LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
IBE INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG BIOMETRIE UND EPIDEMIOLOGIE
DIREKTOR: PROF. DR. RER. NAT. ULRICH MANSMANN

Testing Groups of Genes

Part II: Scoring Gene Ontology Terms

Manuela Hummel, LMU München
Adrian Alexa, MPI Saarbrücken

NGFN-Courses in Practical DNA Microarray Analysis
Heidelberg, March 6, 2008

Biological questions

max planck institut informatik

➢ **Main idea:**

- If you look for candidate genes correlated with a given phenotype it is better to look for interesting gene groups first.
- Grouping the genes into biological predefined clusters can be seen as a filtering: genes from the same group share the same biology.

➢ **Analysis steps:**

1. Derive score for genes (p-value, t-statistic, even gene expression value itself).
2. Map genes to biological groups and compute significance of these groups using a suitable test statistic.
3. Screen the significant biological groups for candidate genes.

➢ **Advantages:**

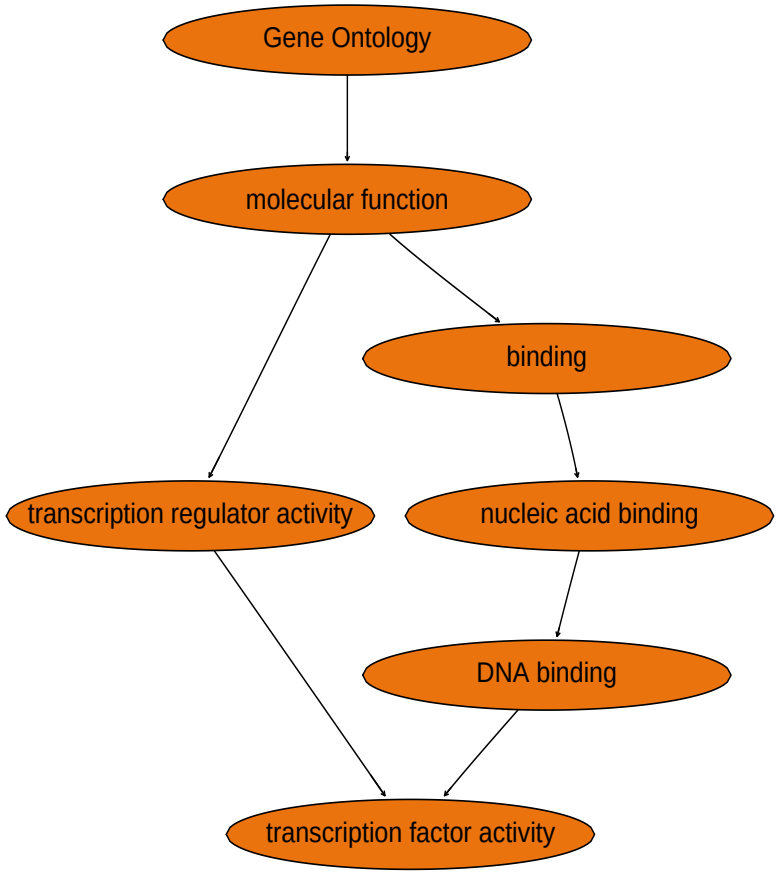
- Easier to find biologically related genes sharing the same pattern.
- Fewer groups to be investigated for differential expression than individual genes.
- Easier to find genes with sensible small change in expression.

Gene Ontology

- The Gene Ontology (GO) is a controlled vocabulary to describe gene and gene product attributes (*http://www.geneontology.org/*)

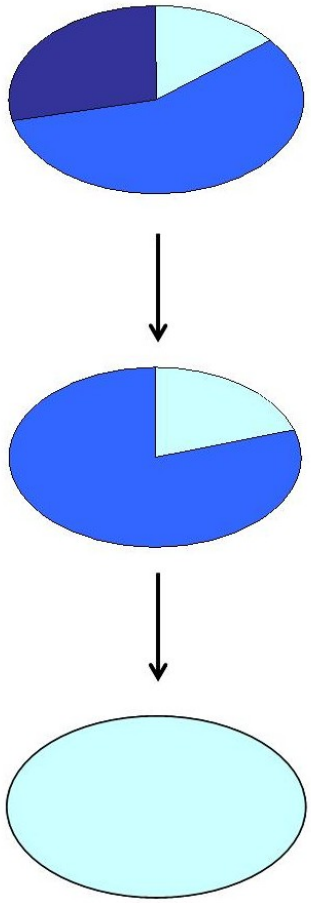
- Three Ontologies
  Molecular Function (7825 terms)
  Biological Process (13860 terms)
  Cellular Component (1993 terms)

- Relations between GO terms are displayed in directed acyclic graphs



Gene Ontology

- Genes known to be associated with some attributes are mapped to corresponding GO terms
- Inheritance
  Each gene associated with some term is also mapped to all its ancestors
- Overlap exists also between unrelated terms
- Not every gene belongs to a leave node
  $\{genes\,in\,the\,leaves\} \neq \{genes\,in\,the\,root\}$



GO Analysis

- Most current tools for GO analysis use tests based on Gene Set Enrichment
  Khatri and Draghici (2005), Rivals et al. (2006)

- Testing thousands of GO terms requires some adjustment for multiple testing

- Recent approaches incorporate the special structure of the Gene Ontology

  – Decorrelating the GO (`elim`, `weight`), *Alexa et al. (2006)*

  – Parent-child approach, *Grossmann et al. (2007)*

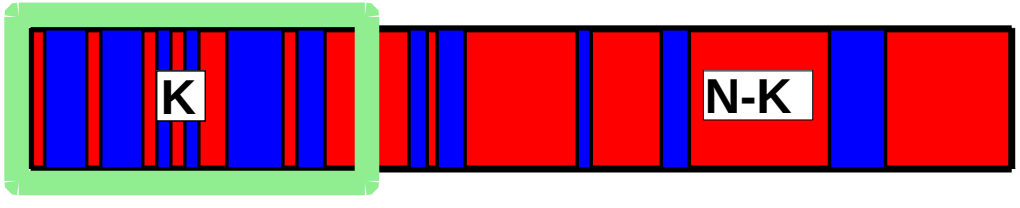
  – Focus-level approach, *Goeman and Mansmann (2008)*

Gene sets enrichment

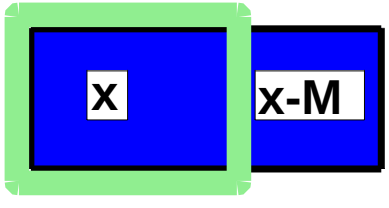
max planck institut informatik

- **Group enrichment:** given a **gene group** with some biological function, analyse the positions of these genes in the **ordered list**. The **gene group** is relevant, if all genes are among the top genes in the **ordered list**.
- **Idea:** Sort genes according to some score (diff. expression) and investigate the ranks of the members of group **A** (the biological function) in this list.
- Define cutoff and count members of group **A** below and above cutoff. Basically, one wants to compare the following ratios:

$$\frac{\mathbf{K}}{\mathbf{N}} \leq \frac{\mathbf{x}}{\mathbf{M}}.$$



N (gene on the microarray)



M (genes in group)

GO scoring: general problem

max planck institut informatik

Given:

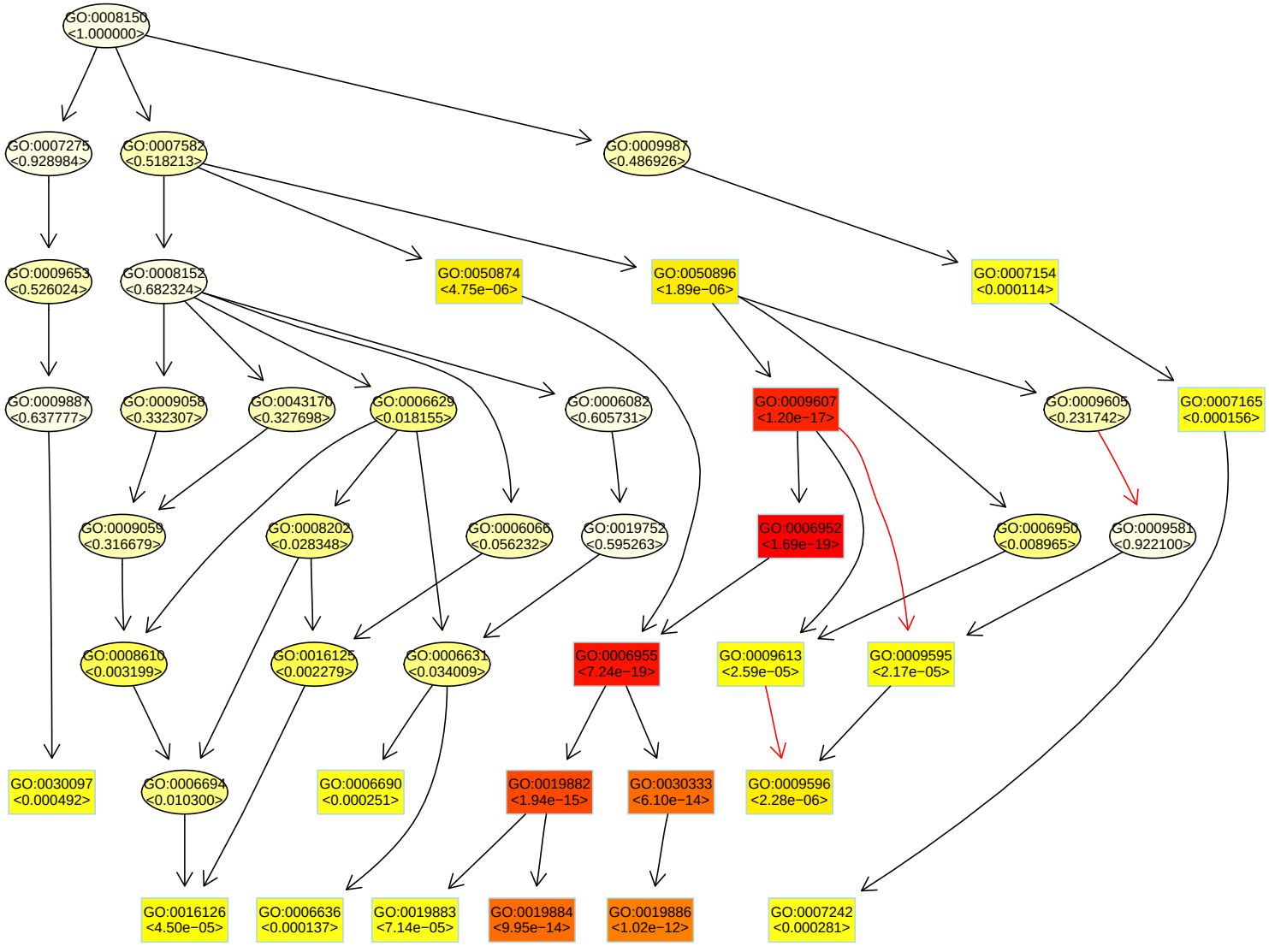
- a directed acyclic graph (GO graph) and a set of items (genes) s.t.:
  - each node in the graph contains some genes
  - the parent of a node contains all the genes of its child
  - a node can contain genes that are not found in the children
- a subset of genes that we call significant genes (differentially expressed genes)

Goal:

- find the nodes from the graph (biological functions) that best represent the significant genes w.r.t some scoring function (some test statistic)

GO independence assumption

max planck institut informatik



Note: The coloring of the nodes represent the *relative* significance of the GO terms: dark red is the most significant, light yellow is the least significant from the graph

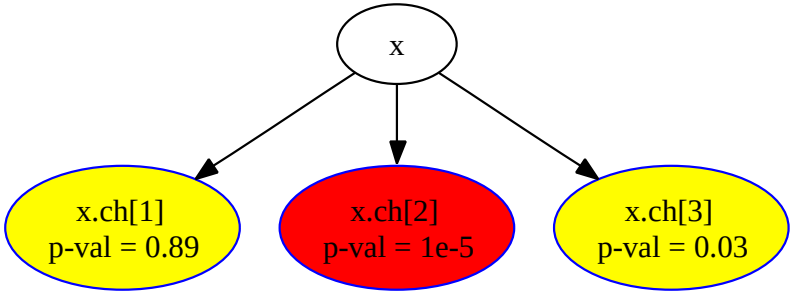
The elim method

max planck institut informatik

The main idea: Test how enriched node x is if we do not consider the genes from its significant children ($x.ch[2]$ in our case).

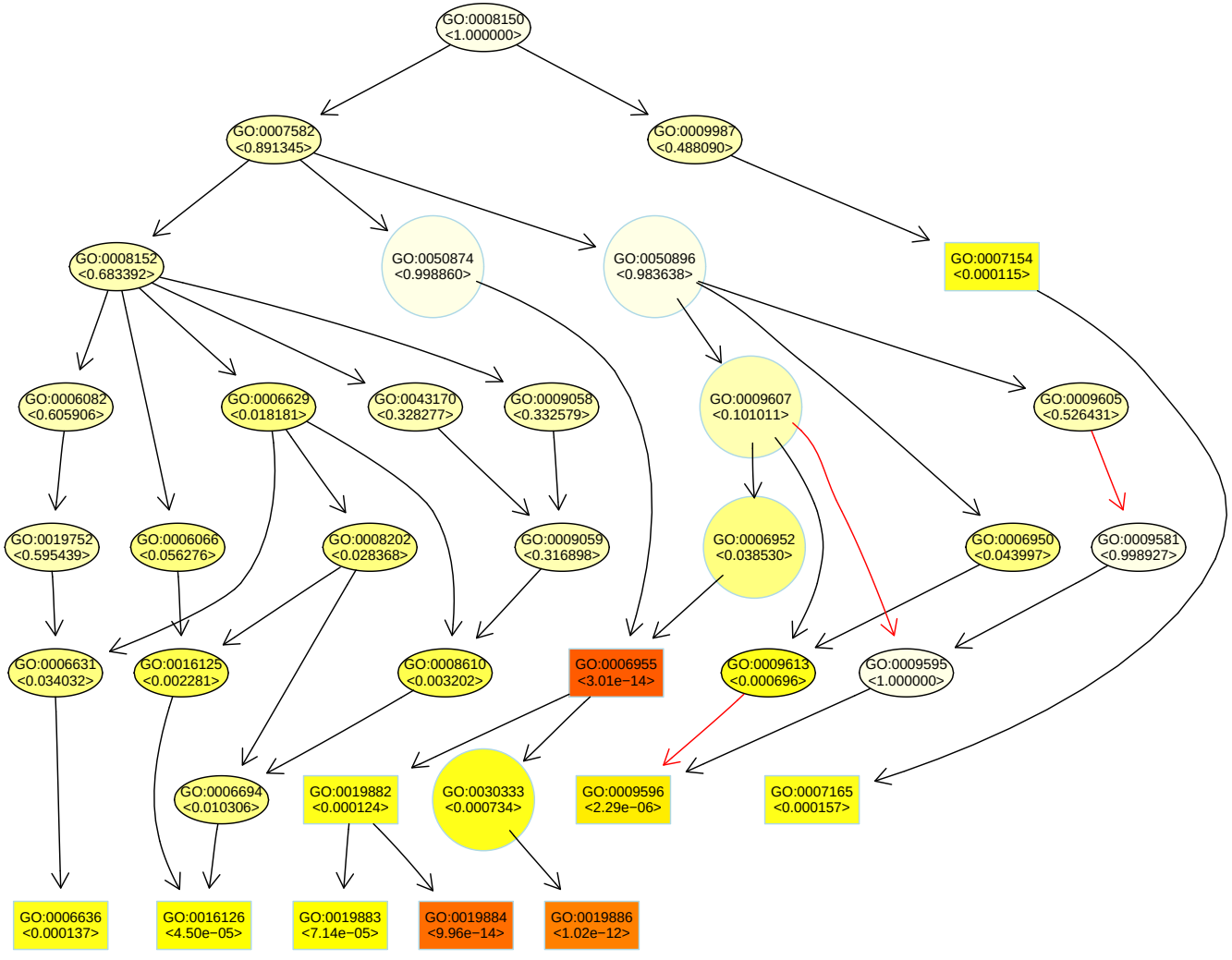
Algorithm:

1. The nodes are processed bottom-up. This assures that all children of node x were investigated before node x itself.
2. Let $removed(x)$ be the set of genes that were removed in a previous step by a node in the lower subgraph induced by node x. Then
   $$genes(x) \leftarrow genes(x) - removed(x).$$
3. The p-value for node x is computed using Fisher's exact test.
4. If node x is found significant, we remove all the genes mapped to this node, from all its ancestors.



elim **result**

max planck institut informatik



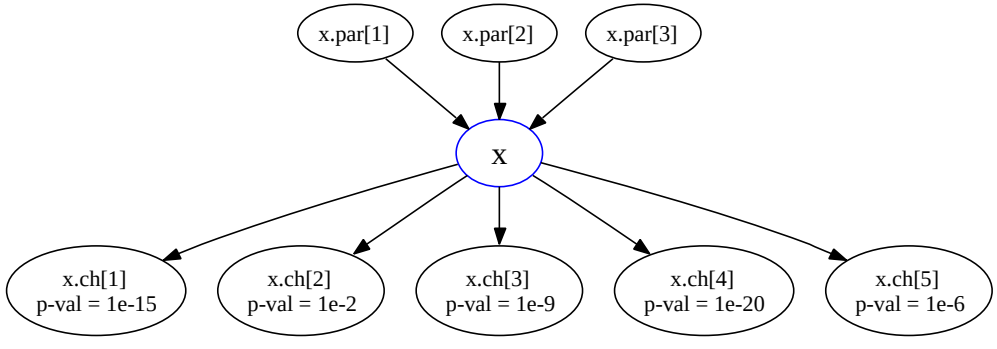
Top 10 significant node (the boxes) obtained with method elim

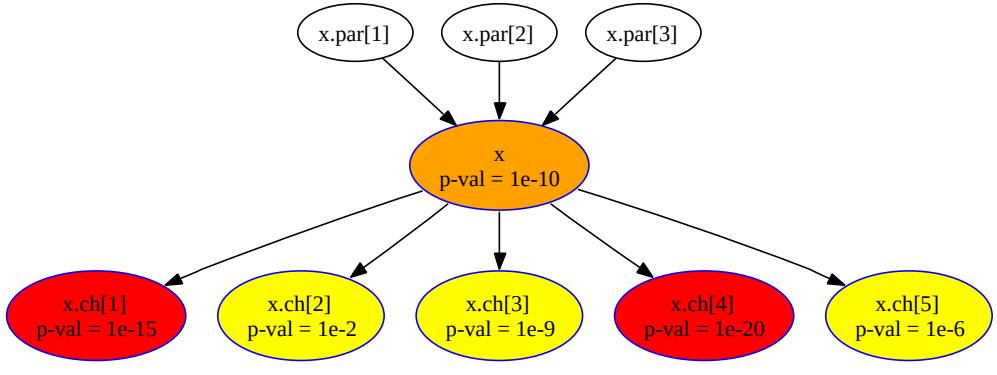
The weight **method**

- We want to decide if node x is better representing the list of interesting genes (is more enriched) than any other node from its neighborhood.
- The main idea: Associate single genes mapped to a node with weights that denote their relevance. The elim algorithm uses 0-1 weights.

Algorithm:

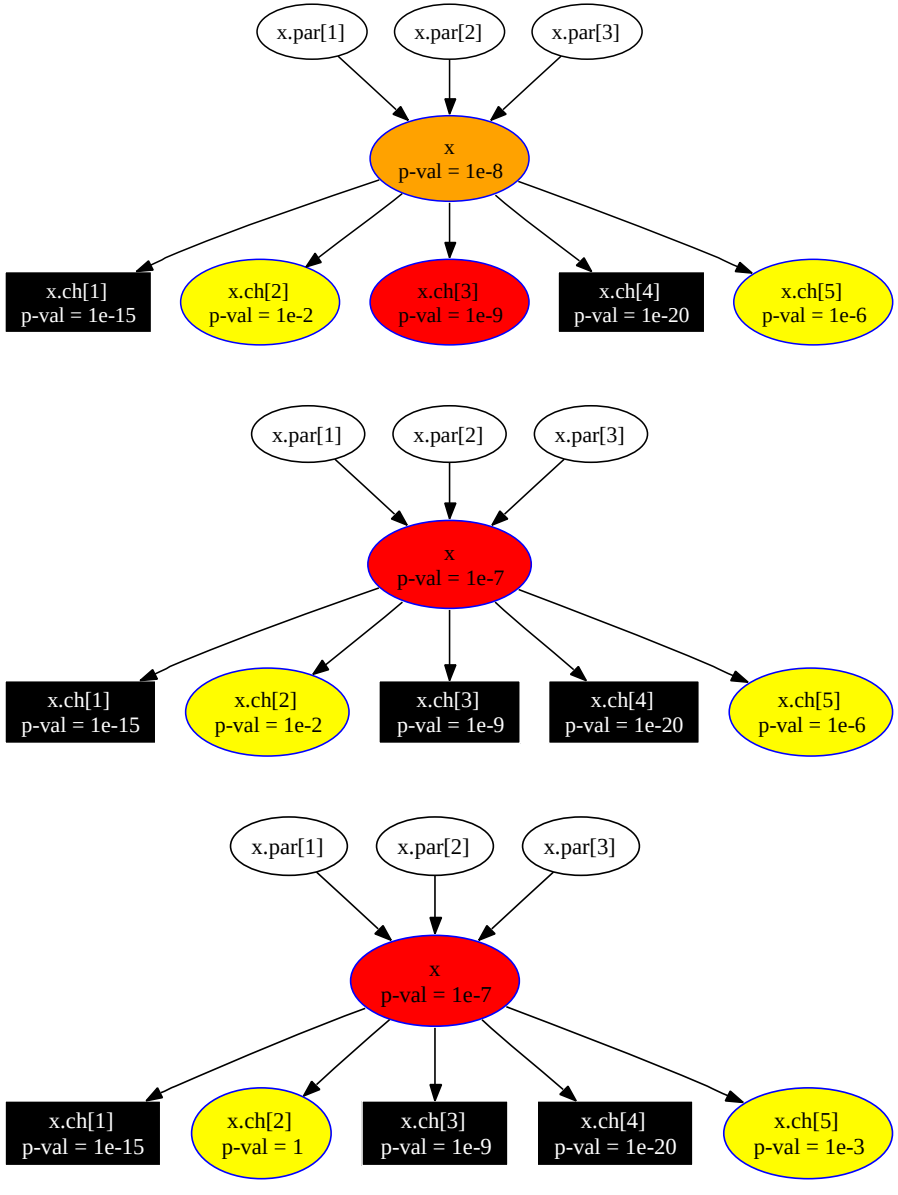
1. Compute the p-value of node x with its current weights. Initially all its genes have weight 1.
2. **CASE I:** Look at the children that are more significant than node x ($x.ch[1]$ and $x.ch[4]$). These children are local optima (colored with red).
3. For each such child down-weight all genes mapped to it in all the ancestors of node x, including x. Mark these children and GOTO step 1.





The weight method

4. **CASE II:** If no child of node x has a p-value less than the current p-value of node x then node x is a local optimum.
5. The genes in these children are down-weighted and the p-values for these nodes are recomputed with the new updated weights.
6. The processing of node x terminates. Its p-value can be changed later, when node x is treated as a child of another node.



- The p-value of a node is computed by applying Fisher's exact test on a weighted contingency table. The quantity

$$|sigGenes \cap genes(u)|$$

is replaced with

$$\left\lceil \sum_{i \in \{sigGenes \cap genes(u)\}} weight[i] \right\rceil .$$

- The weights for node x and one of its children are obtained by

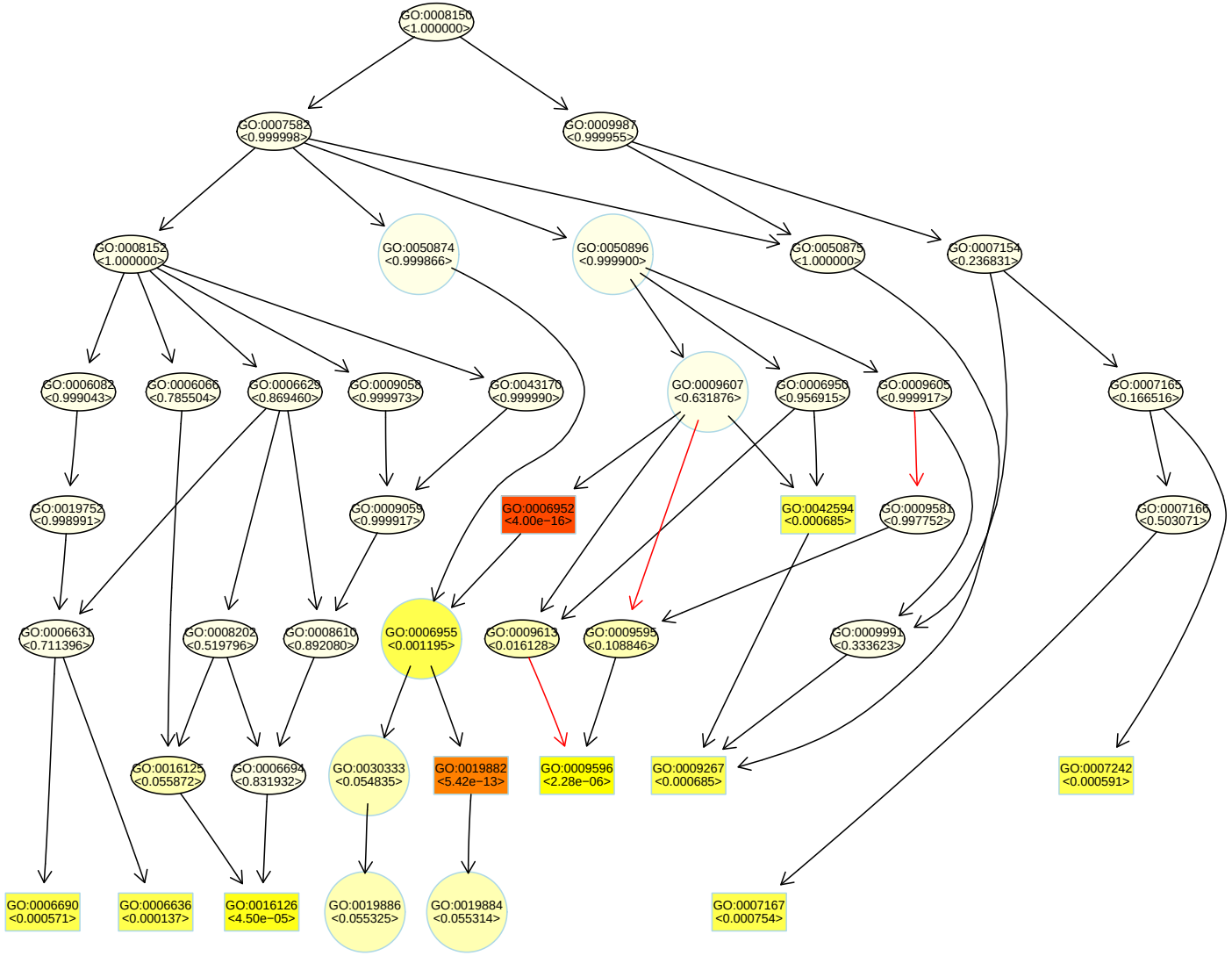
$$\text{sigRatio}(ch, x) = \frac{\log(p\text{-value}(ch))}{\log(p\text{-value}(x))} \quad \text{or} \quad \text{sigRatio}(ch, x) = \frac{p\text{-value}(x)}{p\text{-value}(ch)}$$

If sigRatio() > 1 then node ch is more significant than its parent, node x.

- The weights are updated using vector operators: minimum on the components, the product of the components, etc.

weight **result**

max planck institut informatik



Top 10 significant node (the boxes) obtained with method weight

Algorithms review

➢ classic **algorithm**

- Calculate significance of each GO term independently.
- Adjust pvalues for multiple testing (Bonferroni, FDR, etc.).
- Kolmogorov-Smirnov test can easily be used in this case

➢ elim **algorithm**

- Nodes are processed bottom-up in the GO graph.
- It iteratively removes the genes annotated to significant GO terms from more general GO terms.
- Intuitive and simple to interpret.

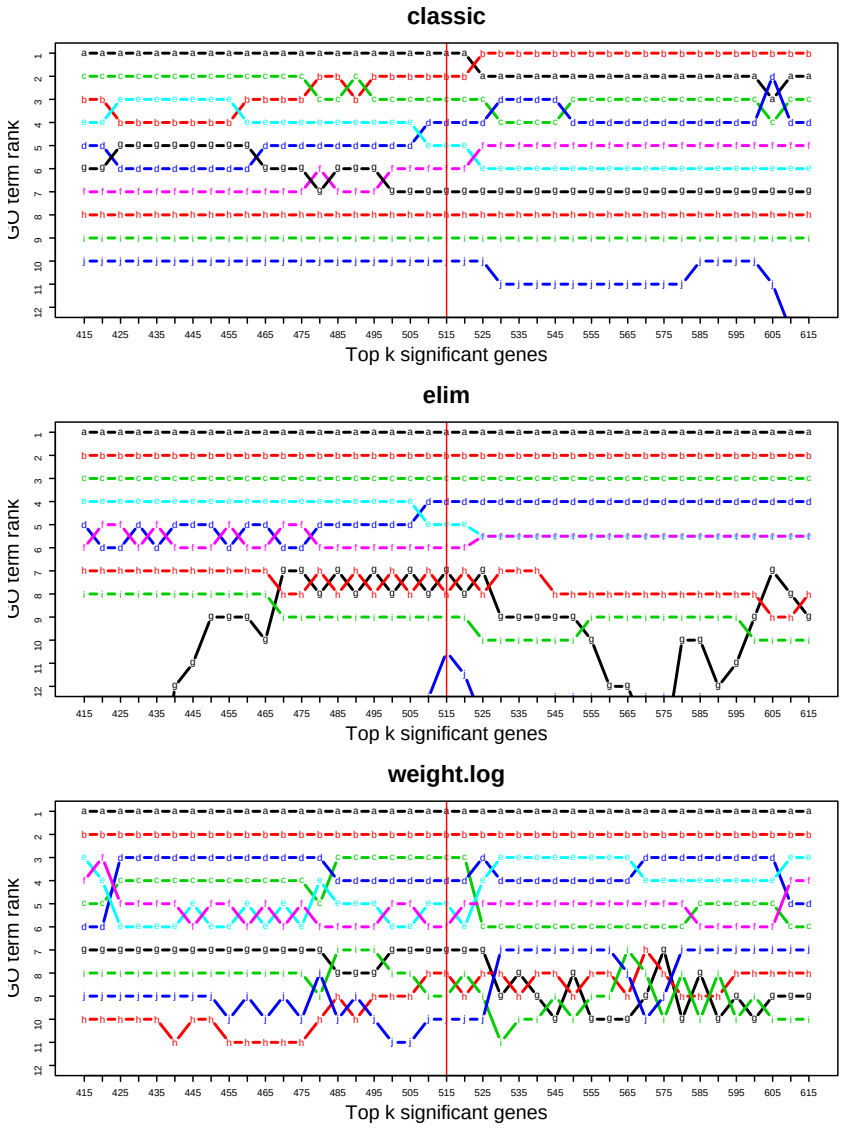
➢ weight **algorithm**

- The genes obtain weights that denote the gene relevance in the significant nodes.
- To decide if a GO term u better represents the interesting genes, the enrichment score of node u is compared with the scores of its children.
- Children with a better score than u better represent the interesting genes; their significance is increased
- Children with a lower score than u have their significance reduced.

Influence of the p-values adjustment

max planck institut informatik

- We had performed a two-stage analysis:
  1. A cutoff is chosen based on the distribution of the genes' scores (p-values adjustment problem). Genes above the cutoff are called *DE genes*.
  2. The enrichment of a set of genes (GO term) is tested based on test statistics that depend on the list of *DE genes*.
- Problem:
  - In real-life cases the list of *DE genes* contains only a small fraction of truly *DE genes*.
  - Is the result of the enrichment analysis hampered by the choice of the cutoff?
- Results:
  - $k = 515$ *DE genes* (all genes with FDR-adjusted p-value $p \leq 0.01$).
  - Variating the cutoff value does not significantly change the order of the most significant GO terms (only small swaps between the GO terms)



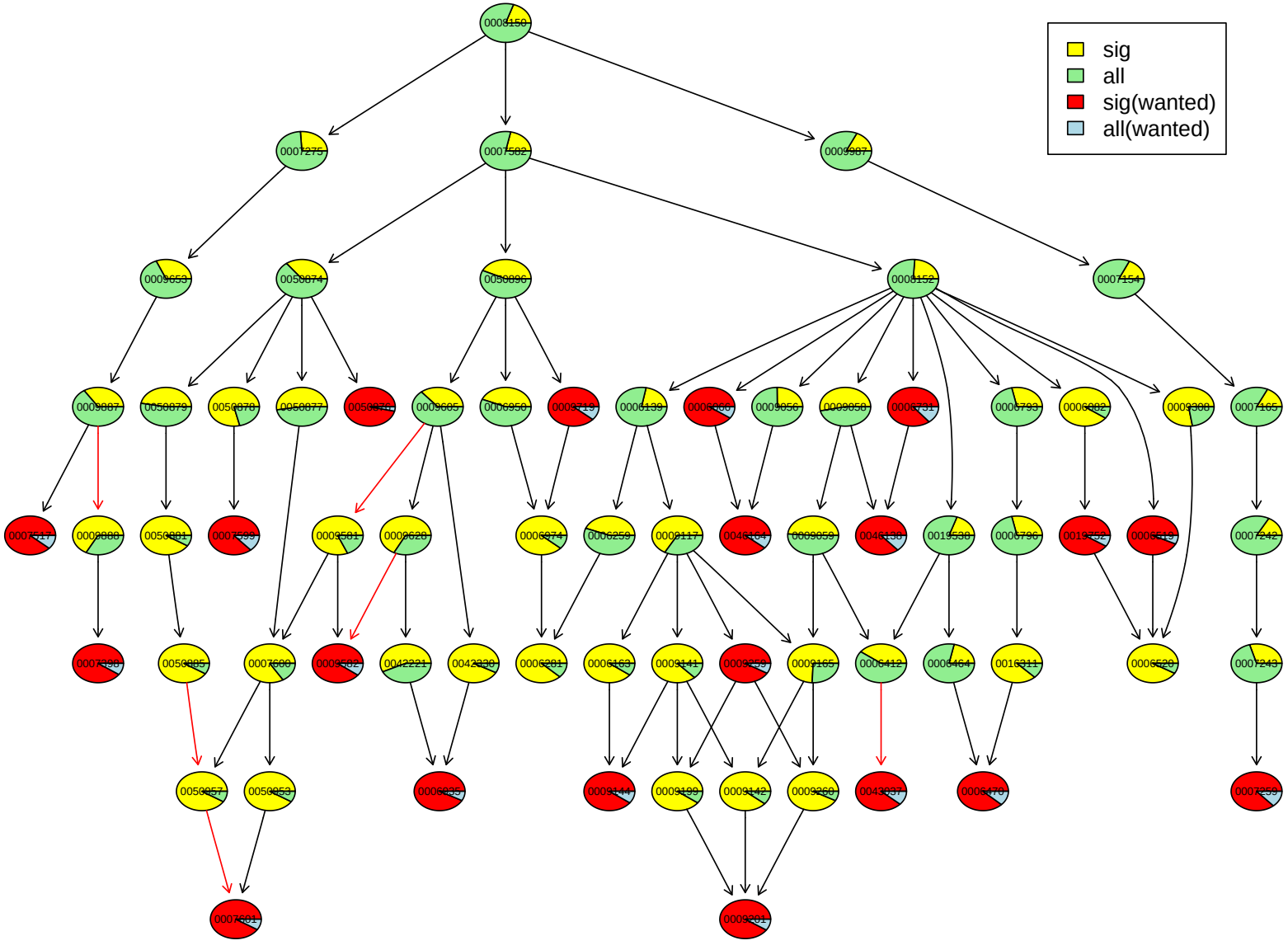
Evaluation on simulated data

max planck institut informatik

- We use the GO graph structure (2311 nodes), and all the genes from HGU95aV2 Affymetrix chip (9623 mapped to the GO graph)
- Select only the nodes that have the no. of mapped genes in some range $(10 \ldots 100)$
- Choose randomly a number of nodes (50 in our case) from the selected nodes. These nodes represent the enriched nodes.
- Set as significant genes all the genes from the enriched nodes.
- Some noise can be introduce:
  - Pick 10% from all significant genes
  - Remove them from the significant list
  - Replace the genes that we removed with other genes
- The goal is to recover as best as possible the enriched nodes.

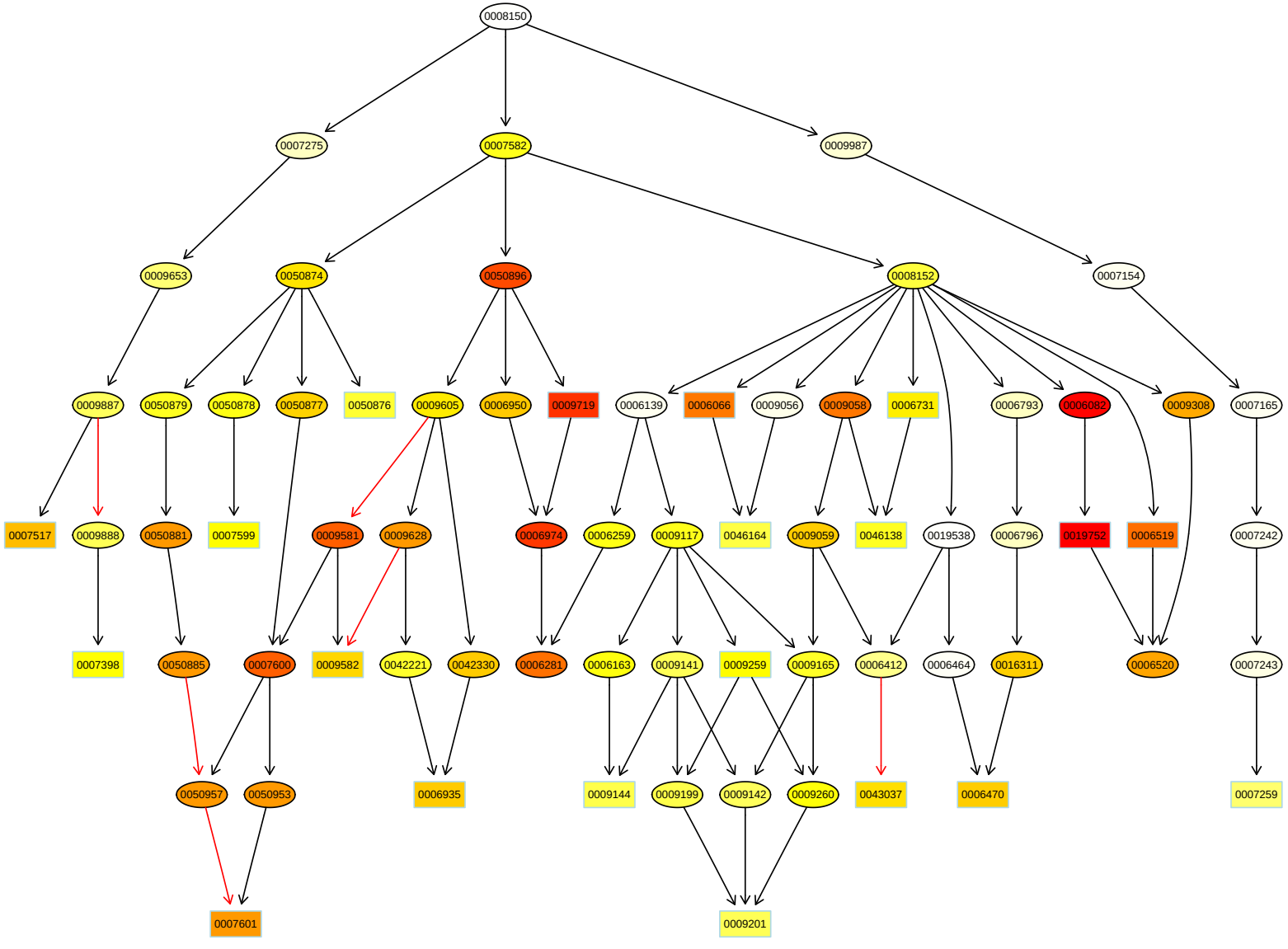
Simulated dataset

max planck institut informatik



Simulated dataset

max planck institut informatik



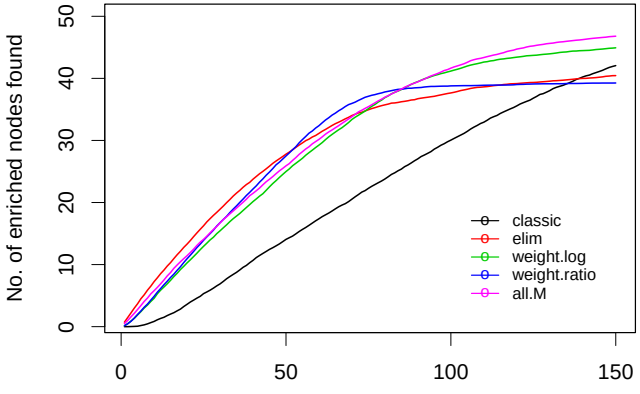
Quality of GO scoring methods

max planck institut informatik

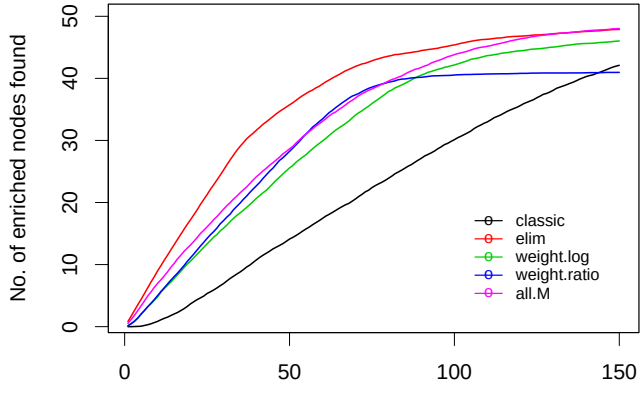
Each curve represents the average of the numbers of preselected GO terms, over 100 simulation runs, that are among the top k GO terms. The left plot represents $score_k^0$ and the right plot represents $score_k^{1p}$.

10 to 50 genes annotated

10% noise level.



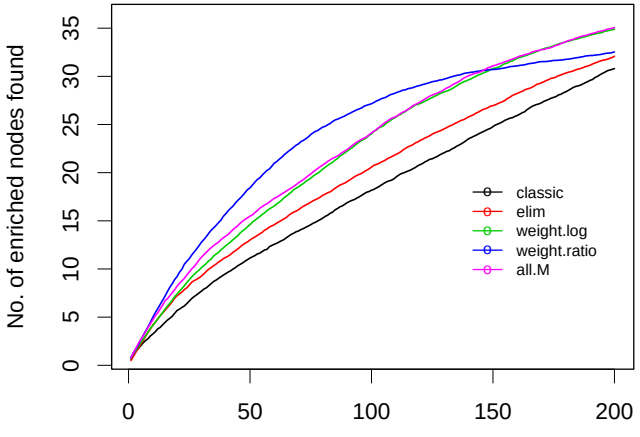
(a) Top k nodes



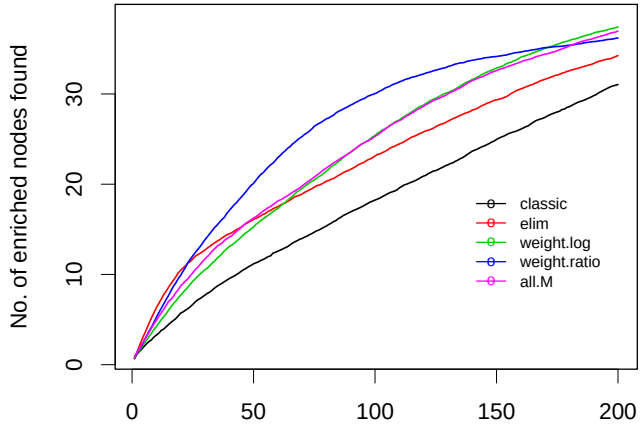
(b) Top k nodes

10 to 1000 genes annotated

40% noise level.



(c) Top k nodes



Top k nodes

Parent-Child Approach

- If many differentially expressed genes are annotated to a GO term it is not surprising that there is also found over-representation in the more specific descendants of the term

- Compute hypergeometric p-values where the reference gene population does not consist of all genes m but rather of only all parental genes $m_{pa(t)}$ of a given GO term t

$$P(X_t \geq x_t | X_{pa(t)} = x_{pa(t)}) = \sum_{k=x}^{\min(x_{pa(t)}, m_t)} \frac{\binom{m_t}{k}\binom{m_{pa(t)} - m_t}{x_{pa(t)} - k}}{\binom{m_{pa(t)}}{x_{pa(t)}}}$$

$pa(t)$: set of parents parents of term t

m_t: number of genes annotated to term t

$m_{pa(t)}$: nr. of genes in either *union* or *intersection* of genes annotated to parents of t

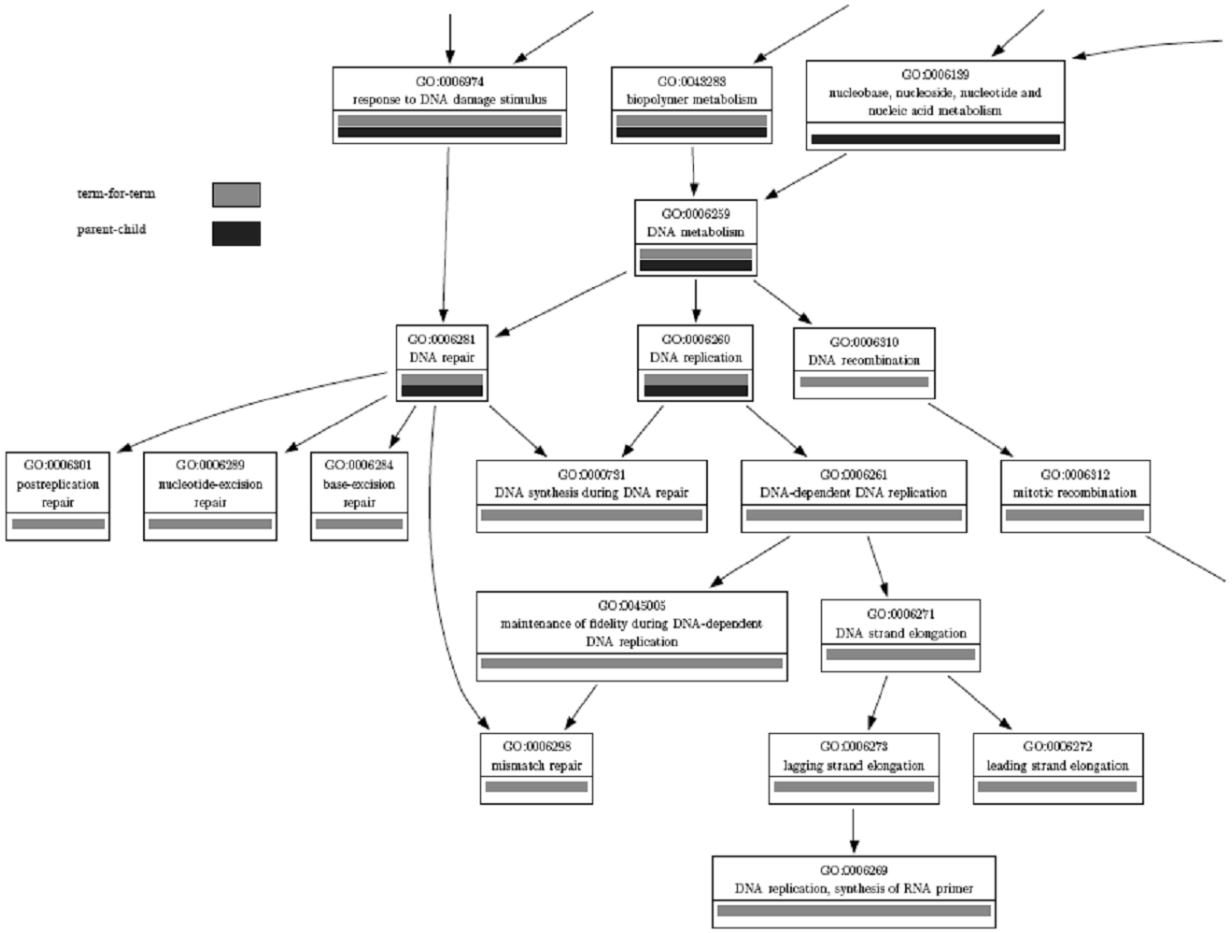
x_t: number of differentially expressed genes annotated to term t

Grossmann et al. (2007)

Parent-Child Approach

- Idea is reverse to `elim` and `weight`: Children nodes might only *inherit significance* from their more general parents

- Focus lies in more general terms

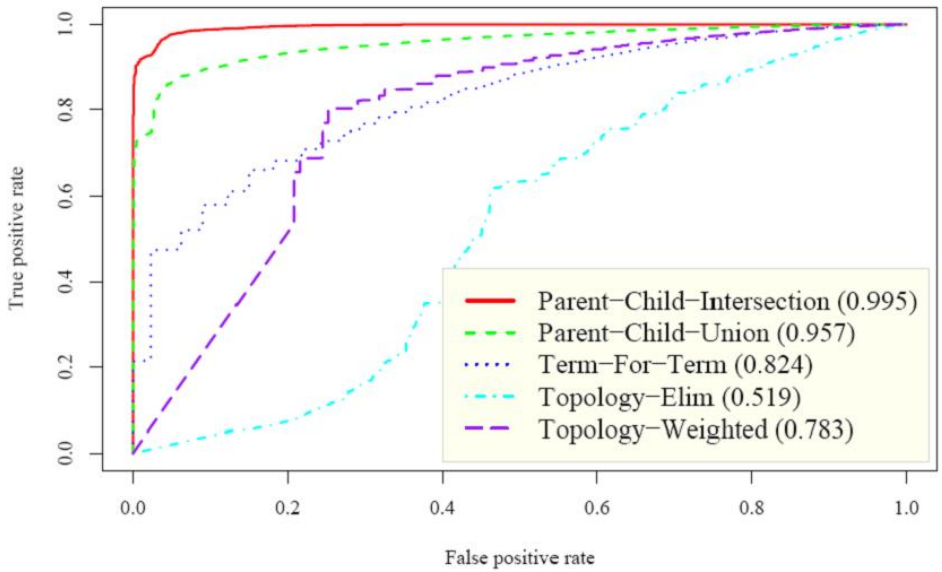


Simulation Study

Similar simulation setup as in *Alexa et al. (2006)*, but

- Pre-selection of terms that actually can achieve a small p-value with the parent-child approach
- Overrepresentation of just one term (out of the preselected)

ROC analysis



→ How to design an objective simulation study . . .?

Focus Level Approach

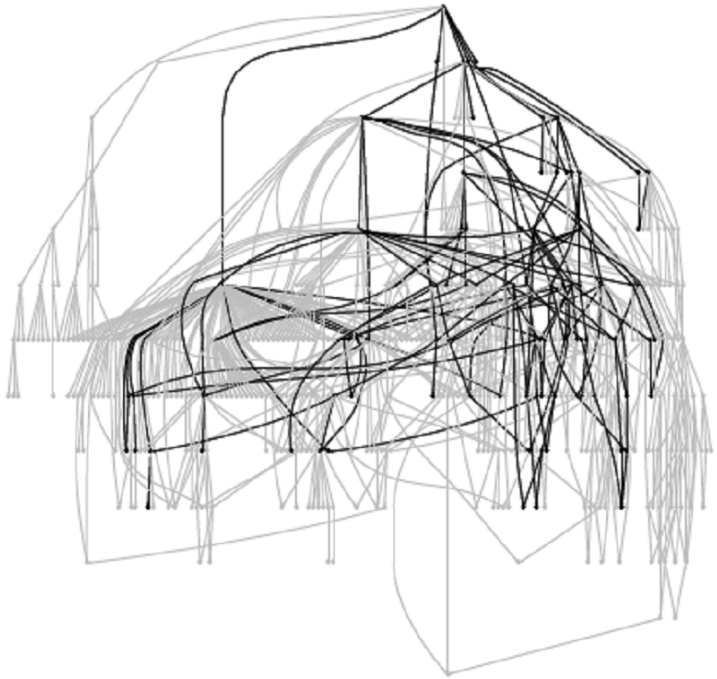
- Again a different idea: Significant terms logically *must* have significant ancestor terms
- Relevance of terms is assessed by global tests
  (e.g. `globaltest` or `GlobalAncova`)
- Multiple testing procedure on the Gene Ontology graph which controls the *family-wise error rate* (FWER): Combines closed testing procedure with correction method of Holm
- Holm correction: very fast but not very efficient
  Closed testing procedure: very efficient in case of correlated test statistics but computationally infeasible

Focus Level Approach

- Choose a focus level – a set of terms H in the middle of the GO graph (as the level of detail that is of most interest)
- Taking each of the terms in H as root nodes, build subgraphs that are closed under intersection
- Iterate:
  1. Test phase: Test the GO terms in H with global tests and correct raw p-values by a Holm's factor (initially $|H|$)
  2. Upward phase: For every hypothesis rejected in the test phase, reject all ancestors
  3. Downward phase: Add those terms to H, for which all parent hypotheses in the closed subgraphs have been rejected
  4. Holm's phase: Recalculate Holm's factor as the number of subgraphs which contain unrejected hypotheses

Focus Level Approach

- Result is a significant subgraph starting from the root
- Leave nodes in the subgraph usually are of most interest



References

1. Alexa A, Rahnenführer J, Lengauer T. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. Bioinformatics 2006; 22(13): 1600-1607.

2. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. Nature Genetics 2000; 25: 25-29.

3. Goeman JJ, Mansmann U. Multiple testing on the directed acyclic graph of Gene Ontology. Bioinformatics 2008; 24(4): 537-544.

4. Grossmann S, Bauer S, Robinson PN, Vingron M. Improved detection of overrepresentation of Gene-Ontology annotations with parent-child analysis. Bioinformatics 2007; 23(22): 3024-3031.

5. Khatri P, Draghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 2005; 21(18): 3587-395.

6. Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test? Bioinformatics 2007; 23(4): 401-407.