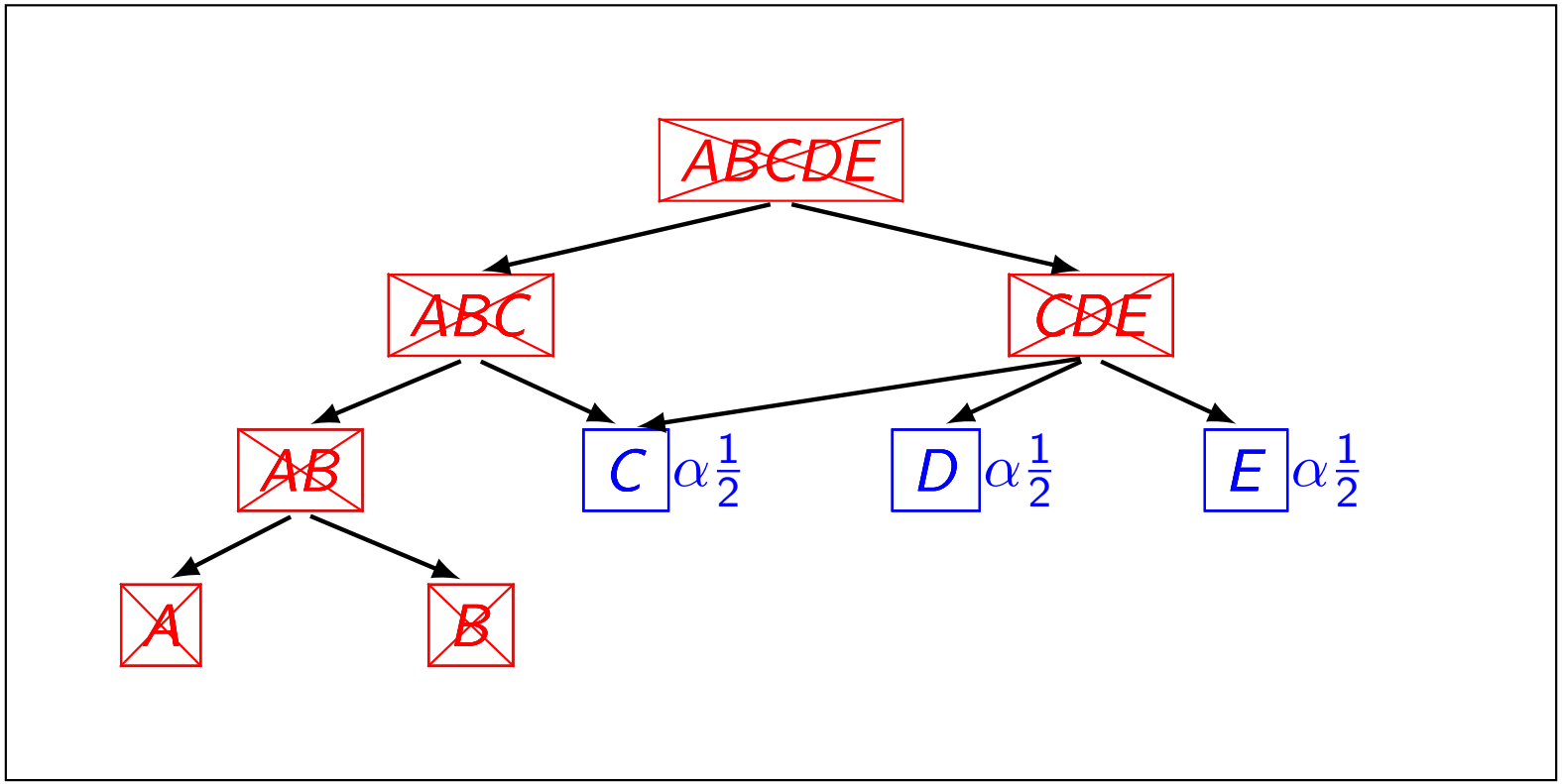
Top down method

If A rejected: nothing changes (2 rejected leaves for congruent set)



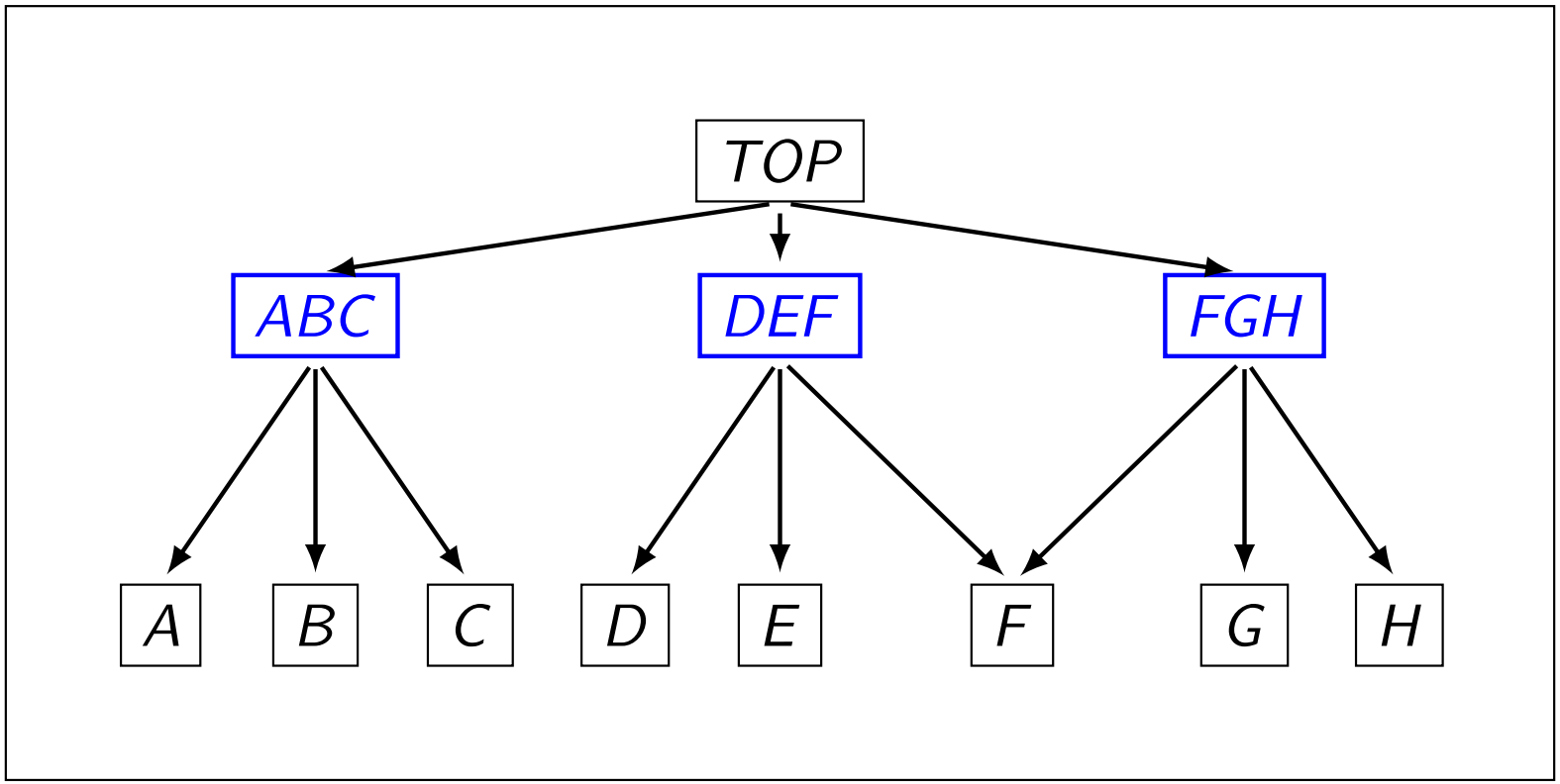
Top down method

If B gets rejected as well: /(5-3)



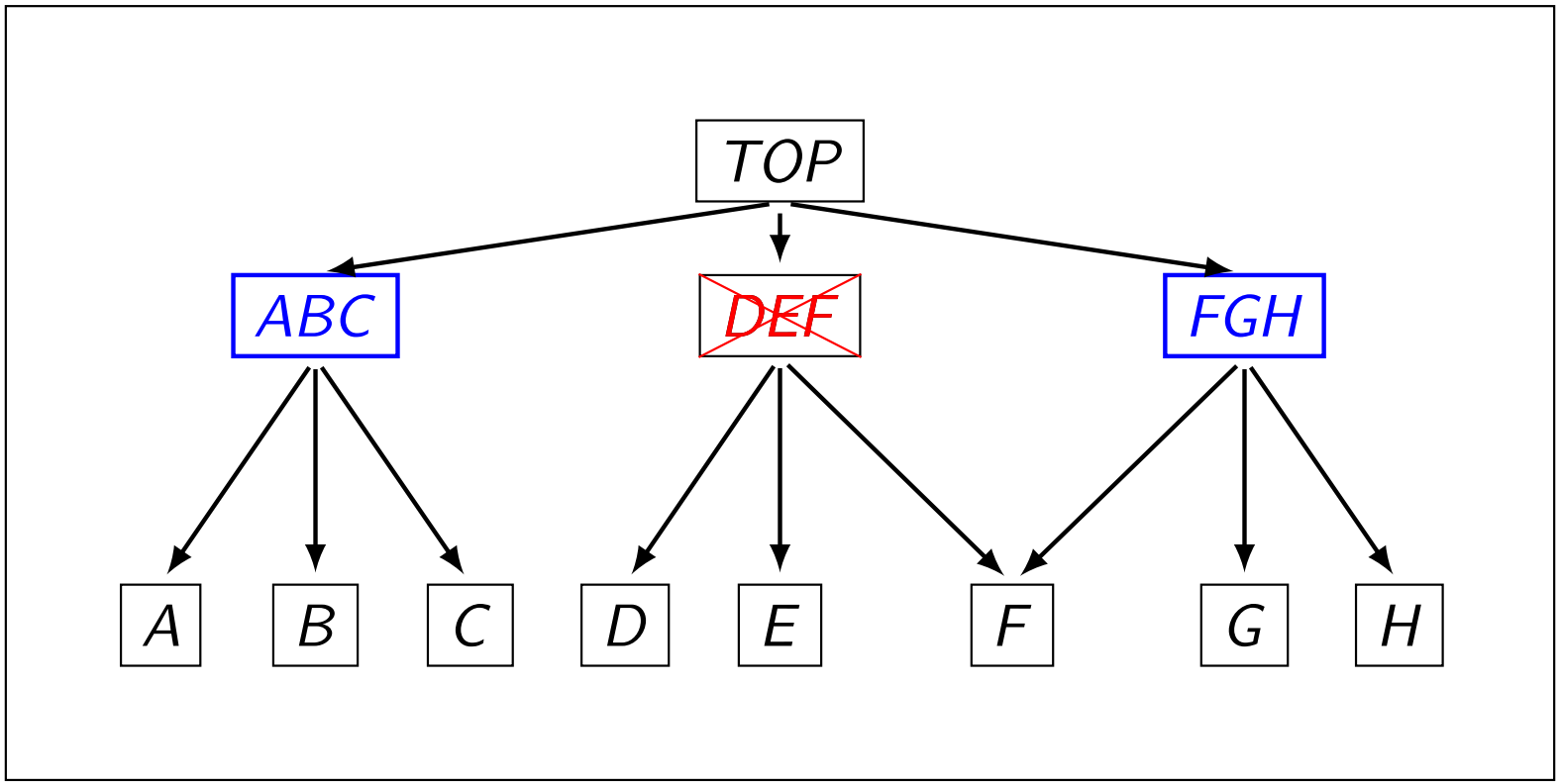
The focus level procedure

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



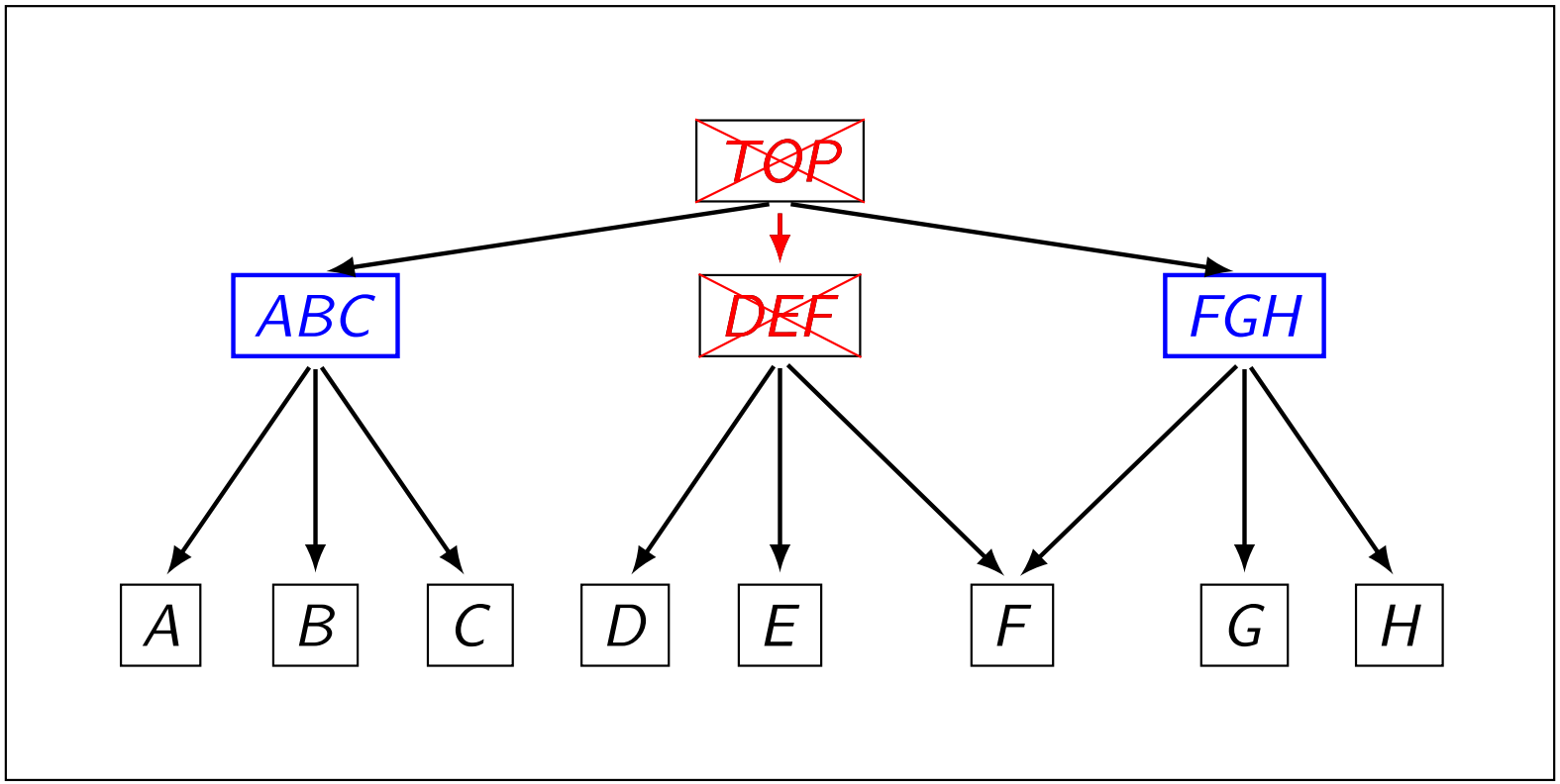
The focus level procedure

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



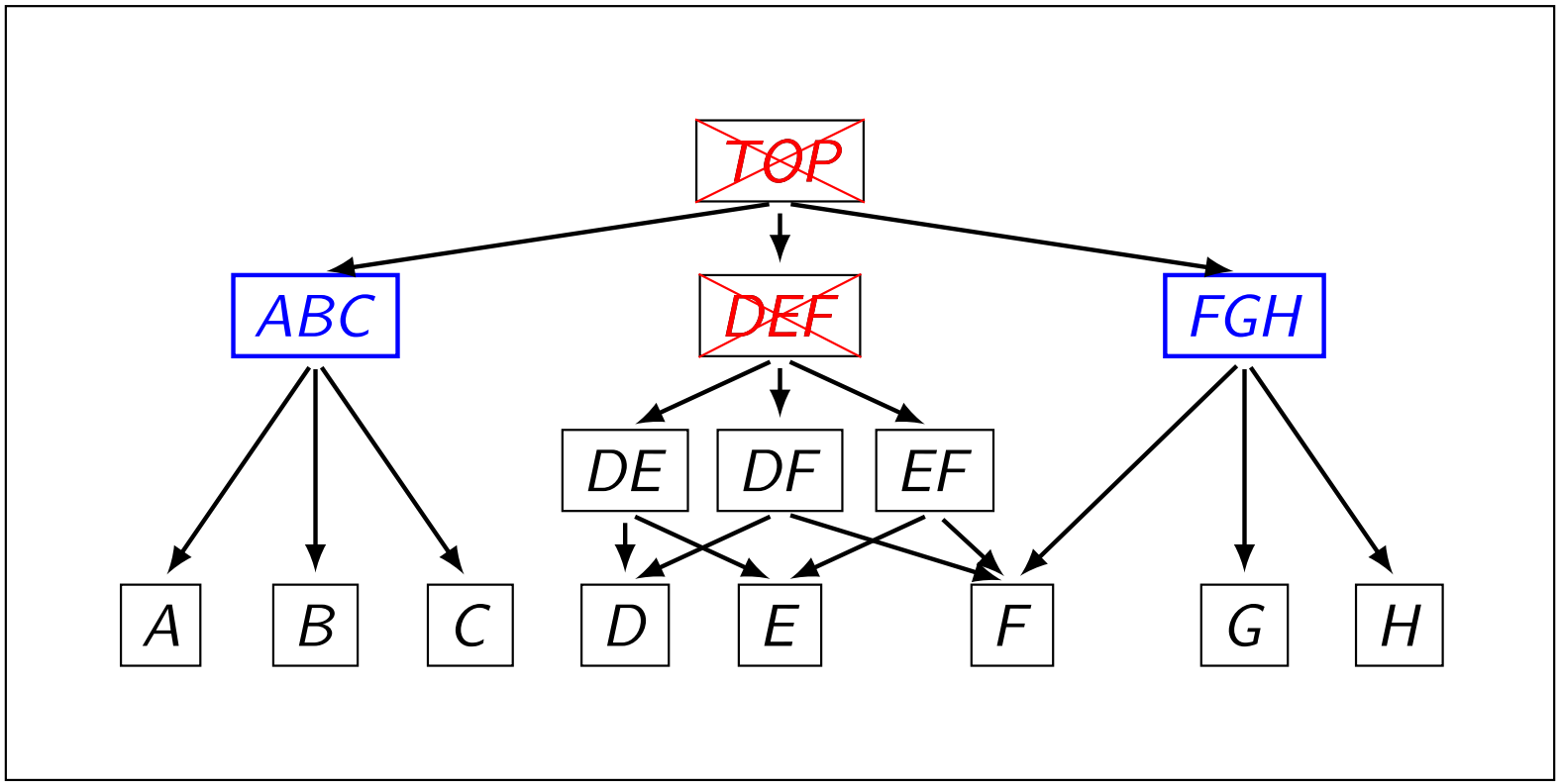
The focus level procedure

Use the bottom-up procedure to propagate significance upwards



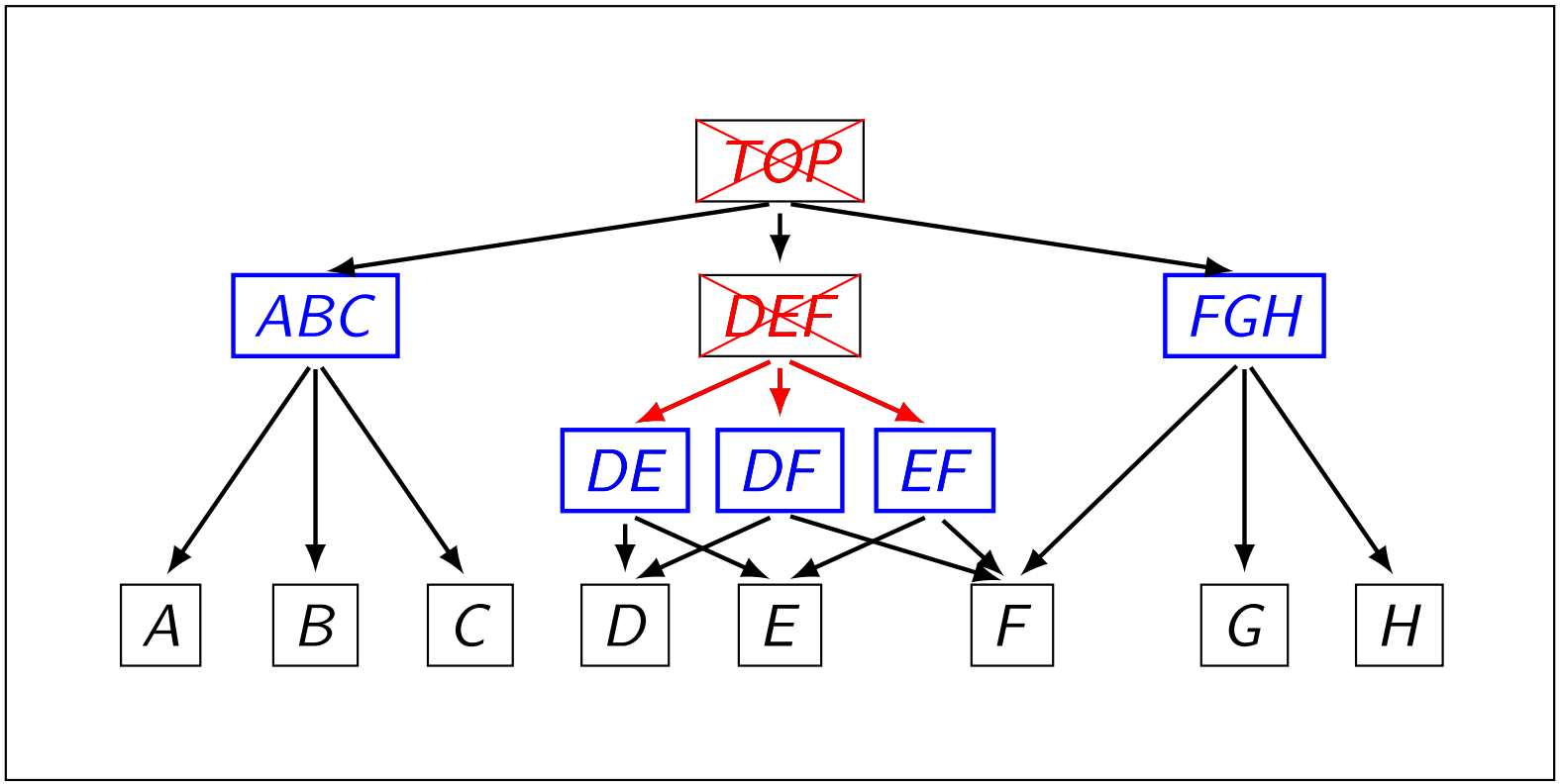
The focus level procedure

Expand the graph below *DEF*



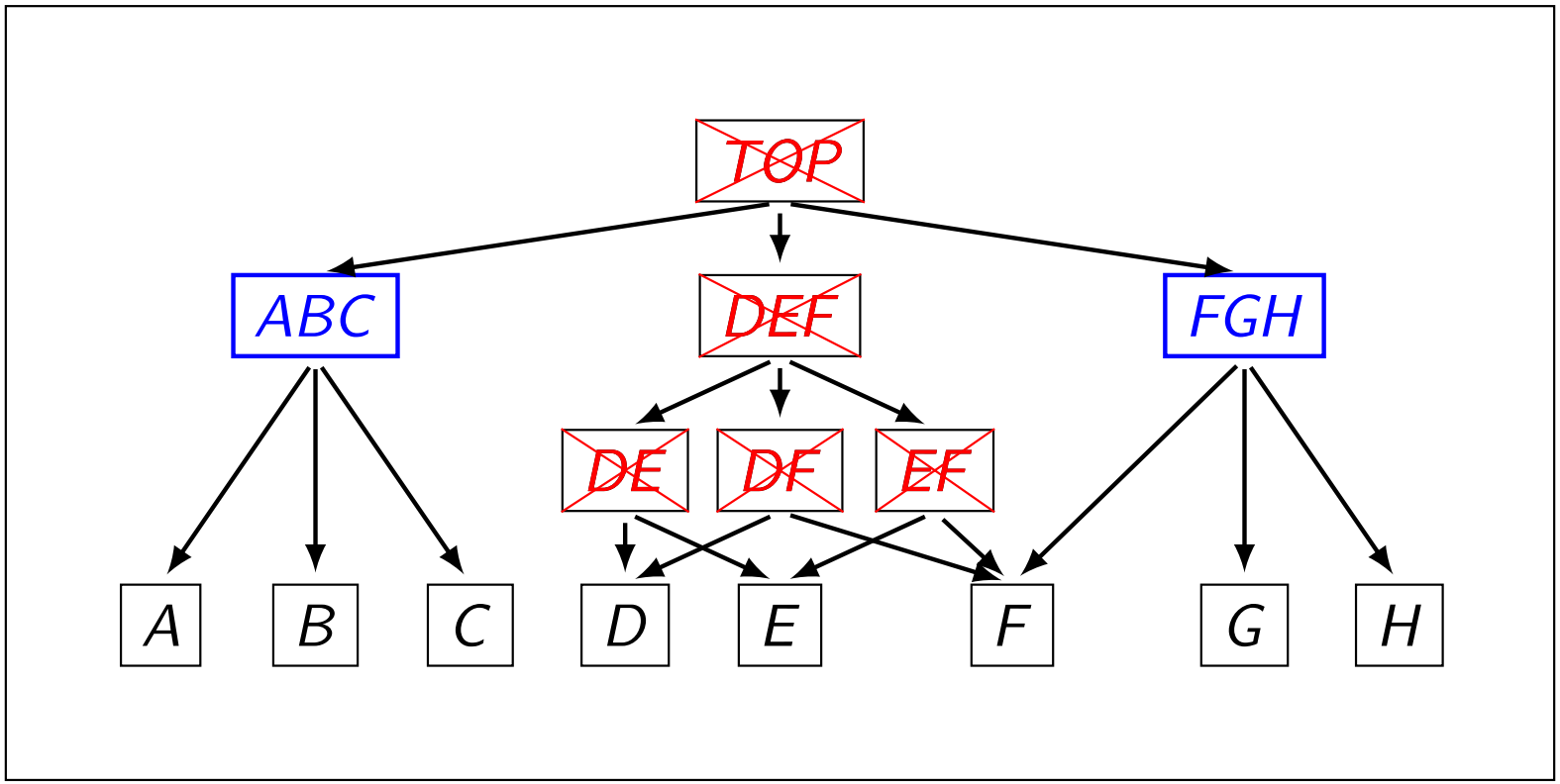
The focus level procedure

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



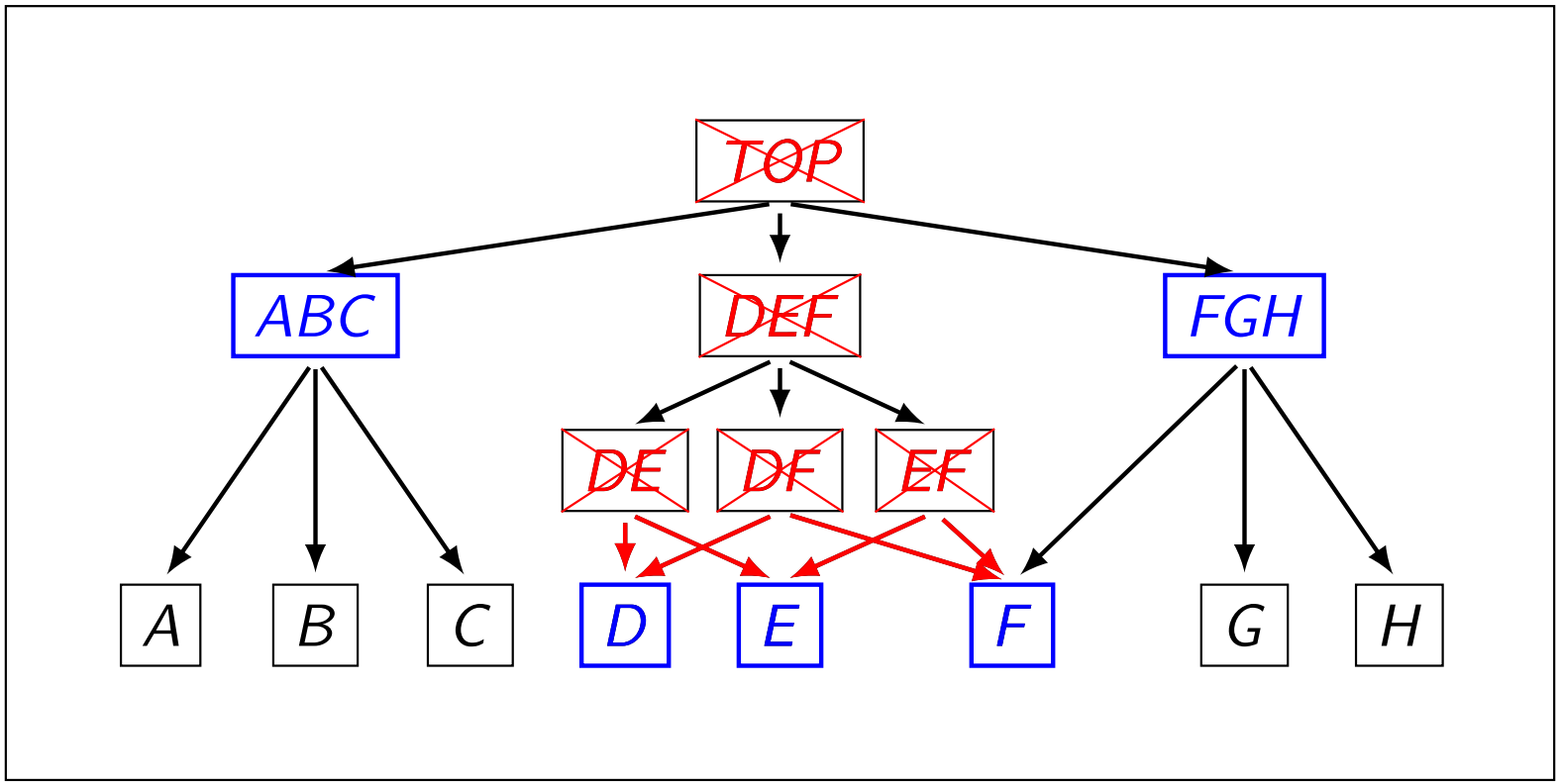
The focus level procedure

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



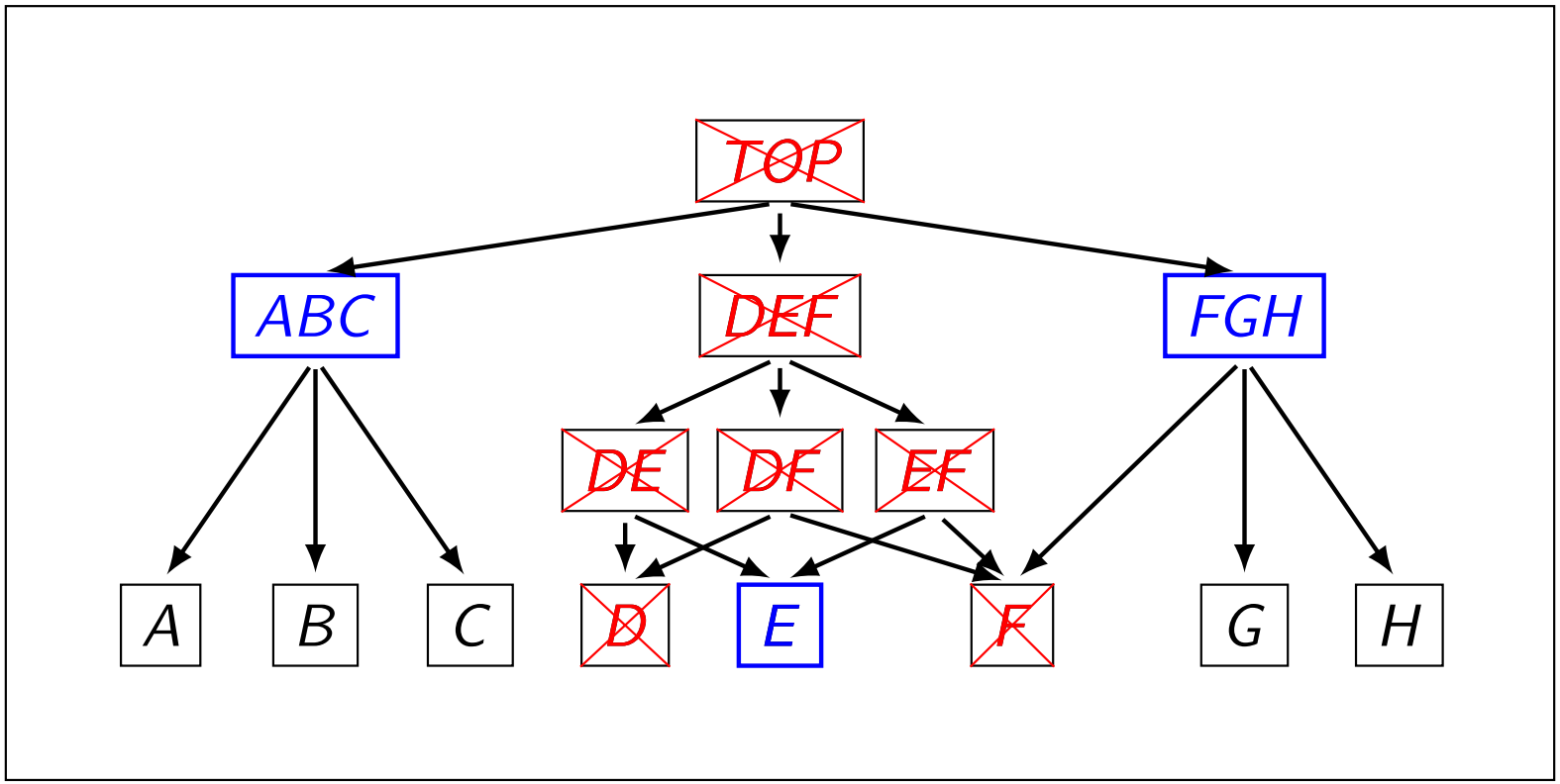
The focus level procedure

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



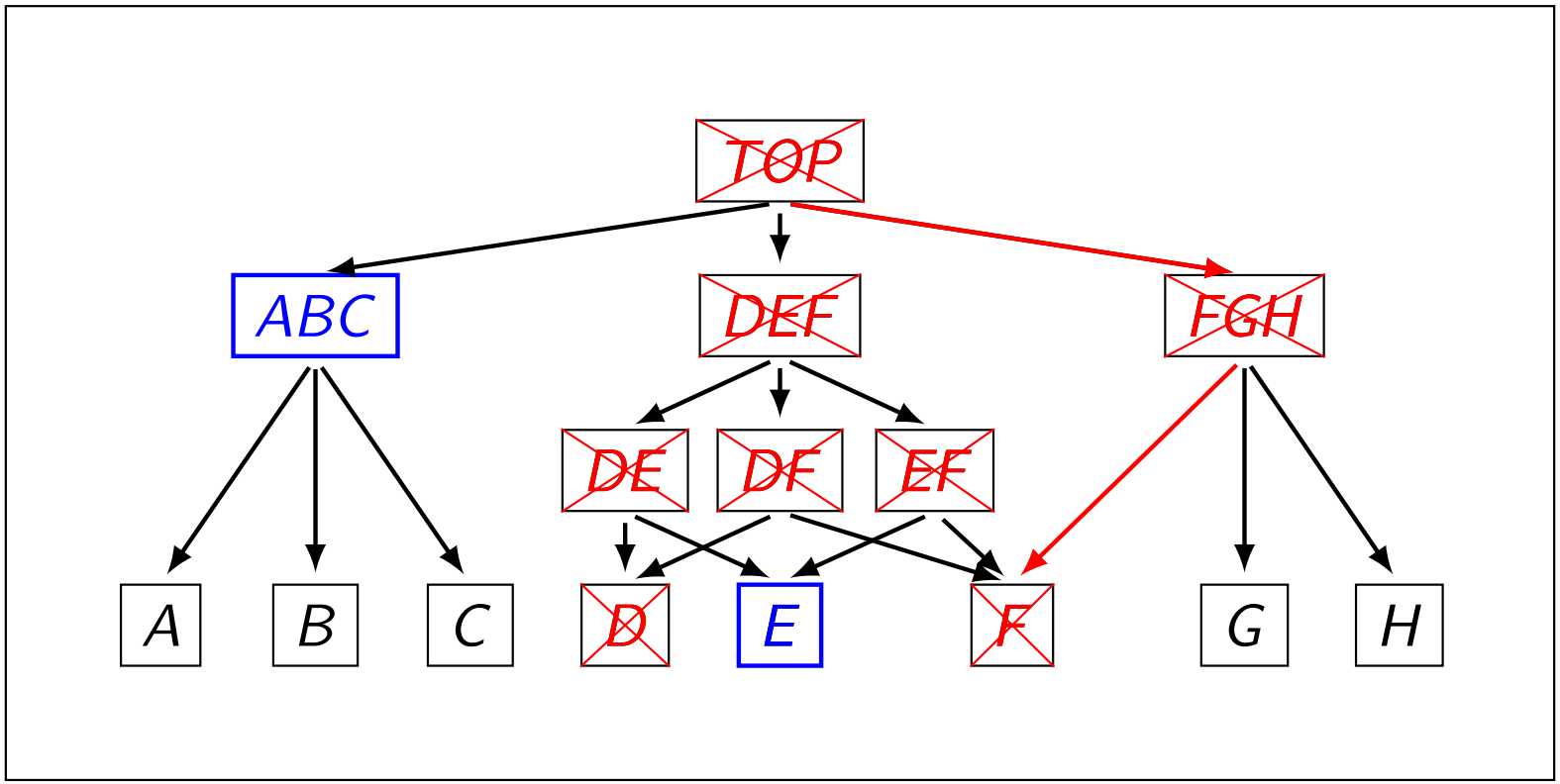
The focus level procedure

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



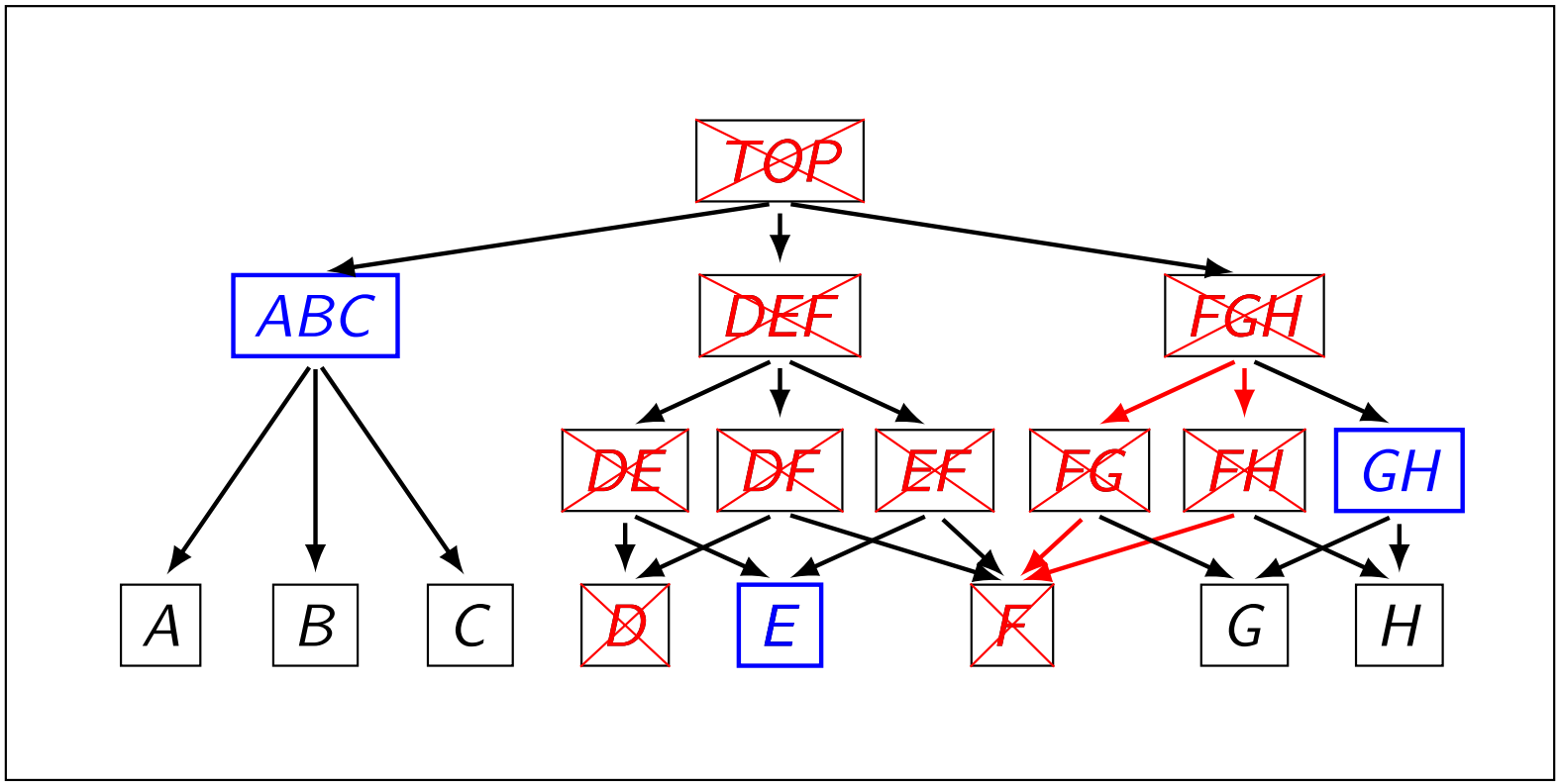
The focus level procedure

Call all ancestors of significant nodes significant



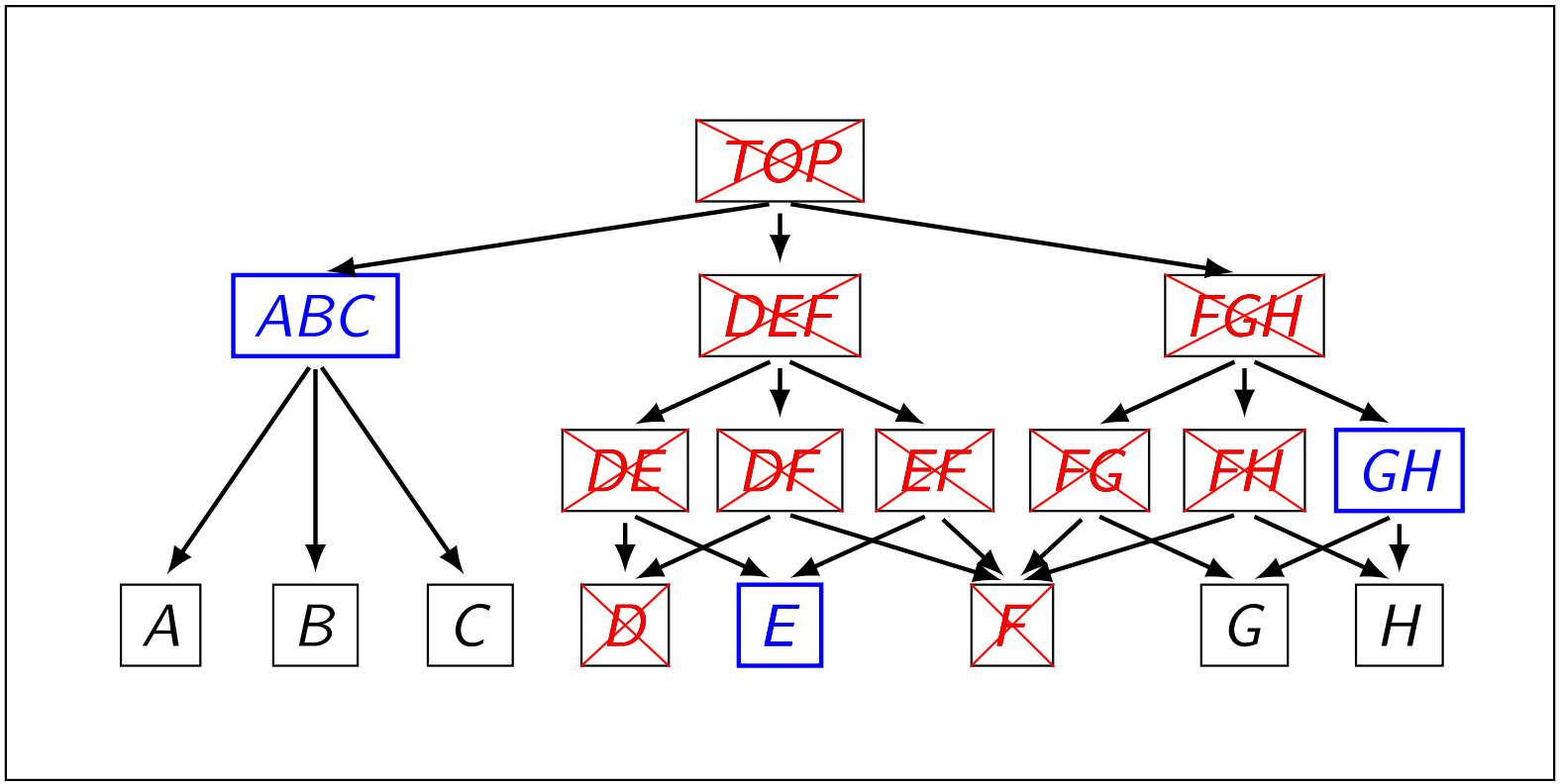
The focus level procedure

Expand subgraphs when necessary



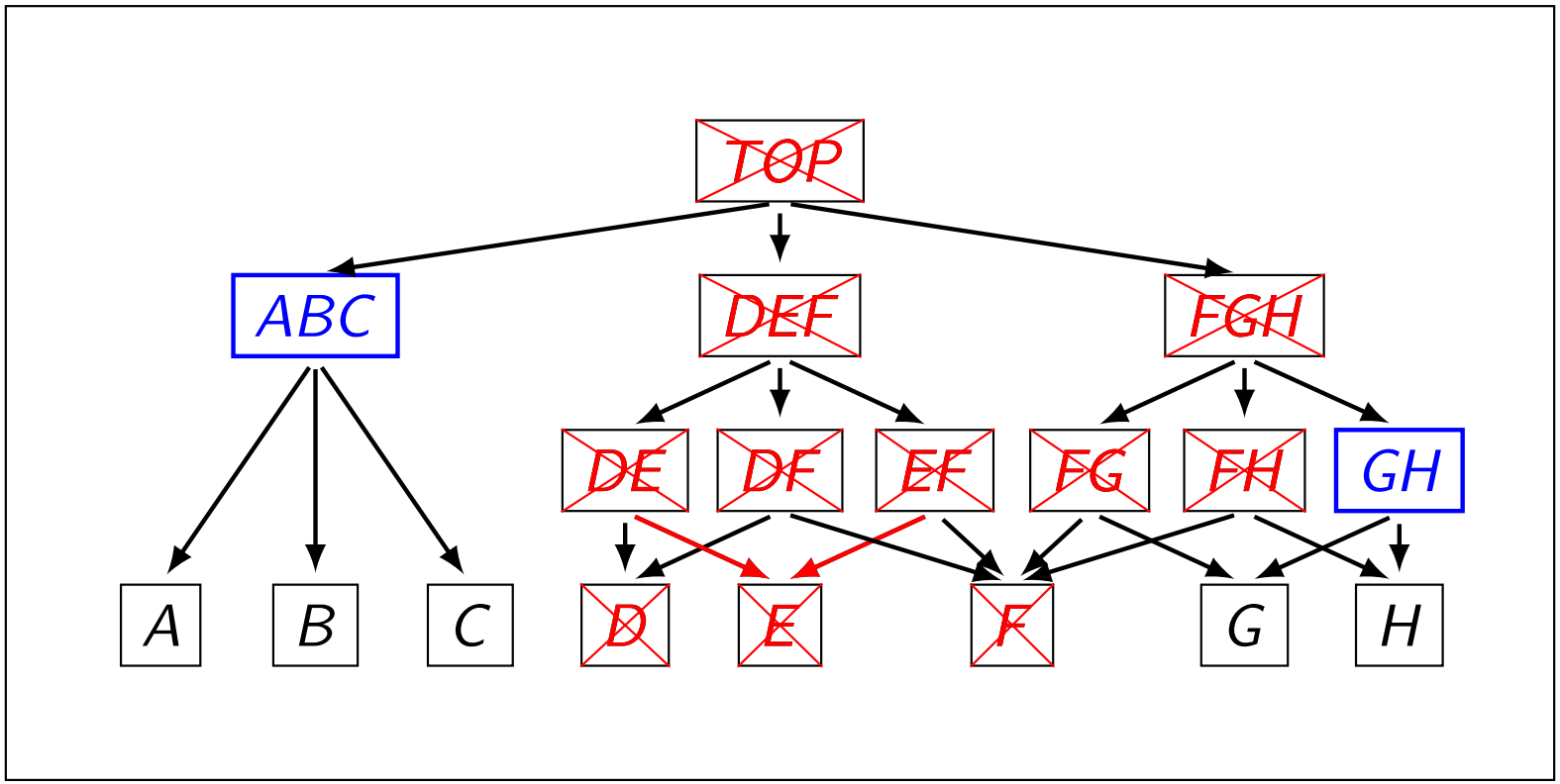
The focus level procedure

Go on until no more significant sets can be found



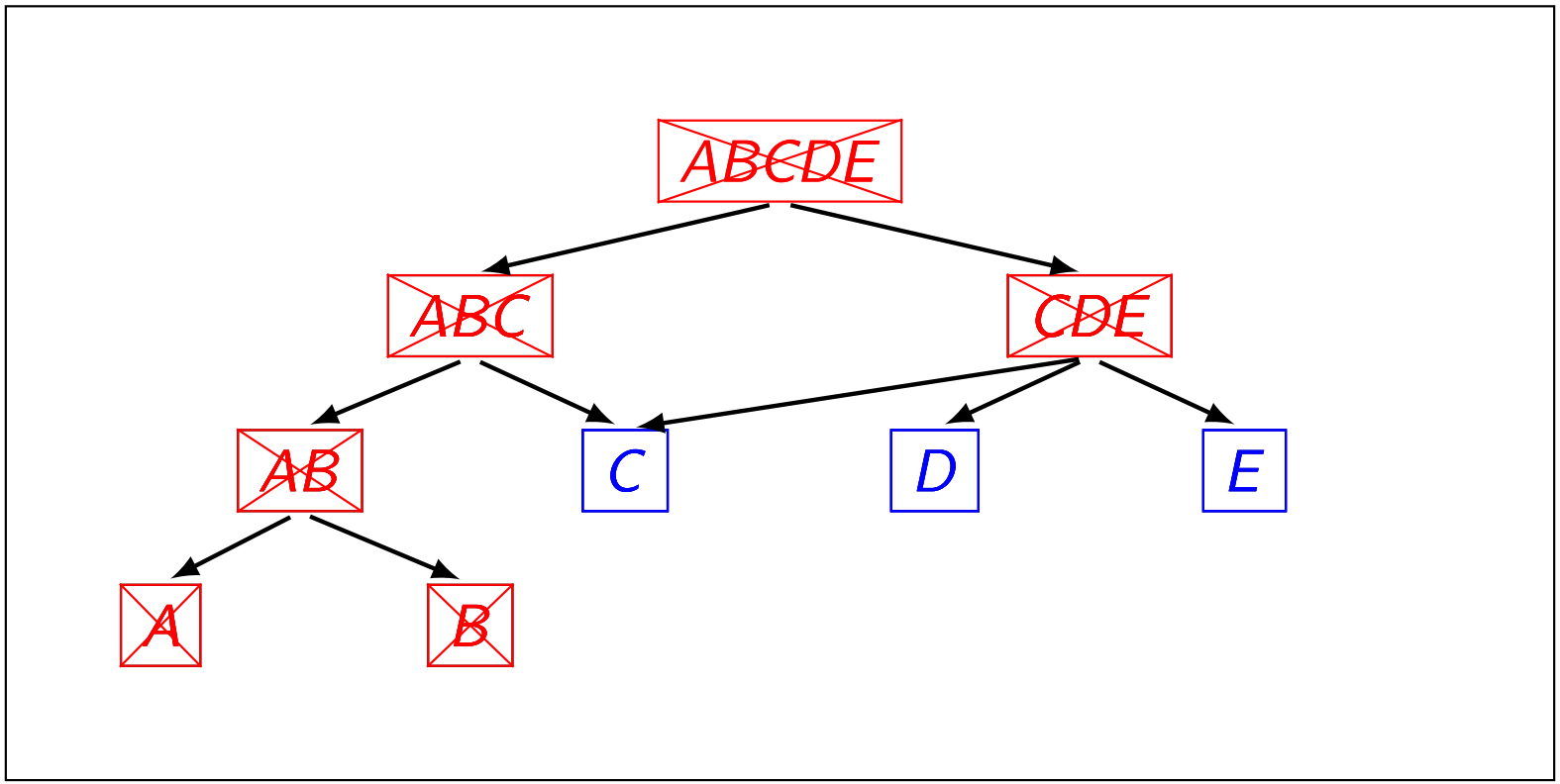
The focus level procedure

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



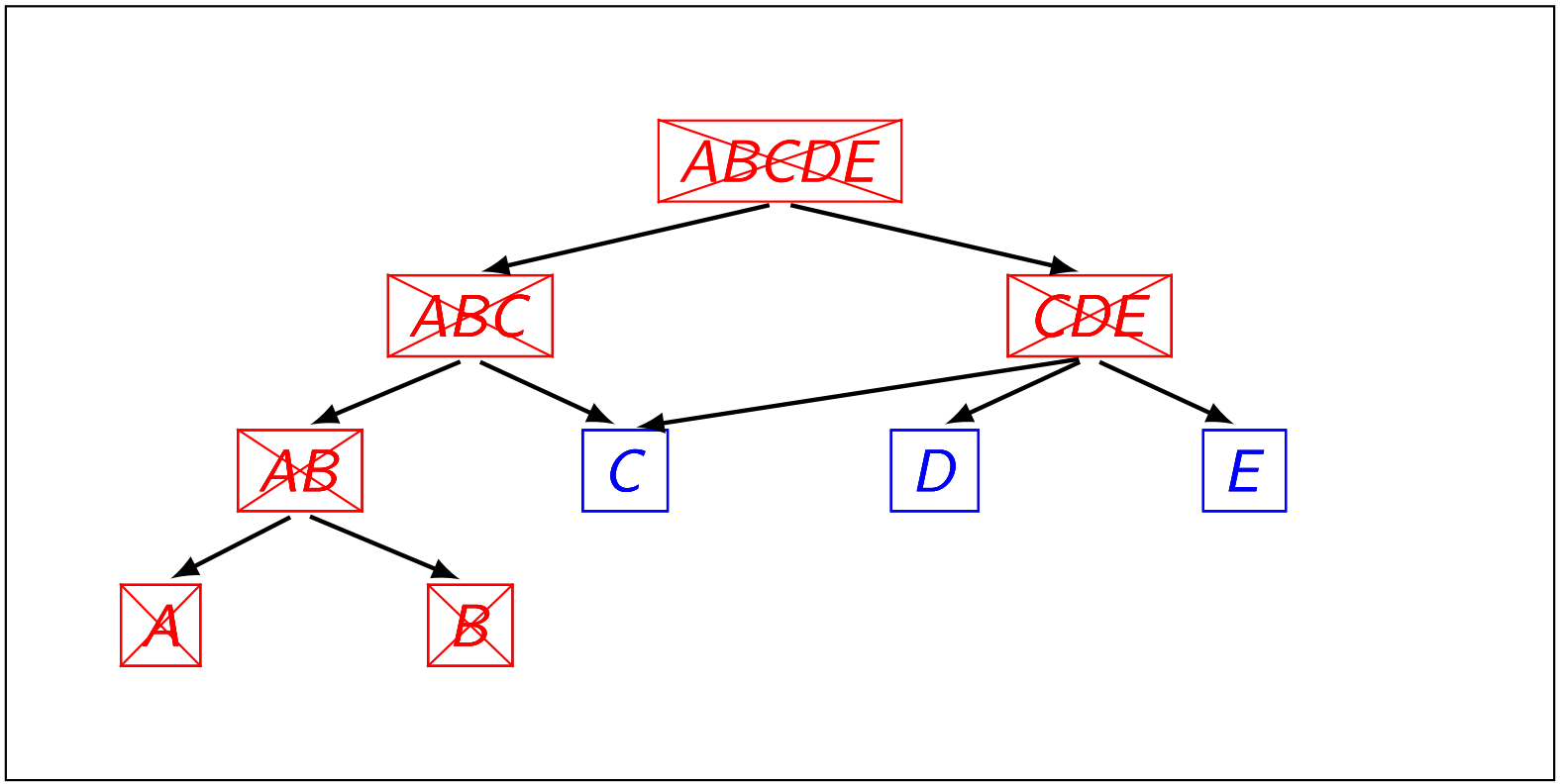
Number of associated genes in a set

Set ABCDE contains with 95% confidence at least 3 non-null genes



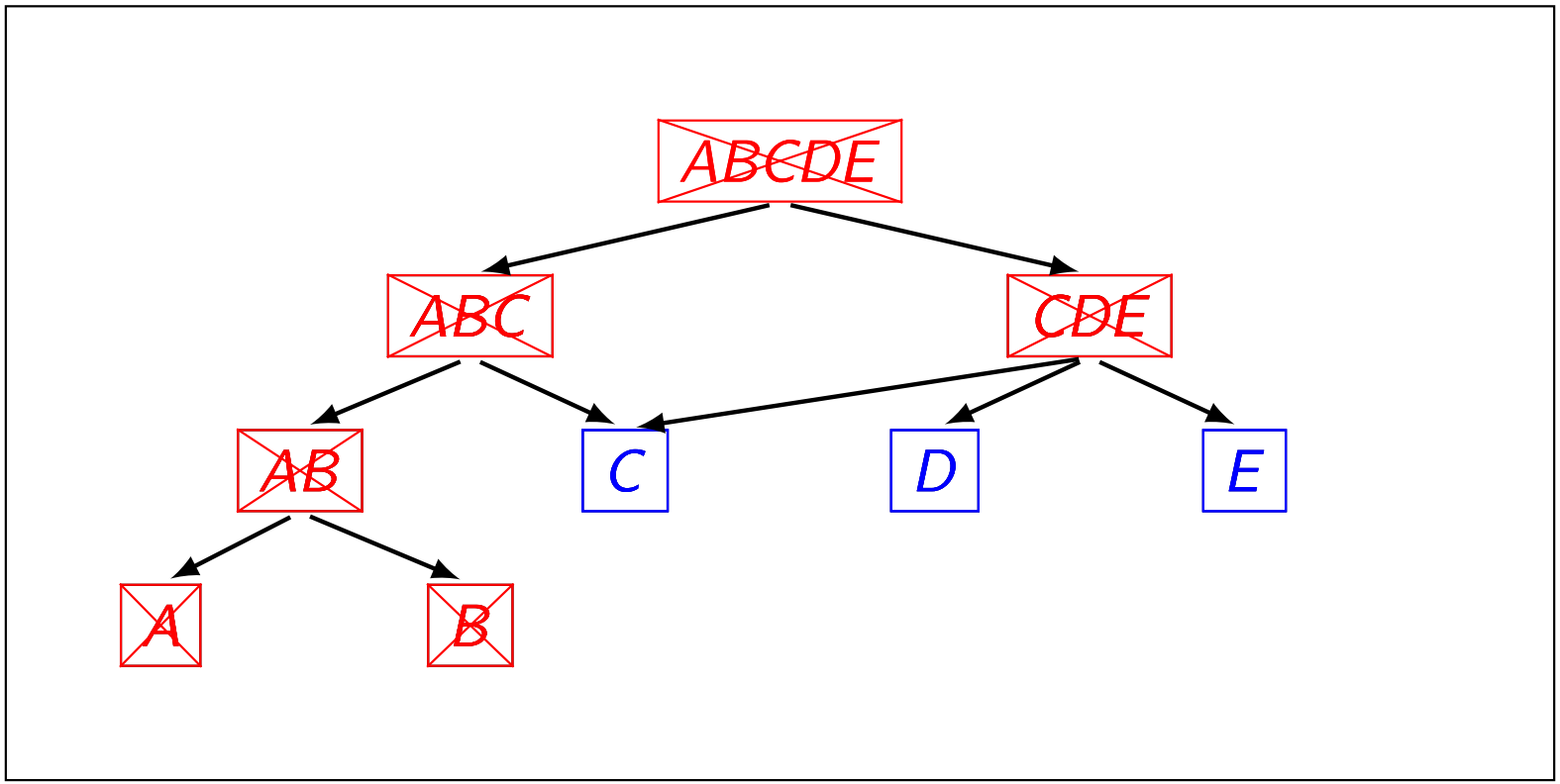
Number of associated genes in a set

Set BCDE contains with 95% confidence at least 2 non-null genes



Number of associated genes in a set

These confidence statements are simultaneous for all sets



Algorithms and software

Algorithms

Finding the correct alpha-levels requires sophisticated algorithms

- exact: integer linear programming (NP-hard)
- approximate: linear programming (polynomial)

`globaltest` package (Bioconductor)

Focus level method

`cherry` package (cran)

Top down method and structured Holm

Which method to use when?

Power against what alternative?

Huge complexity in possible alternatives

Best method

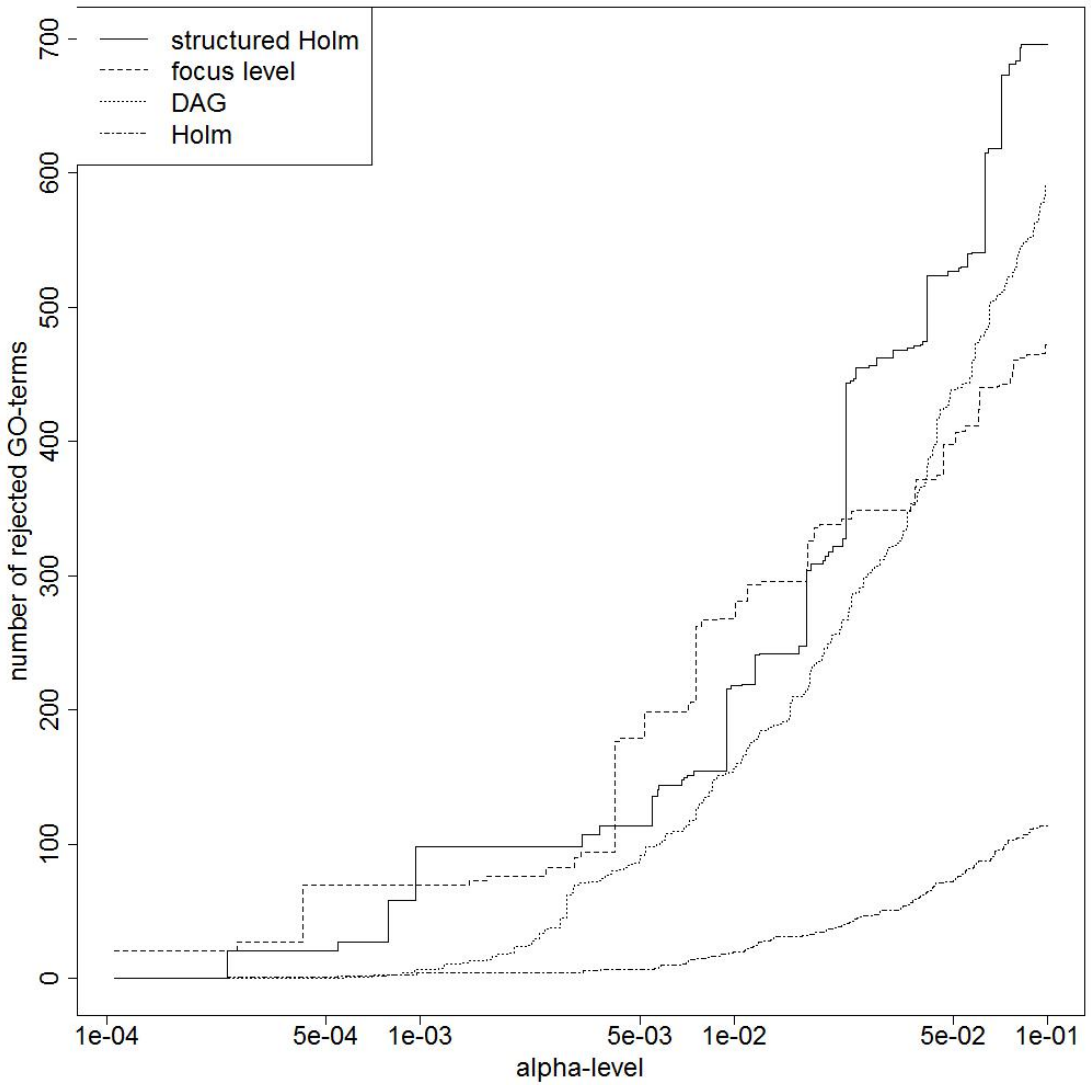
Intuitively: put most alpha immediately on test with most power

- Top down: widespread weak effects
- Focus level: effects concentrated in a branch
- Structured Holm: best guess if no information available

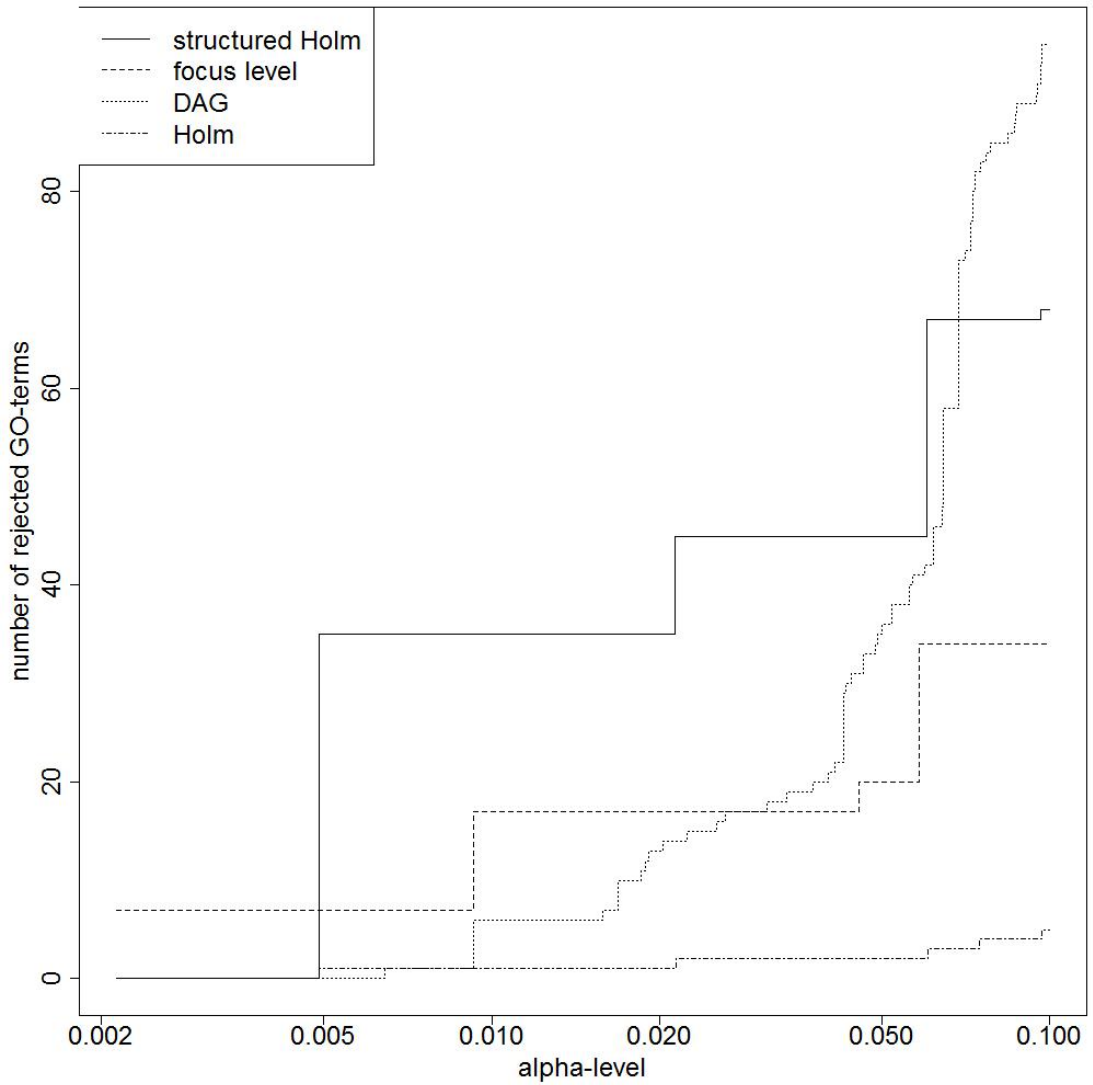
Is this true in practice?

Difficult to simulate or check

Comparison: VDX data



Comparison: Mainz data



Summary: testing in DAGs

DAGs are not exchangeable

Difficult to work with FDR or FDP over a DAG

The sequential rejection principle

- Facilitates formulation of FWER controlling procedures
- Facilitates proof of FWER control
- Allows use of logical relationships
- Easy to formulate multiple testing methods for graph-structured hypotheses

Many options for procedures

Difficult to decide which procedure is best for which situation?

Read more?


Goeman and Mansmann (2008).
Multiple testing in the DAG of gene ontology
Bioinformatics, **24** (4), 537-544.


Goeman and Solari (2010).
The Sequential Rejection Principle of Familywise Error Control.
Annals of Statistics 38 (6) 3782-3810.


Meijer and Goeman (2015a).
A multiple testing method for hypotheses structured in a DAG
Biometrical Journal, **57** (1), 123–143.


Meijer and Goeman (2015b).
Multiple testing of gene sets from Gene Ontology
Briefings in Bioinformatics, in press.