From a gene list to biological function

– *Scoring Gene Ontology terms* –

Adrian Alexa

alexa@mpi-inf.mpg.de

Computational Biology and Applied Algorithmics

Max Planck Institute for Informatics

D-66123 Saarbrcken

Courses in Practical DNA Microarray Analysis, Saarbrücken, September 22, 2005

- Gene sets enrichment
- Scoring GO Terms
- Topology based GO Terms scoring
- Evaluation of scoring methods

- **Gene sets enrichment**
- Scoring GO Terms
- Topology based GO Terms scoring
- Evaluation of scoring methods

- The Microarray experiments provide a **long list of genes**.
- Typical studies analyze genes one by one:
  1. samples are divided into two groups: disease vs. healthy and the genes are ranked according to differential expression.
  2. genes are ordered according to correlation of the expression values with a phenotype measurement.

  These studies result in an **ordered list** of genes.

- **More important is the group enrichment:**
  - given a set of genes with some biological function, analyze the positions of these genes in the ordered list.
  - the biological function is relevant, if all genes are among the top genes in the ordered list.

- Gene sets:
  - Gene Ontology (GO) terms
  - Metabolic pathways
  - MIPS classes
  - Chromosomes
  - Classes defined via transcription factors
  - Gene sets obtained from other previous experiments
- Remark 1:

  The score and the gene set must be chosen independently!
- Remark 2:

  The dependence between gene sets usually make the statistical interpretation of the result harder!

Main idea: Sort genes according to some score and analyze positions of members of the investigated gene group in this list.

- We want to know if the members of group **a** have significantly small ranks (higher in the list). If this is the case, then group **a** is enriched.
- There are basically two approaches:
  1. Define cutoff and count members of group **a** below and above cutoff (parametric test statistic).
  2. Analyze distribution of all ranks of members of group **a** (non-parametric test statistic).

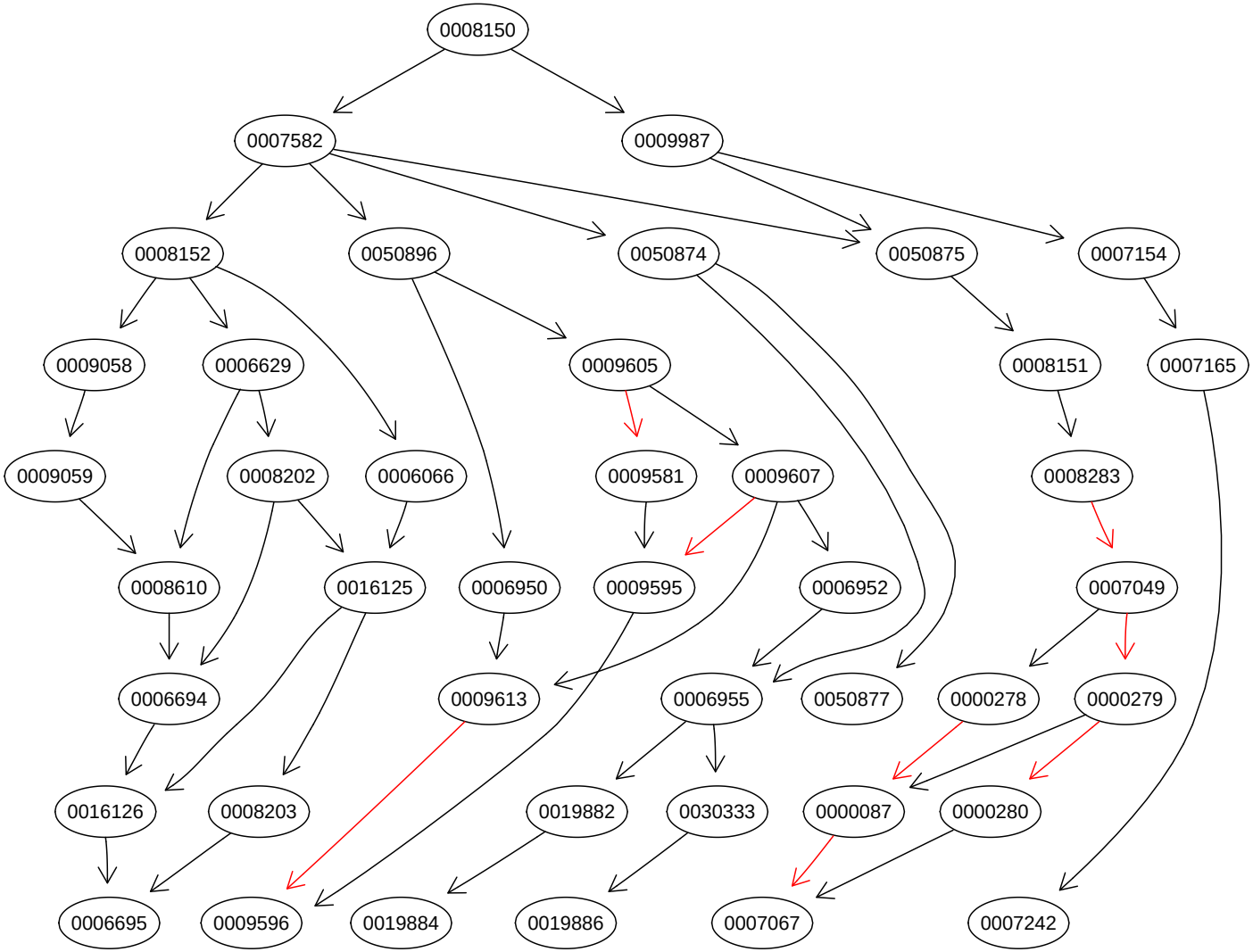
Gene	Score	Group
$\text{gene}_{\sigma(1)}$	score 1	a
$\text{gene}_{\sigma(2)}$	score 2	b
$\text{gene}_{\sigma(3)}$	score 3	a
$\text{gene}_{\sigma(4)}$	score 4	a
.....	.....	.....
$\text{gene}_{\sigma(100)}$	score 100	b
$\text{gene}_{\sigma(101)}$	score 101	a
.....	.....	
$\text{gene}_{\sigma(9905)}$	score 9905	b

- Gene sets enrichment
- **Scoring GO Terms**
- Topology based GO Terms scoring
- Evaluation of scoring methods

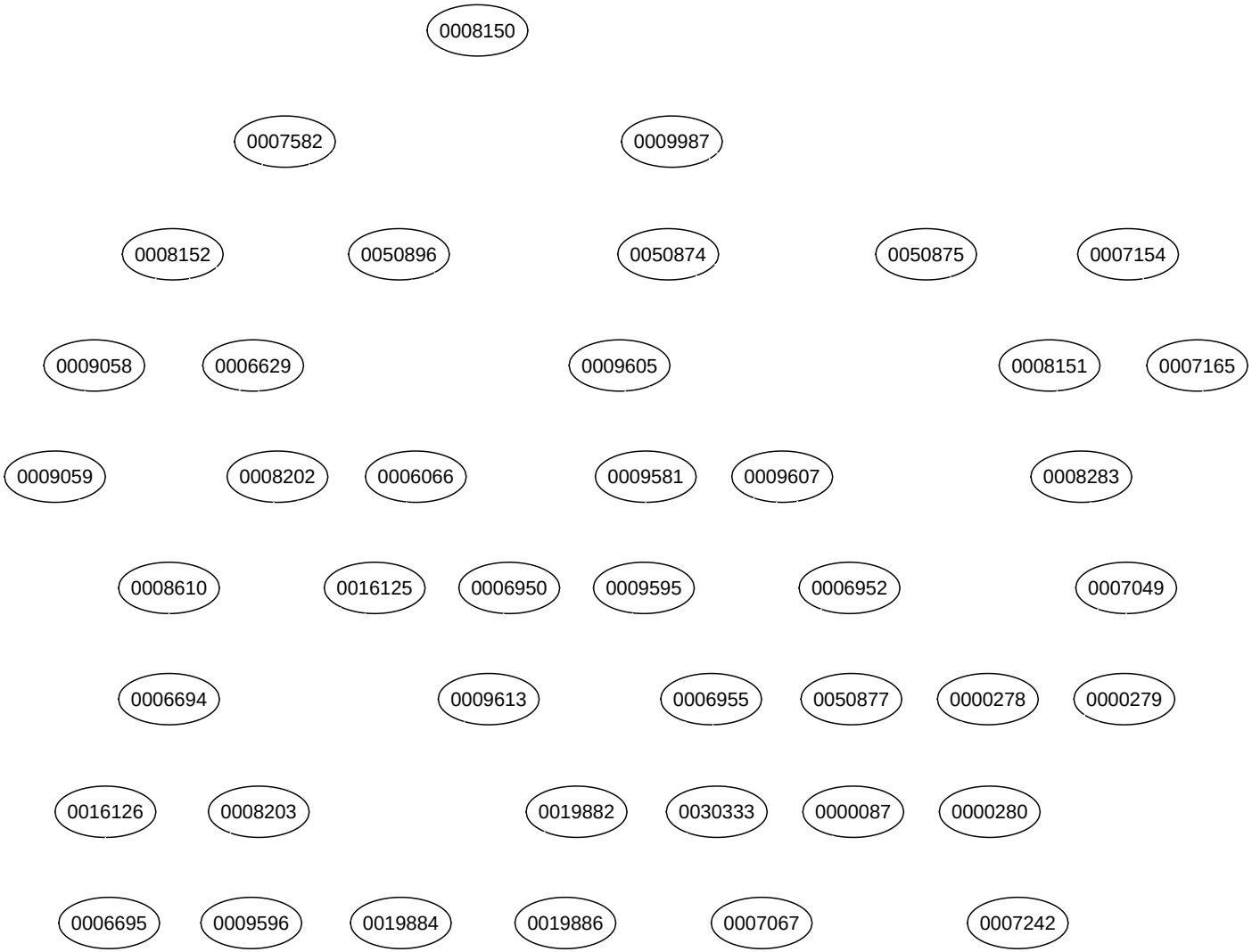
- Obtain the Gene Expression Data from the microarrays experiments (this is the normalized and cleaned data: Long list of genes)
- Select a set of significant genes (use some test statistic: *t-test, permutation-test*)
- Map all the genes to the corresponding GO terms
- Analyze the GO terms for significance (pretty tricky)

  Remark: the GO terms are considered to be independent and the significance is computed for each one separately.

- Khatri P. and Draghici S. (2005). *Ontological analysis of gene expression data: current tools, limitations, and open problems*, Bioinformatics, 21(18):3587-3595.
  - Most used methods: Onto-Express, GOstat, GoMiner, FunSpec, FatiGO, GO::TermFinder
  - Methodically, all known methods are very similar (the accent is put on multiple tests adjustment)



Note: The labels of the nodes are the GO IDs: 0008150 ≅ GO:0008150



Note: The labels of the nodes are the GO IDs: 0008150 ≅ GO:0008150

Small example: suppose that we have a GO term for which we expect ~ 10 genes to be significant.

genes expected	genes in data	
10	10	random
10	12	still random
10	20	better than random
10	40	significant

For computing the significance of a gene set, we can use a *hypergeometric test*:

- N genes are on microarray
- Bio is a GO term
  - M genes $\in Bio$
  - $N - M$ genes $\notin Bio$
- let K be the no. of significant genes
- what is the probability of having exactly x genes from K, of type Bio ?

$$P(X = x|N, M, K) = \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}.$$

- This is the probability of getting exactly x by chance (not what we want)

$$p = 1 - \sum_{i=0}^{x-1} \frac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}}.$$

(also called Fisher's exact test)

The score for a GO term is the degree of independence between the two characteristics: $\mathcal{A}$ = {gene is in the list of significant genes} and $\mathcal{B}$ = {gene is found in the GO term}.

	Significant genes	Not significant genes	Sum
Genes in G	$\lvert \texttt{sigGenes} \cap \texttt{funcGenes} \rvert$	$\lvert \overline{\texttt{sigGenes}} \cap \texttt{funcGenes} \rvert$	$\lvert \texttt{funcGenes} \rvert$
Genes in $\overline{G}$	$\lvert \texttt{sigGenes} \cap \overline{\texttt{funcGenes}} \rvert$	$\lvert \overline{\texttt{sigGenes}} \cap \overline{\texttt{funcGenes}} \rvert$	$\lvert \overline{\texttt{funcGenes}} \rvert$
Sum	$\lvert \texttt{sigGenes} \rvert$	$\lvert \overline{\texttt{sigGenes}} \rvert$	$\lvert \texttt{allGenes} \rvert$

Testing the independence of two groups in the above contingency table corresponds to Fisher's exact test.

	GO:0006955	GO:0009059
Term name	immune response	macromolecule biosynthesis
Definition	Any process involved in the immunological reaction of an organism to an immunogenic stimulus	The formation from simpler components of macromolecules, large molecules including proteins, nucleic acids and carbohydrates
Ontology	BP	BP
# mapped genes	780	568

- The genes are sorted based on a two sided t-test statistic. There are a total of 9905 genes on the array.
- A cutoff of 559 is chosen (the number of genes which are found significant at a level $\alpha = 0.01$ test after a Bonfferoni adjustment procedure is employed).

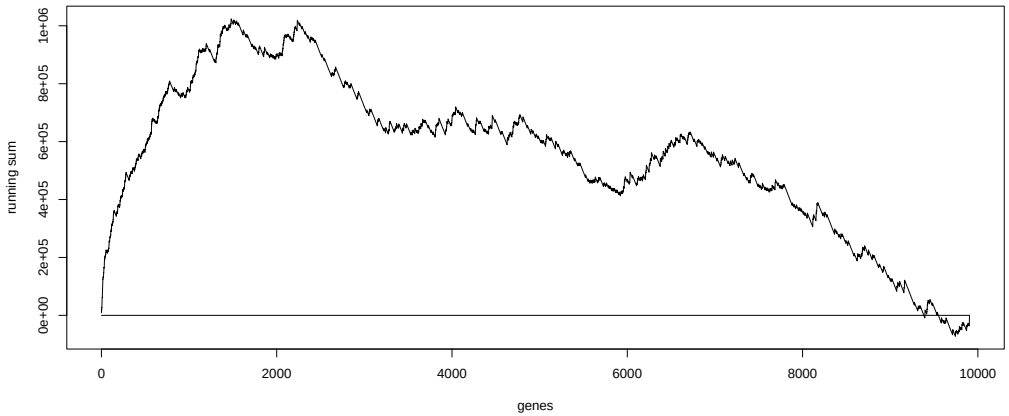
Contingency table for GO:0006955

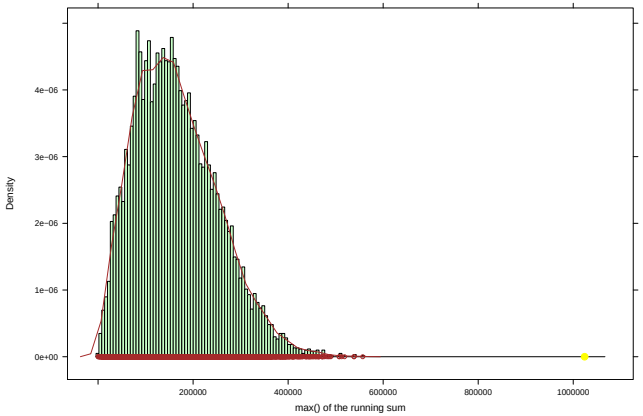
	Significant genes	Not significant genes	Sum
Genes in G	107	673	780
Genes in $\overline{G}$	452	8673	9125
Sum	559	9346	9905

Contingency table for GO:0009059

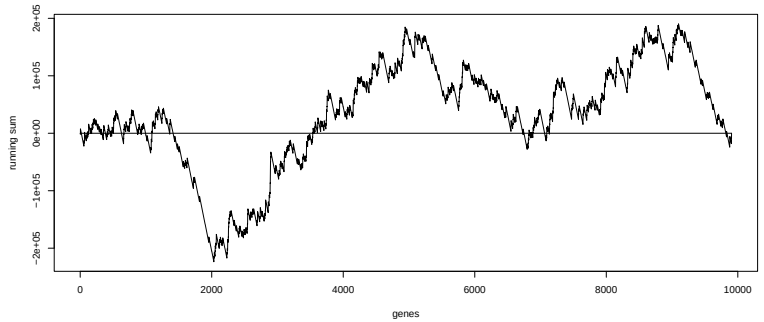
	Significant genes	Not significant genes	Sum
Genes in G	35	533	568
Genes in $\overline{G}$	524	8813	9337
Sum	559	9346	9905

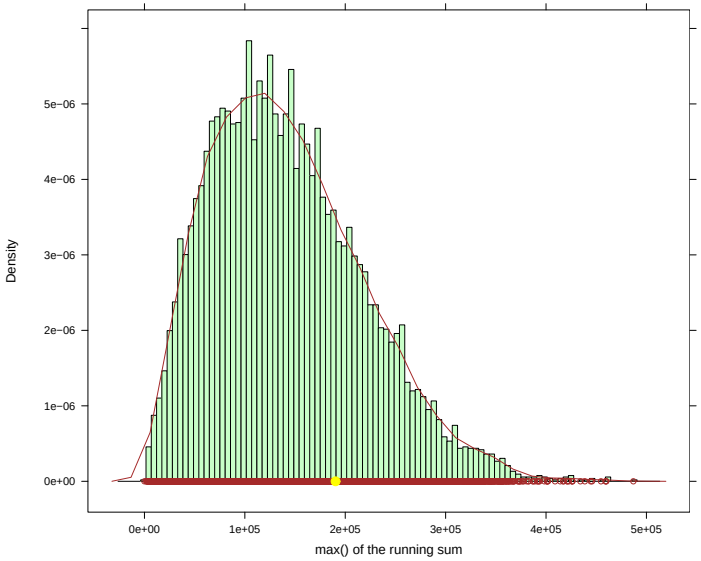
	GO:0006955	GO:0009059
Observed	107	33
Expected	44.020	32.055
Standard deviation	6.186	5.339
raw p-value (Fisher)	7.3e-19	0.3166
adj p-value (Fisher)	7.3e-15	1
raw p-value (Z score)	1.2e-24	0.291





The p-value for GO:0006955 is 0





The$p$-value for GO:0009059 0.2492

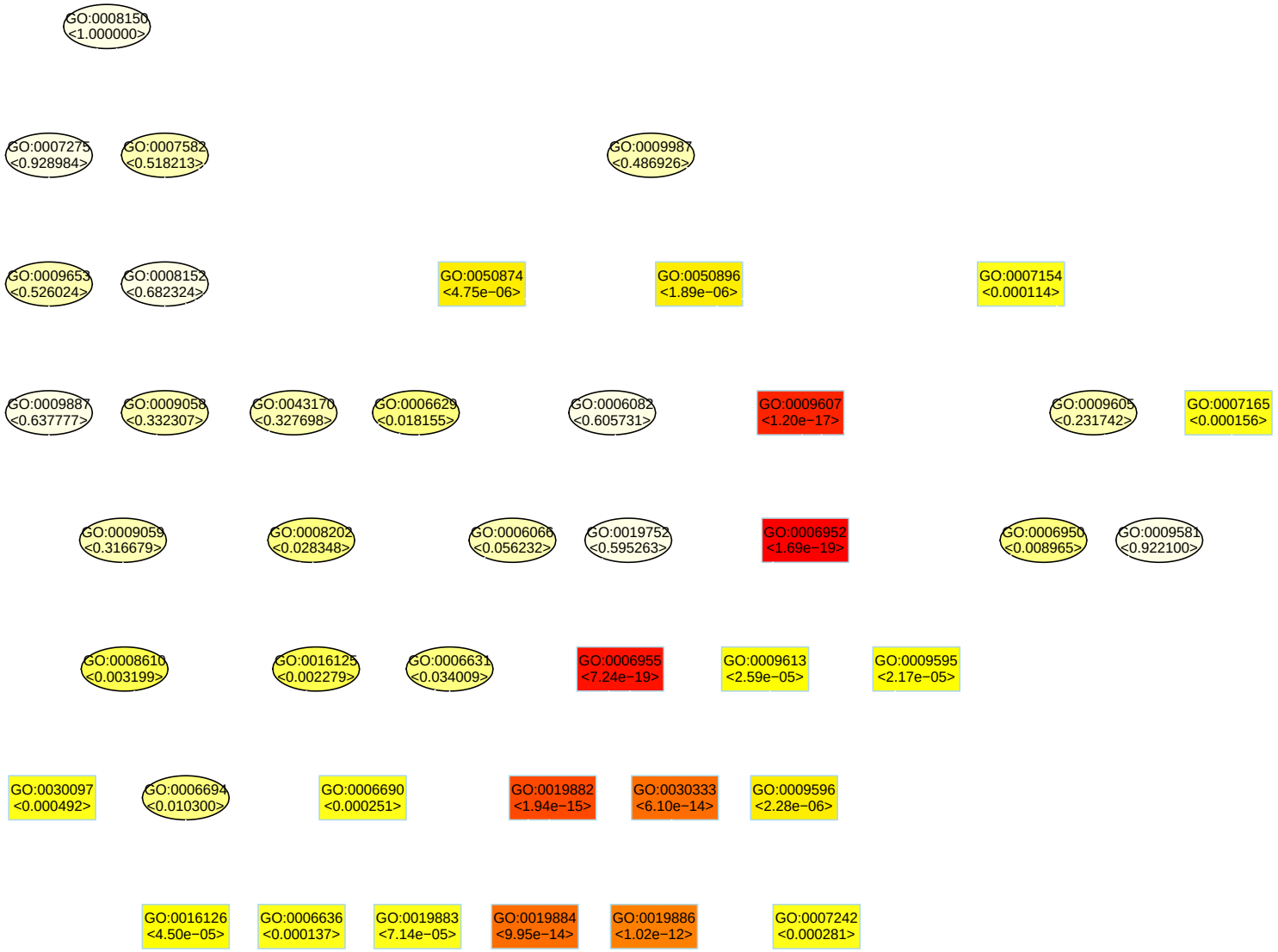
- Gene sets enrichment
- Scoring GO Terms
- **Topology based GO Terms scoring**
- Evaluation of scoring methods

Given:

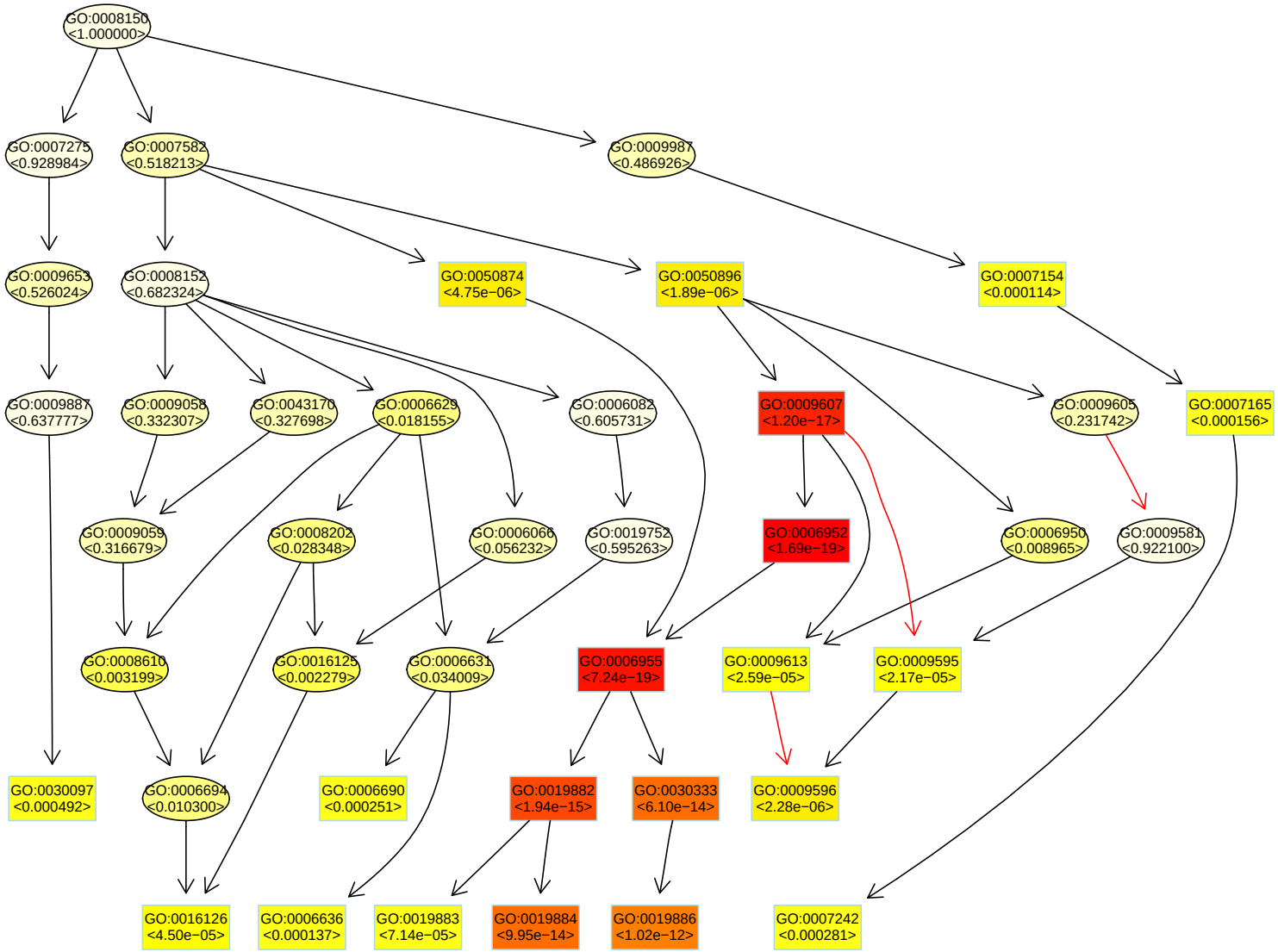
- a directed acyclic graph (GO graph) and a set of items (genes) s.t.:
  - each node in the graph contains some genes
  - the parent of a node contains all the genes of its child
  - a node can contain genes that are not found in the children
- a subset of genes that we call significant genes (differentially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

Goal:

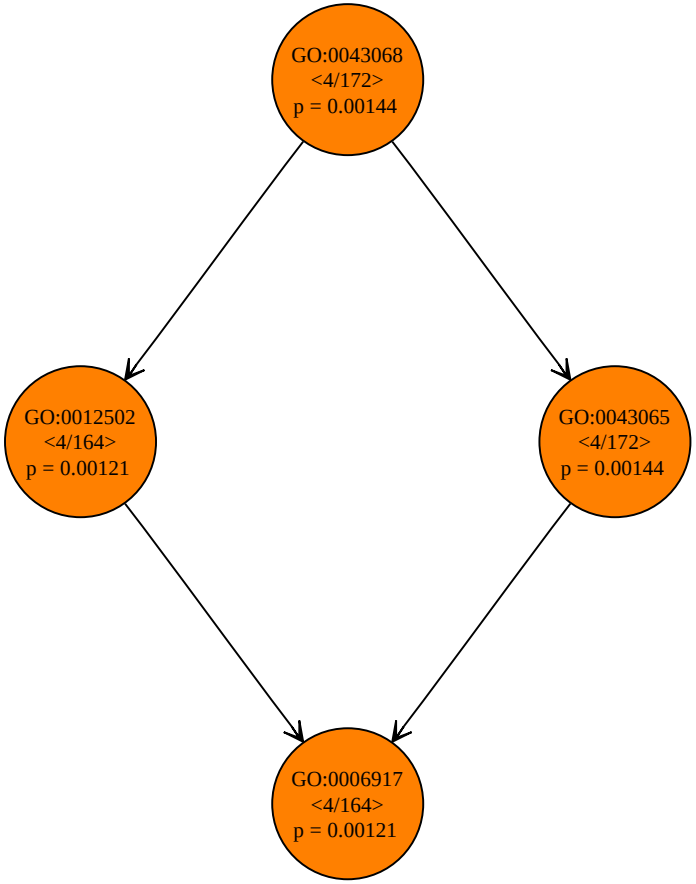
- find the nodes from the graph (biological functions) that best represent the significant genes w.r.t some scoring function (some test statistic)

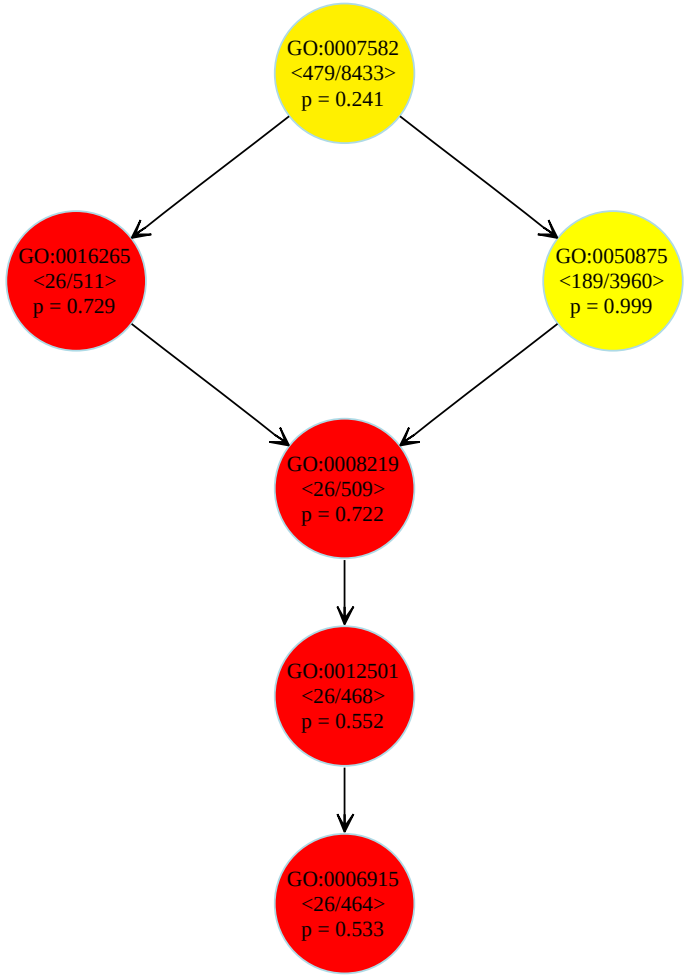


Note: The coloring of the nodes represent the *relative* significance of the GO terms: dark red is the most significant, light yellow is the least significant from the graph



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



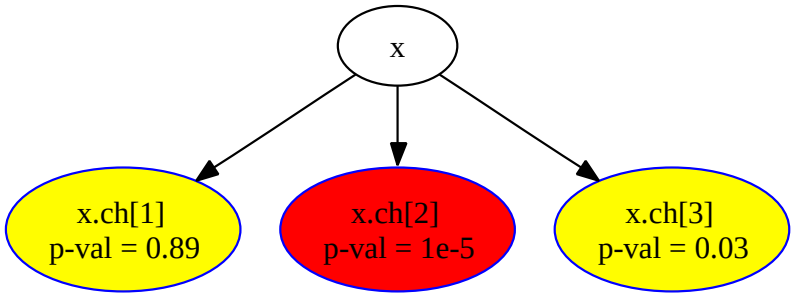


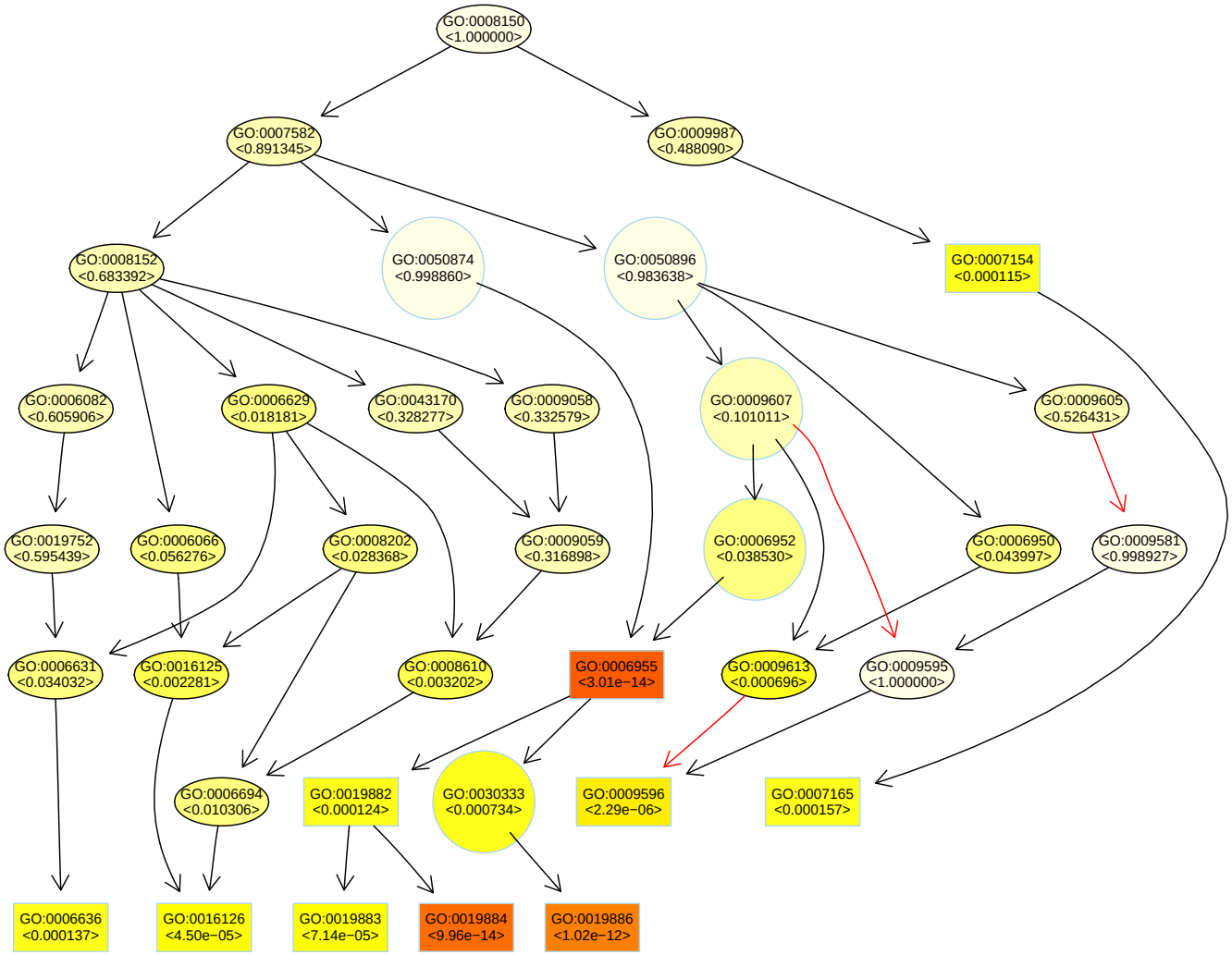
For each GO term the counts and the p-values are displayed. $< x/y >$ denotes that out of y genes mapped to the node, x belong to the list of interesting genes.

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.



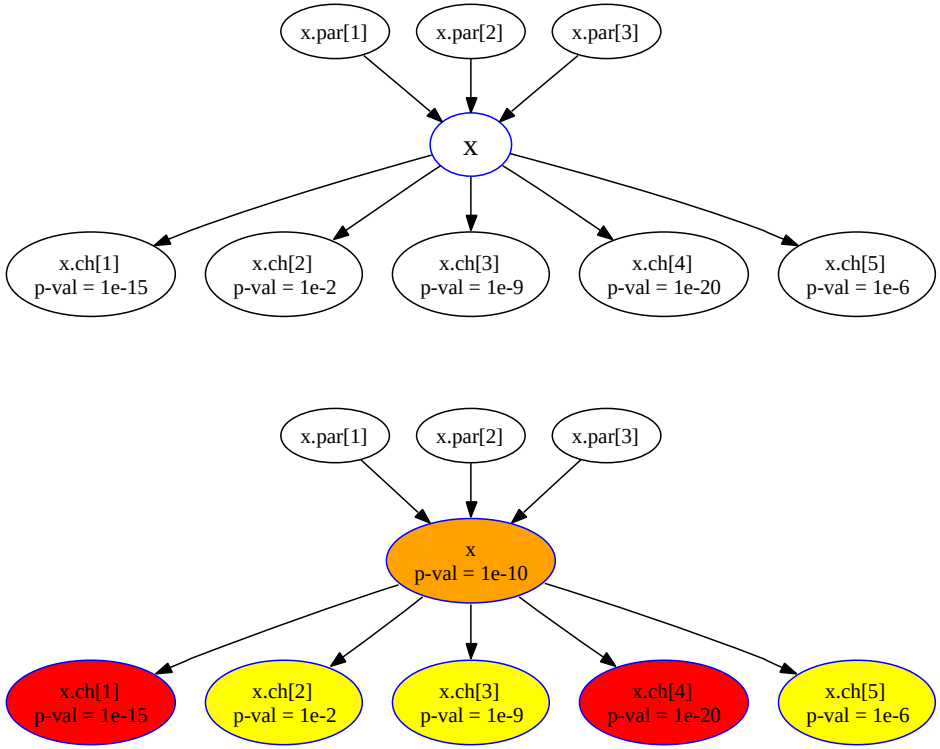


Top 10 significant node (the boxes) obtained with method elim

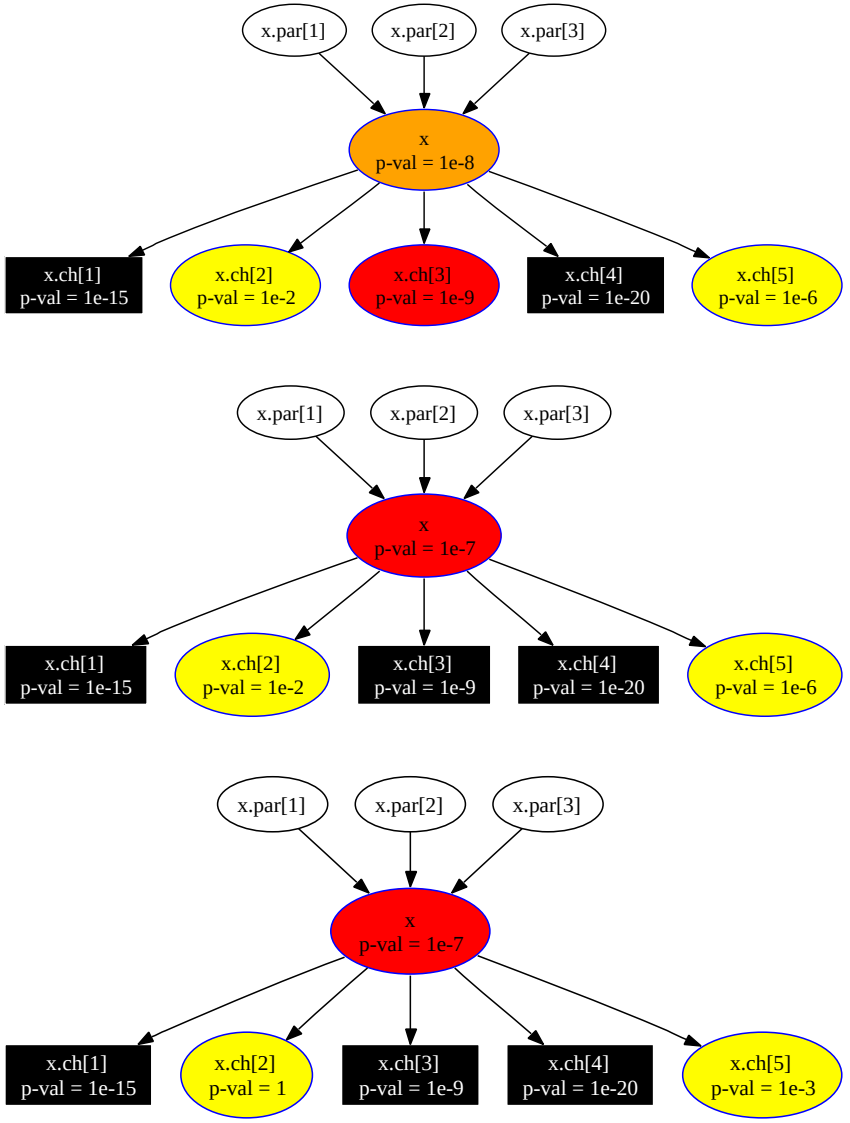
- We want to decide if node x is better representing the list of interesting genes (is more enriched) than any other node from its neighborhood.
- The main idea: Associate single genes mapped to a node with weights that denote their relevance. The elim algorithm uses 0-1 weights.

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.



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

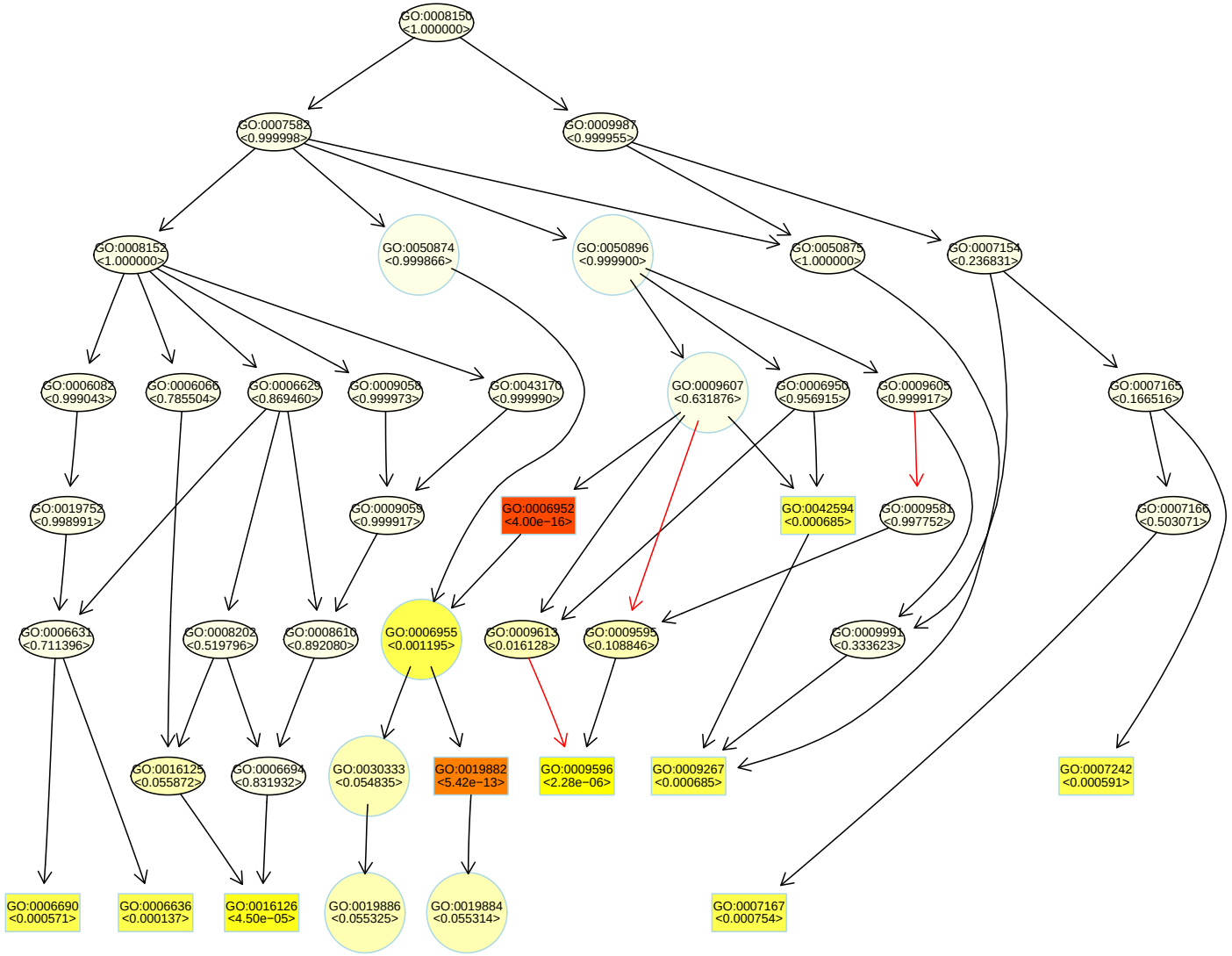
$$\left\lceil \sum_{i \in \{sigGenes \cap genes(u)\}} weight[i] \right\rceil .$$

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

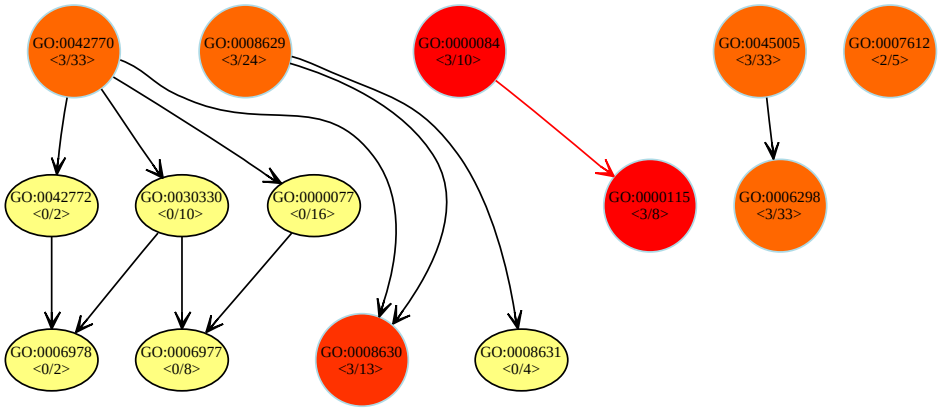
$$\text{sigRatio}(ch, x) = \frac{\log(p\text{-value}(ch))}{\log(p\text{-value}(x))} \quad \text{or} \quad \text{sigRatio}(ch, x) = \frac{p\text{-value}(x)}{p\text{-value}(ch)}$$

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

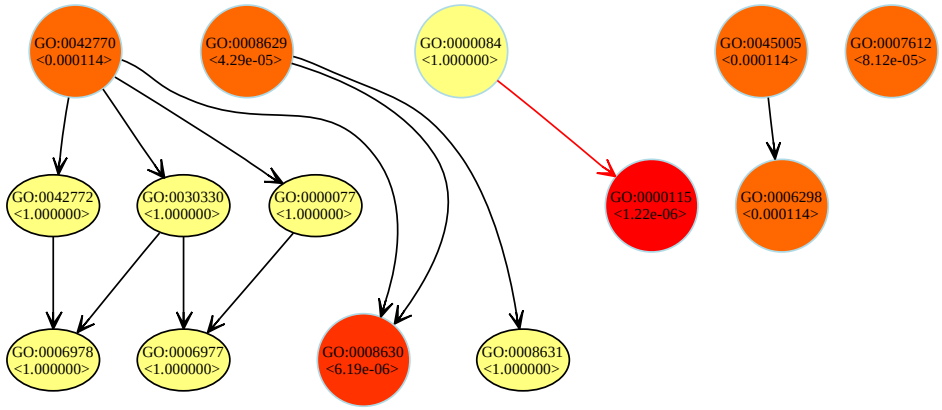
- The weights are updated using vector operators: minimum on the components, the product of the components, etc.



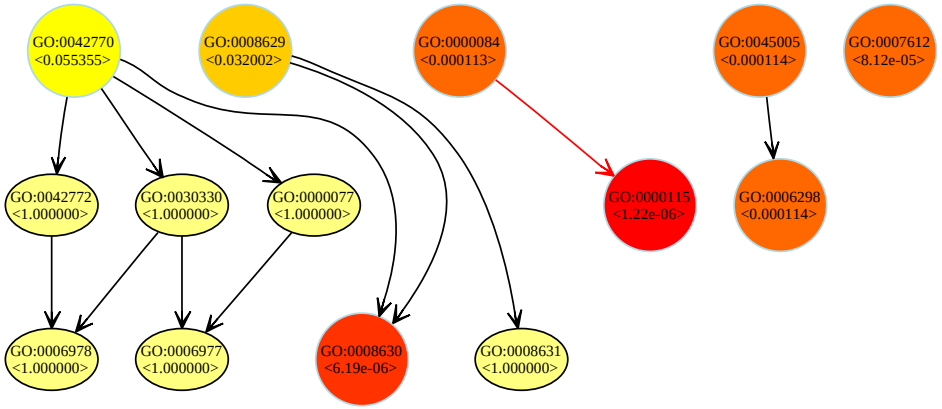
Top 10 significant node (the boxes) obtained with method weight



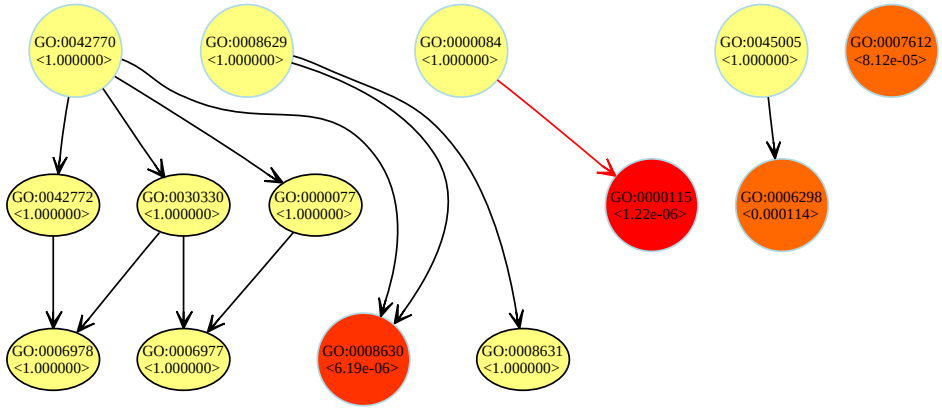
classic method



elim method



weight method



elim method (slightly modified)

	classic	elim	weight.log	weight.ratio
classic	1.000	0.310	0.226	−0.102
elim	0.310	1.000	−0.006	0.388
weight.log	0.226	−0.006	1.000	0.462
weight.ratio	−0.102	0.388	0.462	1.000

Rank correlation for a sample of significant GO terms.

- For each method we retrieve the 100 most significant GO terms.
- The union set of all resulting GO terms is compiled. There are 138 distinct GO terms in this case.
- For these GO terms we retrieve the raw p-values assigned by each method forming a matrix with 4 columns, one column for each method, and 147 rows.

Since the correlation between the results of the algorithms is rather small, we can combine all the algorithms into an ensemble method.

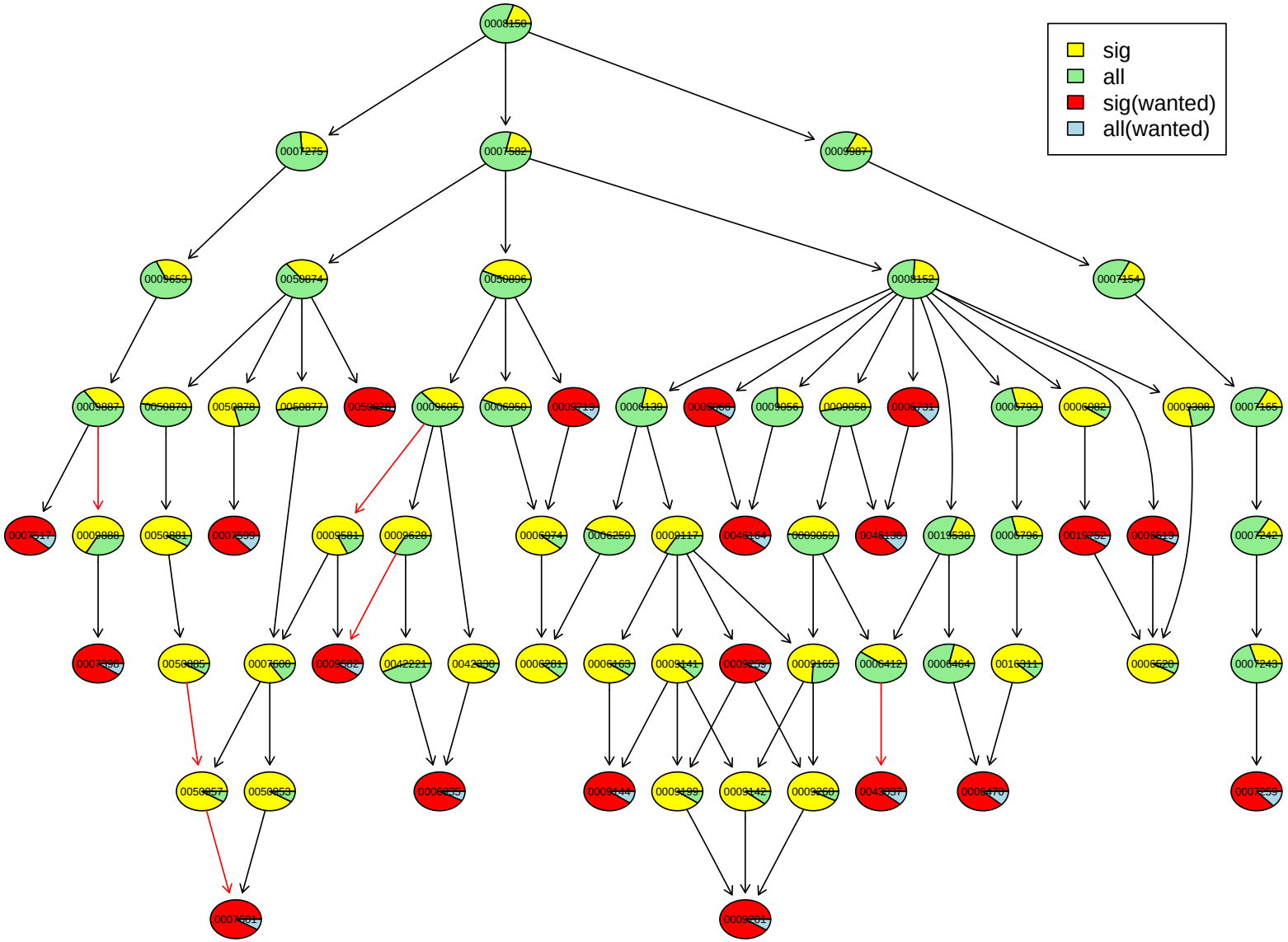
	GO ID	Term	Observed	Expected	Annotated	p-values				
						classic	elim	weight.log	weight.ratio	all.M
1	GO:0006952	defense response	112	46.913	836	6.1e−15	1.000	1.0e−11	5.4e-12	1.5e−05
2	GO:0006955	immune response	102	42.816	763	2.0e−13	5.9e−09	9.3e−09	1.000	3.2e−10
3	GO:0009607	response to biotic stimul...	116	54.264	967	2.4e−12	1.000	9.3e−07	1.000	1.9e−05
4	GO:0019882	antigen presentation	17	1.683	30	1.2e−10	0.647	2.5e−10	5.9e-08	0.00062
5	GO:0030333	antigen processing	17	1.796	32	4.2e−10	0.647	3.5e−10	0.757	0.00083
6	GO:0019884	antigen presentation, exo...	12	0.898	16	4.1e−09	1.2e−08	3.0e−06	1.000	4.6e−08
7	GO:0019886	antigen processing, exoge...	12	1.01	18	3.2e−08	7.6e−08	9.9e−05	1.000	3.8e−07
8	GO:0009605	response to external stim...	127	79.235	1412	3.2e−05	1.000	0.0020	1.000	0.92887
9	GO:0050874	organismal physiological ...	129	89.897	1602	0.012	1.000	0.0071	1.000	1.00000
10	GO:0016126	sterol biosynthesis	9	1.515	27	0.019	0.047	0.0187	0.062	0.11467
11	GO:0050896	response to stimulus	137	98.146	1749	0.020	1.000	0.0726	1.000	0.87163

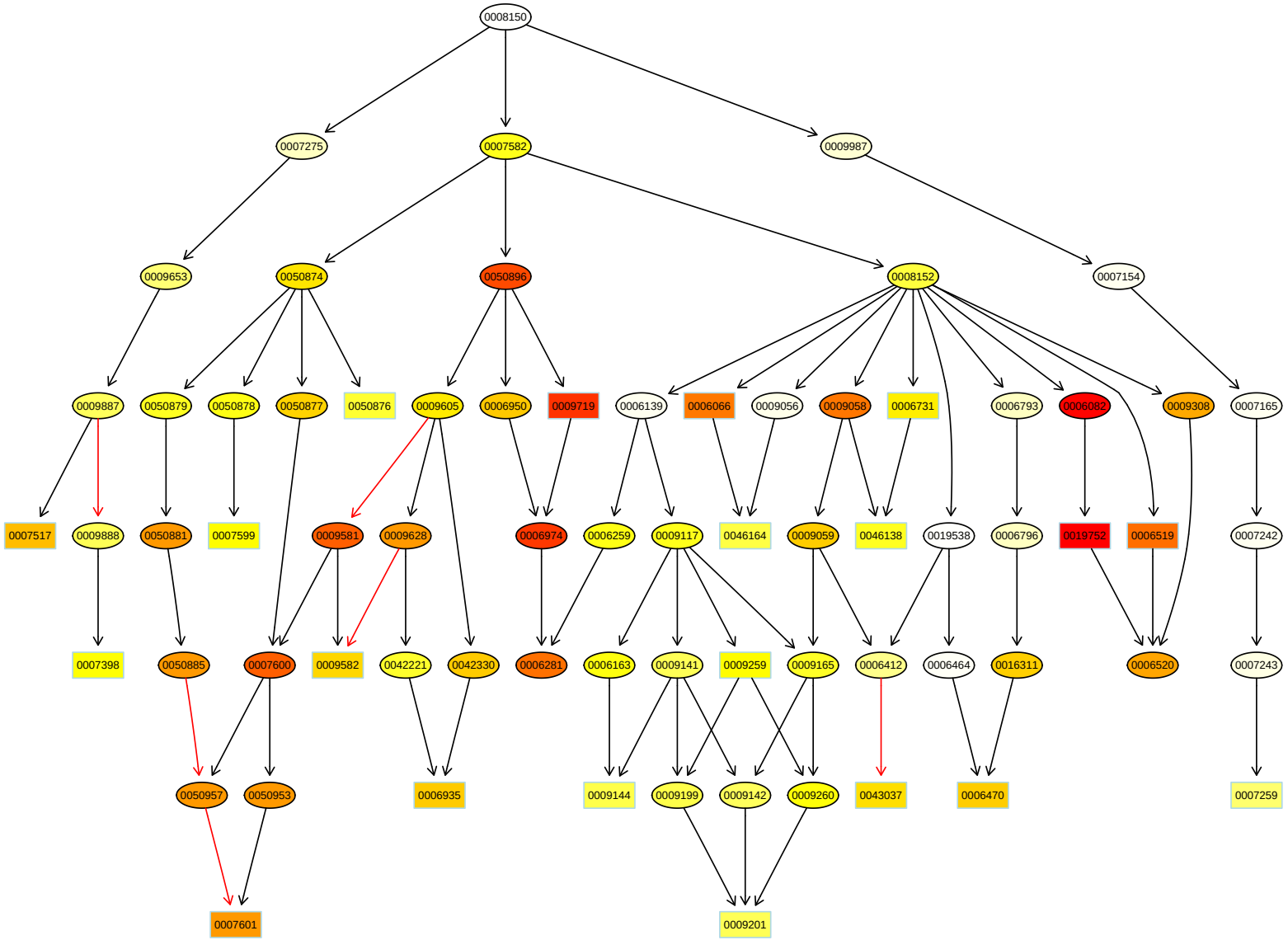
Statistics for significant GO terms for the ALL data set. The column *Expected* represents the expected number of interesting genes mapped to the GO term if the interesting genes were randomly distributed over all GO terms.

- Gene sets enrichment
- Scoring GO Terms
- Topology based GO Terms scoring
- **Evaluation of scoring methods**

- 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.


sig
all
sig(wanted)
all(wanted)


0008150
0007275
0007582
0009987
0009653
0050874
0050896
0008152
0007154
0009887
0050879
0050878
0050877
0050876
0009605
0006950
0009719
0006139
0006066
0009056
0009058
0006731
0006793
0006082
0009308
0007165
0007517
0009888
0050881
0007599
0009581
0009628
0006974
0006259
0009117
0046164
0009059
0046138
0019538
0006796
0019752
0006519
0007242
0007398
0050885
0007600
0009582
0042221
0042330
0006281
0006163
0009141
0009259
0009165
0006412
0006464
0016311
0006520
0007243
0050957
0050953
0006935
0009144
0009199
0009142
0009260
0043037
0006470
0007259
0007601
0009201

- To assess the performance of each method $\mathcal{M}$ the following scores are used:

$$score_k^0(\mathcal{M}) = |top_k(\mathcal{M}) \cap enriched| \,.$$

  i.e. the number of *enriched nodes* found among the top k nodes.

- To get more insight into how each method accounts for the topology of the graph, the following scores are defined:

$$score_k^1(\mathcal{M}) = \left|level_k^1(\mathcal{M}) \cap enriched\right| ,$$

$$score_k^{1p}(\mathcal{M}) = \left|level_k^{1p}(\mathcal{M}) \cap enriched\right|$$

  with

$$level_k^1 = top_k(\mathcal{M}) \cup parents(top_k(\mathcal{M})) \cup children(top_k(\mathcal{M})),$$

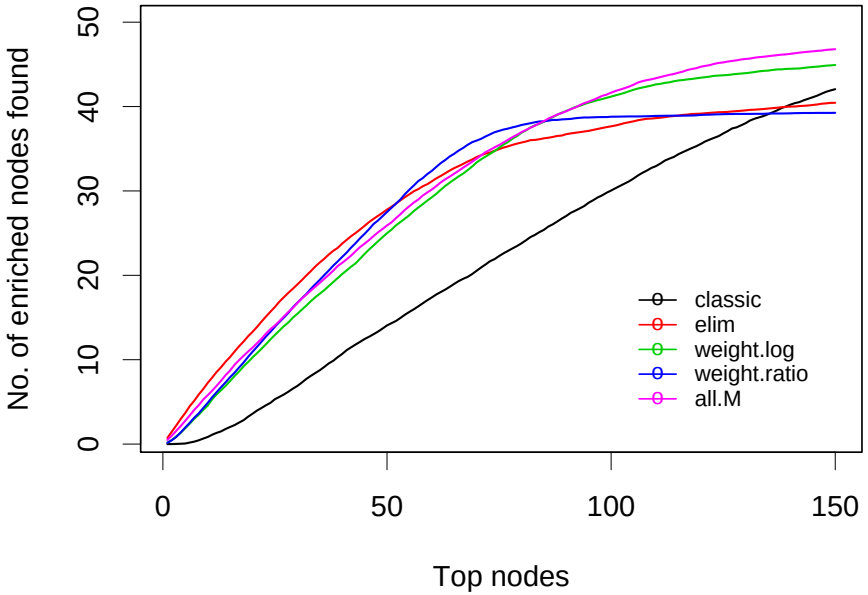
$$level_k^{1p} = top_k(\mathcal{M}) \cup parents(top_k(\mathcal{M})).$$

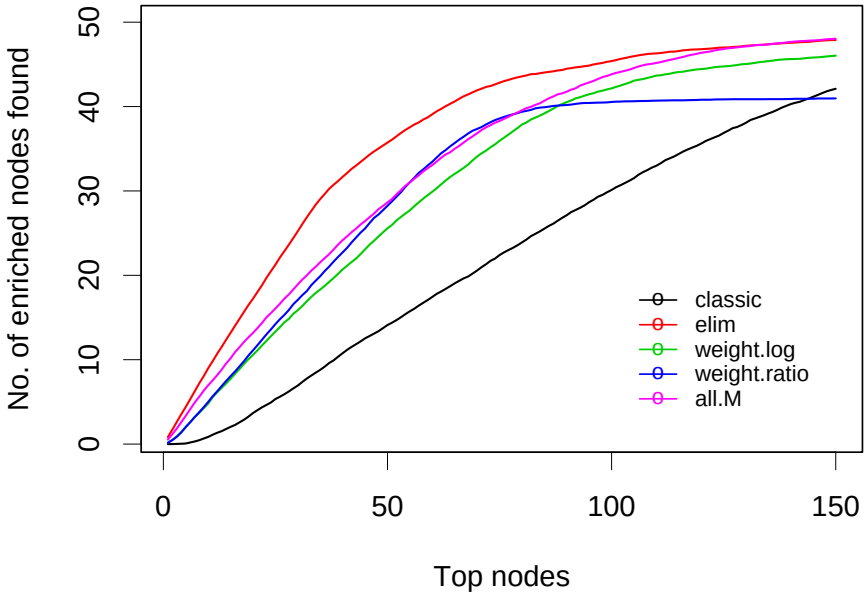
- Methods that obtain a higher score better retrieve the true enriched nodes.

k	class	weight.log	weight.ratio	elim	all.M
25	5.5	13	14	17	15.5
50	**14.5**	**25.5**	**28**	**27.5**	**28.5**
75	22.5	35.5	38	31	38
100	31	42	39.5	33.5	43.5

k	Score	class	weight.log	weight.ratio	elim	all.M
	0	**14.5**	**25.5**	**28**	**27.5**	**28.5**
	1p	15	26	29	40	31
50	1	23	32	35	41	36
	2p	15	26	29	43	31
	2	29	36	39	45	40

Average numbers of correctly identified *enriched nodes* over 100 simulation runs with 50 true *enriched nodes*, 10% noise level, and between 10 and 50 genes annotated to the *enriched nodes*.





The average performance of the algorithms for 100 simulation runs, 50 enriched nodes, 10 to 50 genes annotated, 10% noise level. The left plot represents $score_k^0$ and the right plot represents $score_k^{1p}$.