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 Saarbrücken

Courses in Practical DNA Microarray Analysis, Saarbrücken, 28th of September, 2006

- Gene set enrichment
  - Parametric based tests [Khatri and Draghici, 2005]
  - Distribution based tests [Subramanian, A., *et al.*, 2005]
- Gene Ontology terms scoring
  - classic method
  - elim method
  - weight method
- Evaluation and stability of the methods
  - Discrimination into B-cell and T-cell type leukemias [Chiaretti, S., *et al.*, 2004]
  - Discrimination based on minimal residual disease (MRD) [Cario, G., *et al.*, 2005]
  - Factor analysis for prostate cancer progression
  - Influence of the p-value adjustment
  - Evaluation on simulated data
- Conclusions

- **Gene set enrichment**
  - Parametric based tests [Khatri and Draghici, 2005]
  - Distribution based tests [Subramanian, A., *et al.*, 2005]
- Gene Ontology terms scoring
- Evaluation and stability of the methods
- Conclusions

- 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.

- 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.

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

Enrichment 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.

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

Enrichment 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).

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

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

The score for a GO term is the degree of independence between the two properties:

$$\mathcal{A} = \{\text{gene is in the list of significant genes}\}$$
$$\mathcal{B} = \{\text{gene is found in the GO term}\}.$$

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

Testing the independence of two groups in the above contingency table corresponds to Fisher's exact test [Khatri and Draghici, 2005].

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

genes expected		
10 10 10 10		

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

genes expected	genes in data	
10 10 10 10	10	random

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		
10		

Background on Fisher's exact test

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		

Small example: suppose that we have a GO term for which we expect $\sim$ 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

Background on Fisher's exact test

Small example: suppose that we have a GO term for which we expect $\sim$ 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

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 ?

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 ?

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

Small example: suppose that we have a GO term for which we expect $\sim$ 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)

Small example: suppose that we have a GO term for which we expect $\sim$ 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)

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

(similar 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

Discriminating B-cell and T-cell [Chiaretti, S., et al., 2004]

- ALL dataset consists of 128 microarrays (95 patients with B-cell ALL and 33 patients with T-cell ALL).
- The Affymetrix HGU95aV2 chip used contain 12625 probes (9231 probes are annotated to BP) which induce a GO graph containing 2677 nodes.
- 515 differentially expressed genes (two-sided t-test, FDR-adjusted p-values, level $\alpha = 0.01$).

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

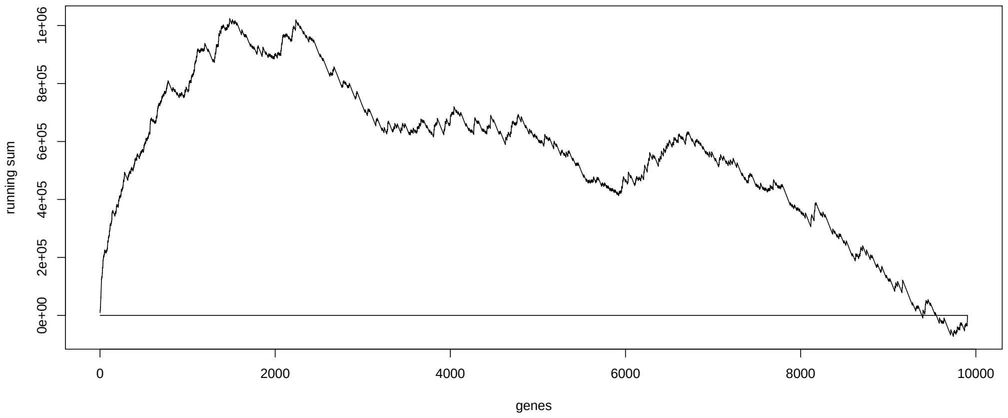
Contingency table for GO:0009059

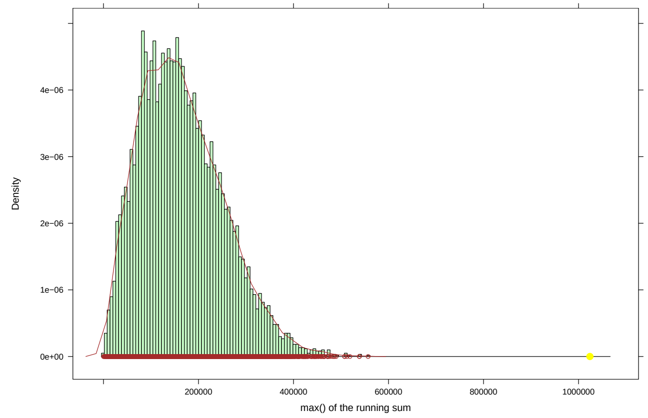
	Significant genes	Not significant genes	Sum
Genes in G	35	533	568
Genes in $\overline{G}$	524	8813	9337
Sum	559	9346	9905

	GO:0006955	GO:0009059
Observed	107	35
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

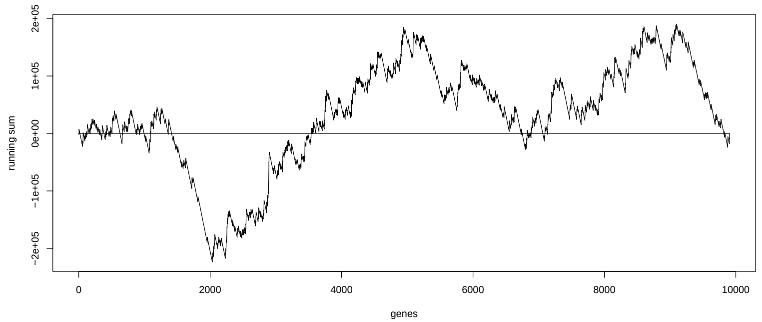
- Fixing a cutoff and looking only at the top genes can be sometimes misleading. Also the position of the genes is not considered in the previous approach. The information embedded in the genes below the cutoff is not used. We want to analyze the distribution of all ranks of members of group **a**.

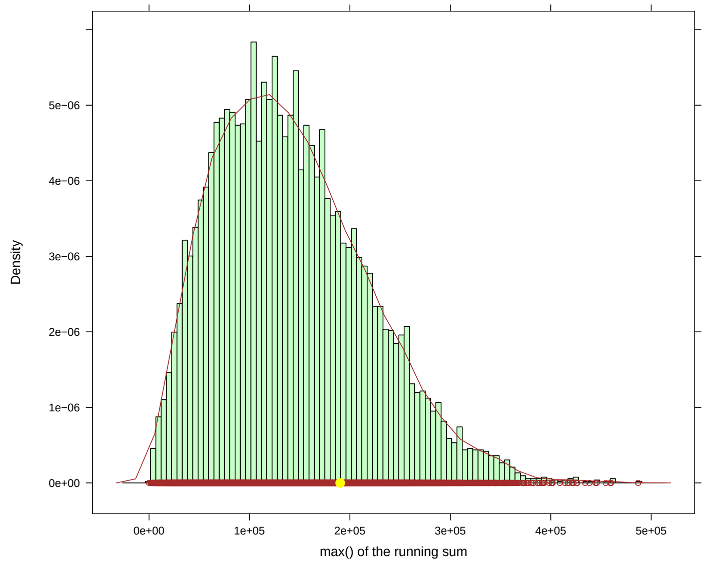
GO example (Kolmogorov-Smirnov)





The p-value for GO:0006955 is 0





The$p$-value for GO:0009059 0.2492

- Gene set enrichment
- **Gene Ontology terms scoring**
  - classic method
  - elim method
  - weight method
- Evaluation and stability of the methods
- Conclusions

GO scoring: general problem

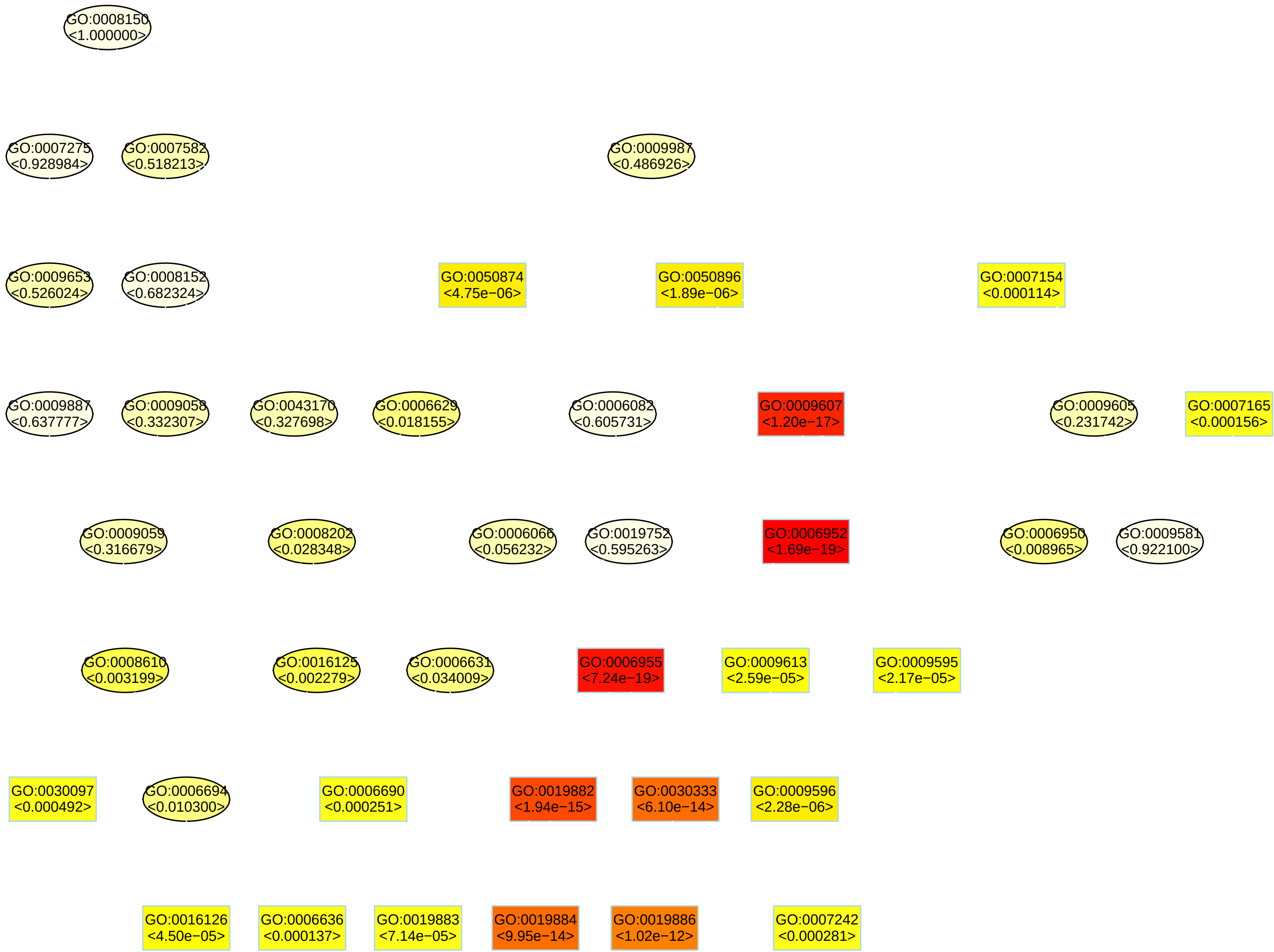
Given:

- 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)

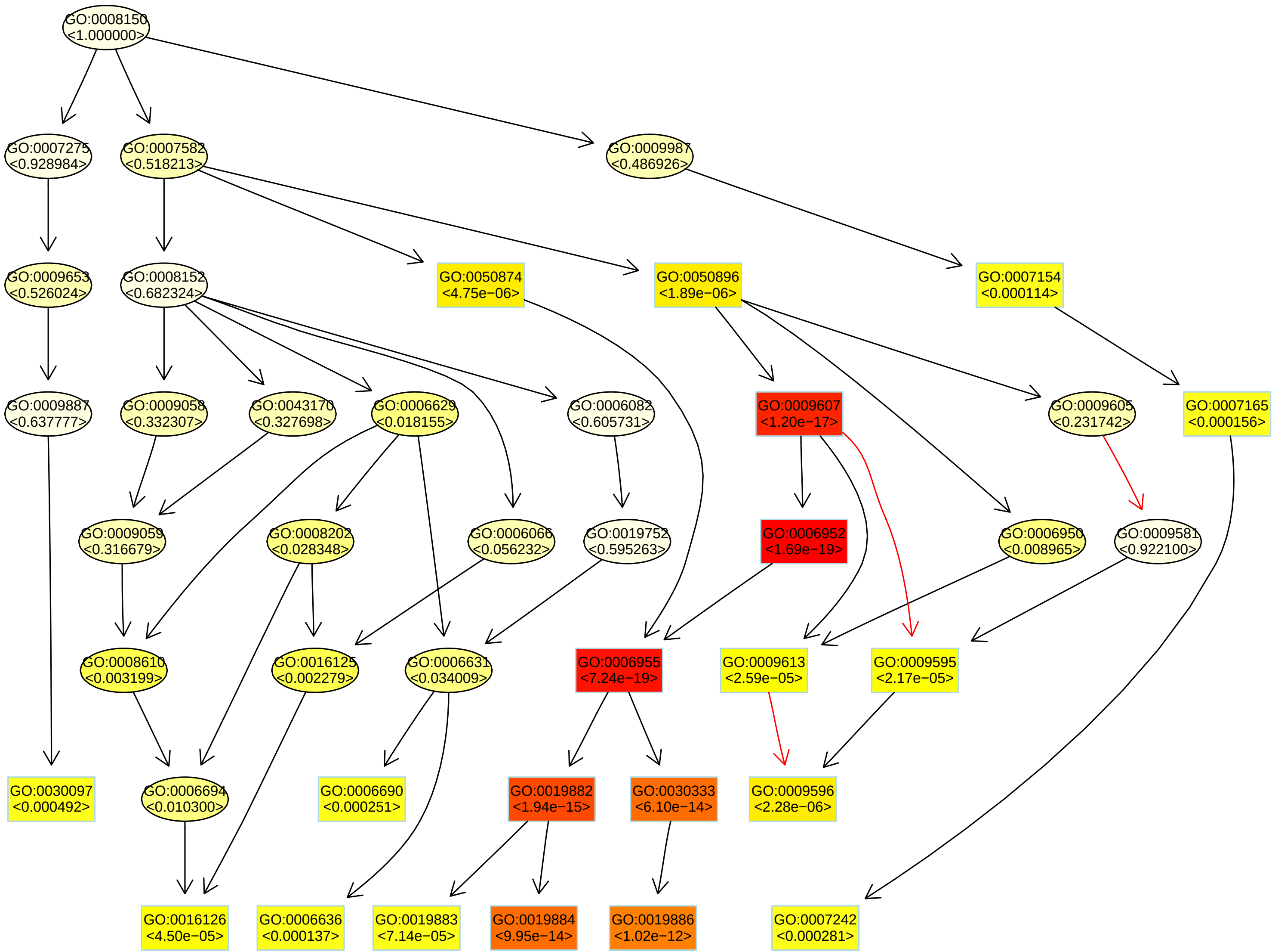
Goal:

- 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



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

➢ classic **algorithm**

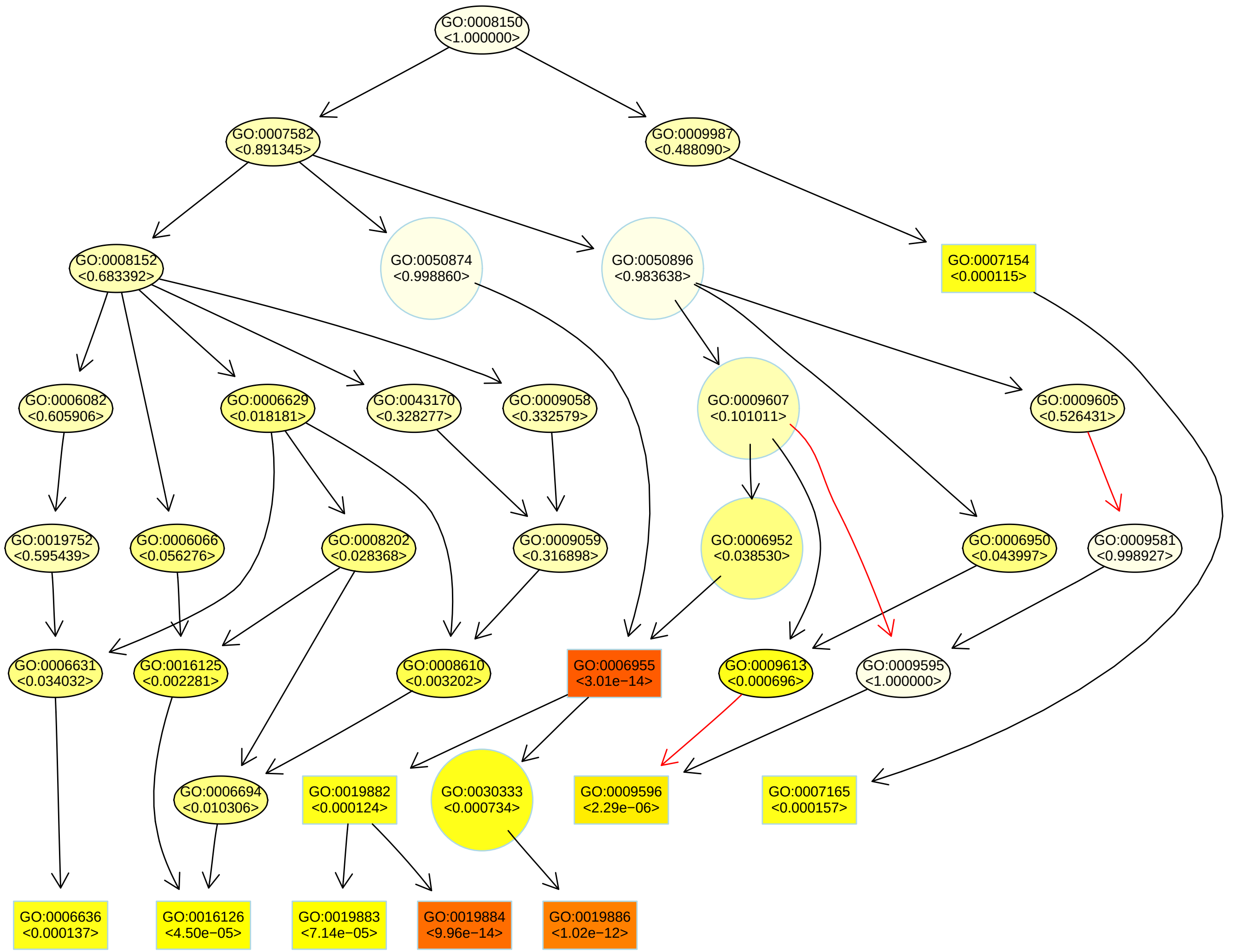
- Calculate significance of each GO term independently.
- Adjust pvalues for multiple testing (Bonferroni, FDR, etc.).
- Kolmogorov-Smirnov test can easily be used in this case

➢ elim **algorithm**

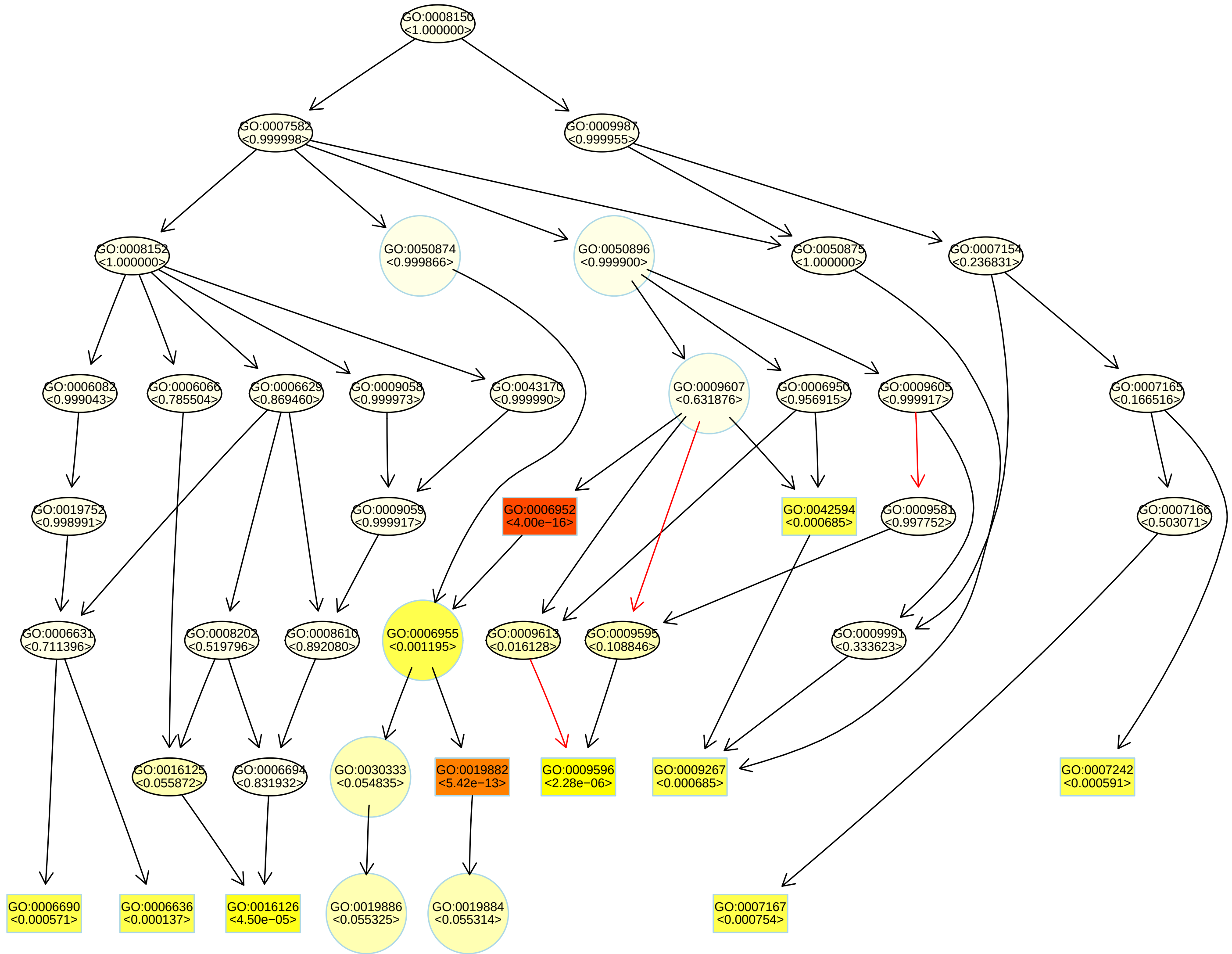
- Nodes are processed bottom-up in the GO graph.
- It iteratively removes the genes annotated to significant GO terms from more general GO terms.
- Intuitive and simple to interpret.

➢ 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.



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



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

- Gene set enrichment
- Gene Ontology terms scoring
- **Evaluation and stability of the methods**
  - Discrimination into B-cell and T-cell type leukemias [Chiaretti, S., *et al.*, 2004]
  - Discrimination based on minimal residual disease (MRD) [Cario, G., *et al.*, 2005]
  - Influence of the p-value adjustment
  - Evaluation on simulated data
- Conclusions

➢ **Discriminating B-cell and T-cell** [Chiaretti, S., *et al.*, 2004]

- ALL dataset consists of 128 microarrays (95 patients with B-cell ALL and 33 patients with T-cell ALL).
- The Affymetrix HGU95aV2 chip used contain 12625 probes (9231 probes are annotated to BP) which induce a GO graph containing 2677 nodes.
- 515 differentially expressed genes (two-sided t-test, FDR-adjusted p-values, level $\alpha = 0.01$).

➢ **Discriminating B-cell and T-cell** [Chiaretti, S., *et al.*, 2004]

- ALL dataset consists of 128 microarrays (95 patients with B-cell ALL and 33 patients with T-cell ALL).
- The Affymetrix HGU95aV2 chip used contain 12625 probes (9231 probes are annotated to BP) which induce a GO graph containing 2677 nodes.
- 515 differentially expressed genes (two-sided t-test, FDR-adjusted p-values, level $\alpha = 0.01$).

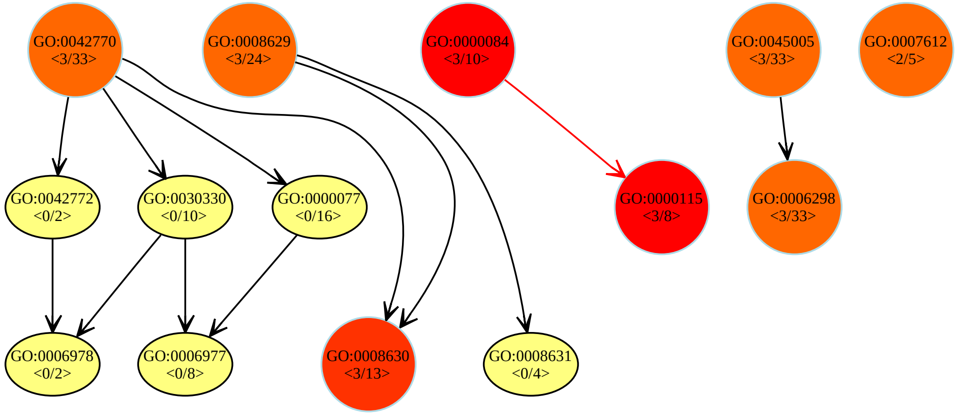
➢ **Discriminating the load level of minimal residual disease (MRD)** [Cario, G., *et al.*, 2005]

- ALL dataset consists of 51 microarrays (30 patients with detectable MRD (MRD-SR) and 21 patients with high MRD load (MRD-HR)).
- Two color chip provides (after preprocessing) 13236 genes (6853 genes are annotated to BP) which induce a GO graph containing 2733 nodes.
- 682 differentially expressed genes (two-sided t-test, FDR-adjusted p-values, level $\alpha = 0.01$)

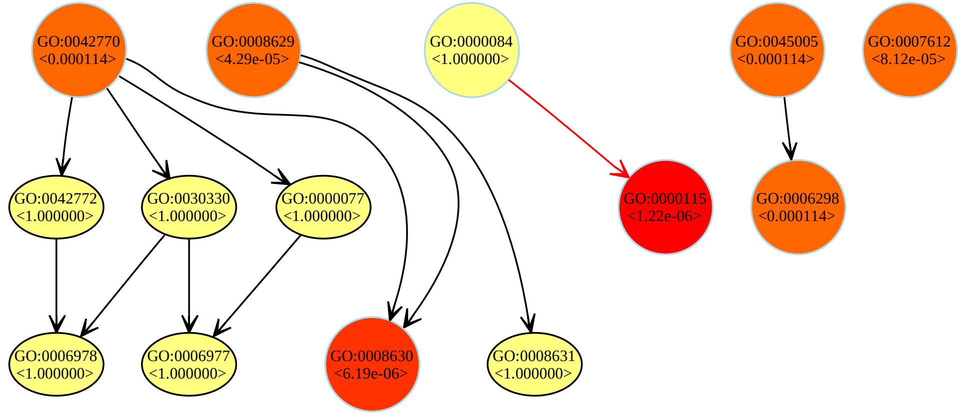
	GO ID	Term	Observed	Expected	Annotated	p-values						
						classic	elim	weight.ratio	weight.log	weight.01	KS	all.M
1	GO:0019882	antigen presentation	22	2.287	41	1.6e-17	0.2821	1.6e-17	1.6e-17	1.6e-17	1e-04	2.8e-14
2	GO:0006952	defense response	107	47.143	845	8.3e-17	0.0065	1.1e-06	1.4e-09	1.7e-06	1e-04	1.7e-08
3	GO:0030333	antigen processing	20	2.12	38	7.8e-16	1.0000	7.8e-16	7.8e-16	7.8e-16	1e-04	8.2e-13
4	GO:0006955	immune response	98	43.293	776	2.7e-15	5.9e-06	0.024	3.0e-05	3.8e-05	1e-04	8.5e-07
5	GO:0019884	antigen presentation, exogenou...	14	1.004	18	5.9e-15	5.9e-15	0.054	2.2e-10	5.9e-15	1e-04	1.9e-11
6	GO:0009607	response to biotic stimulus	112	53.949	967	9.5e-15	0.6873	0.404	1.0e-05	0.945	1e-04	0.00012
7	GO:0019886	antigen processing, exogenous ...	14	1.116	20	6.8e-14	6.8e-14	0.054	1.5e-11	6.8e-14	1e-04	4.8e-11
8	GO:0009596	detection of pest, pathogen or...	9	0.725	13	2.9e-09	2.9e-09	2.9e-09	2.9e-09	3.6e-08	1e-04	4.7e-09
9	GO:0009595	detection of biotic stimulus	9	0.893	16	3.9e-08	1.0000	0.107	1.0e-05	0.055	1e-04	0.00119
10	GO:0016126	sterol biosynthesis	9	1.395	25	4.5e-06	0.0015	4.5e-06	4.5e-06	4.5e-06	0.0016	1.4e-05

	GO ID	Term	Observed	Expected	Annotated	p-values						
						classic	elim	weight.ratio	weight.log	weight.01	KS	all.M
1	GO:0019884	antigen presentation, exogenou...	6	1.095	11	0.00028	0.00028	0.00028	0.00028	0.00028	0.0022	0.00028
2	GO:0009887	organogenesis	85	59.512	598	0.00032	0.00158	0.02427	0.00624	0.04707	0.0003	0.00514
3	GO:0007155	cell adhesion	58	37.319	375	0.00036	0.00036	0.00029	0.00031	0.00058	0.0005	0.00040
4	GO:0019886	antigen processing, exogenous ...	6	1.194	12	0.00052	0.00052	0.00052	0.00052	0.00052	0.0038	0.00052
5	GO:0000187	activation of MAPK activity	7	1.692	17	0.00075	0.00075	0.00075	0.00075	0.00075	0.0062	0.00075
6	GO:0043406	positive regulation of MAPK ac...	7	1.692	17	0.00075	1.00000	0.07989	0.00805	0.00805	0.0078	0.02077
7	GO:0007275	development	141	110.864	1114	0.00079	0.16380	0.30040	0.08667	0.22699	1e-04	0.05985
8	GO:0048513	organ development	87	62.995	633	0.00082	0.86056	0.23651	0.02928	0.09564	0.0003	0.05416
9	GO:0007422	peripheral nervous system deve...	5	0.896	9	0.00086	0.00086	0.00086	0.00086	0.00086	0.0029	0.00086
10	GO:0042438	melanin biosynthesis	4	0.597	6	0.00124	1.00000	0.02758	0.02758	0.02758	0.0056	0.03040

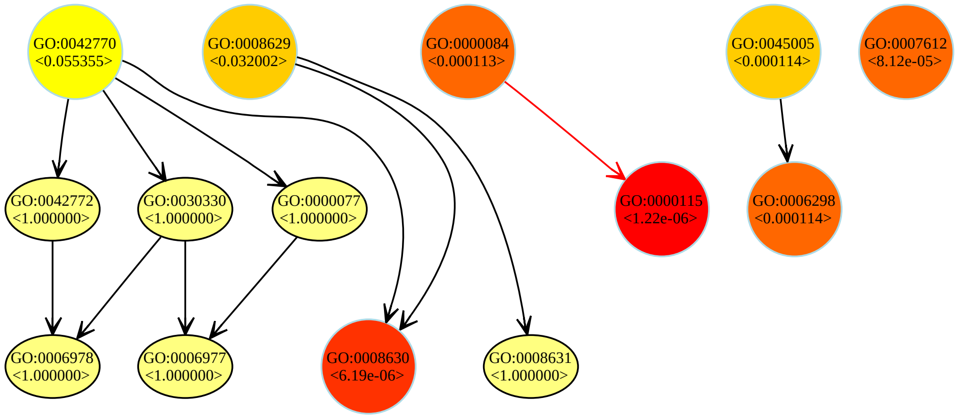
Advantages & Disadvantages for ALL



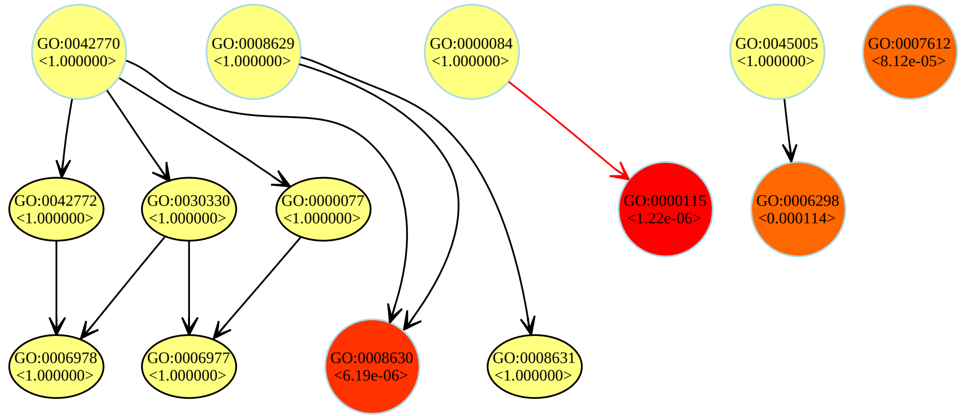
classic method



elim method



weight method

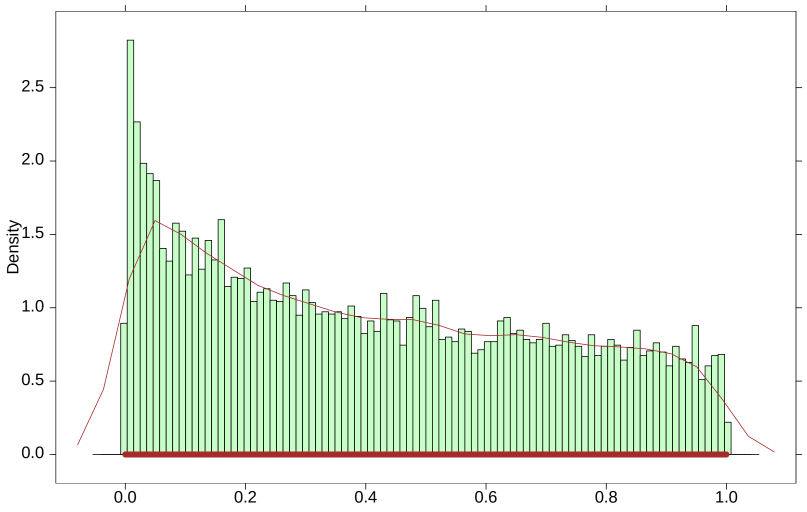


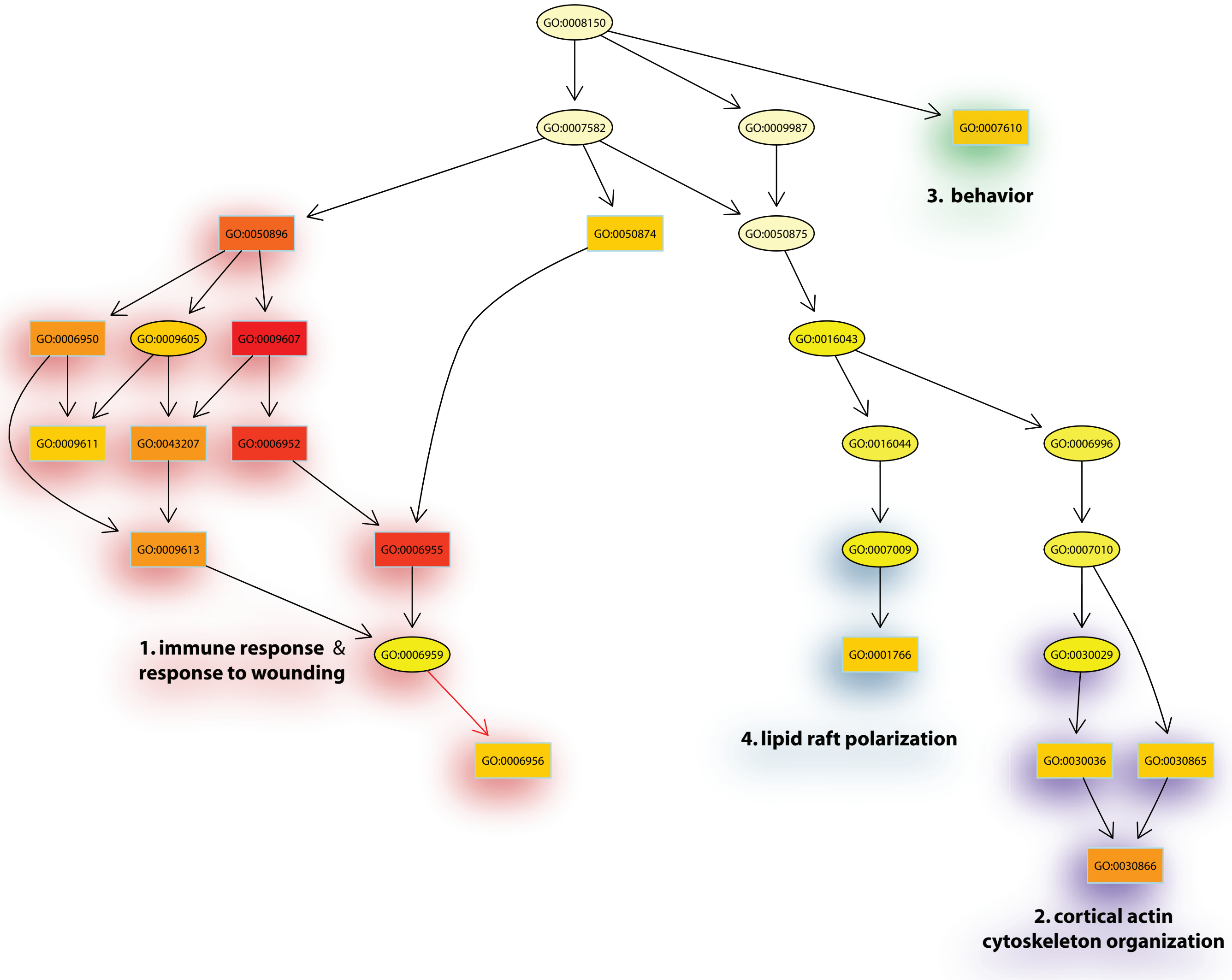
elim method (slightly modified)

➢ GO interaction effect analysis

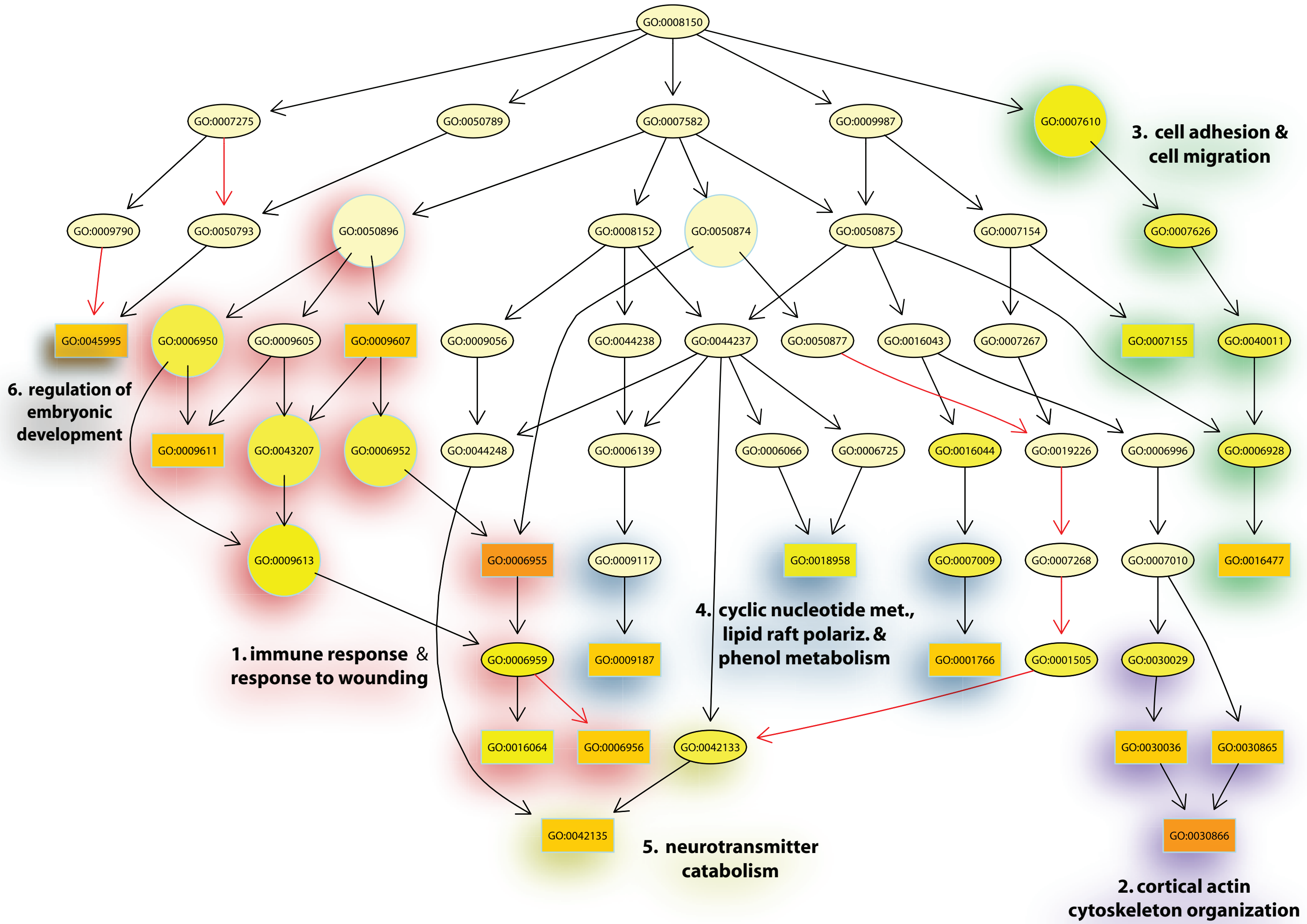
- The dataset consists of 23 microarrays (4 patients with a synergetic effect).
- The Affymetrix HGU133a chip used contain 22283 probes (7774 probes are annotated to BP) which induce a GO graph containing 2429 and 3944 edges.
- Genes were filtered such that the expression values on more than 25% of the samples are over 6.5.
- 337 differentially expressed genes (significance of α_3 coefficient of the linear model, raw p-values, level $\alpha = 0.01$).
- Test for interaction effect: $H_0 : \alpha_3 = 0$ vs $H_1 : \alpha_3 \neq 0$ based on the following linear model:

$$\log(g) = \alpha_0 + \alpha_1 I_{hypo} + \alpha_2 I_{chorm8} + \alpha_3 I_{hypo} I_{hypo} + \epsilon$$





Top 15 significant node (the boxes) obtained with method classic



Top 15 significant node (the boxes) obtained with method weight

Influence of the p-values adjustment

- 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*.

Influence of the p-values adjustment

- 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?

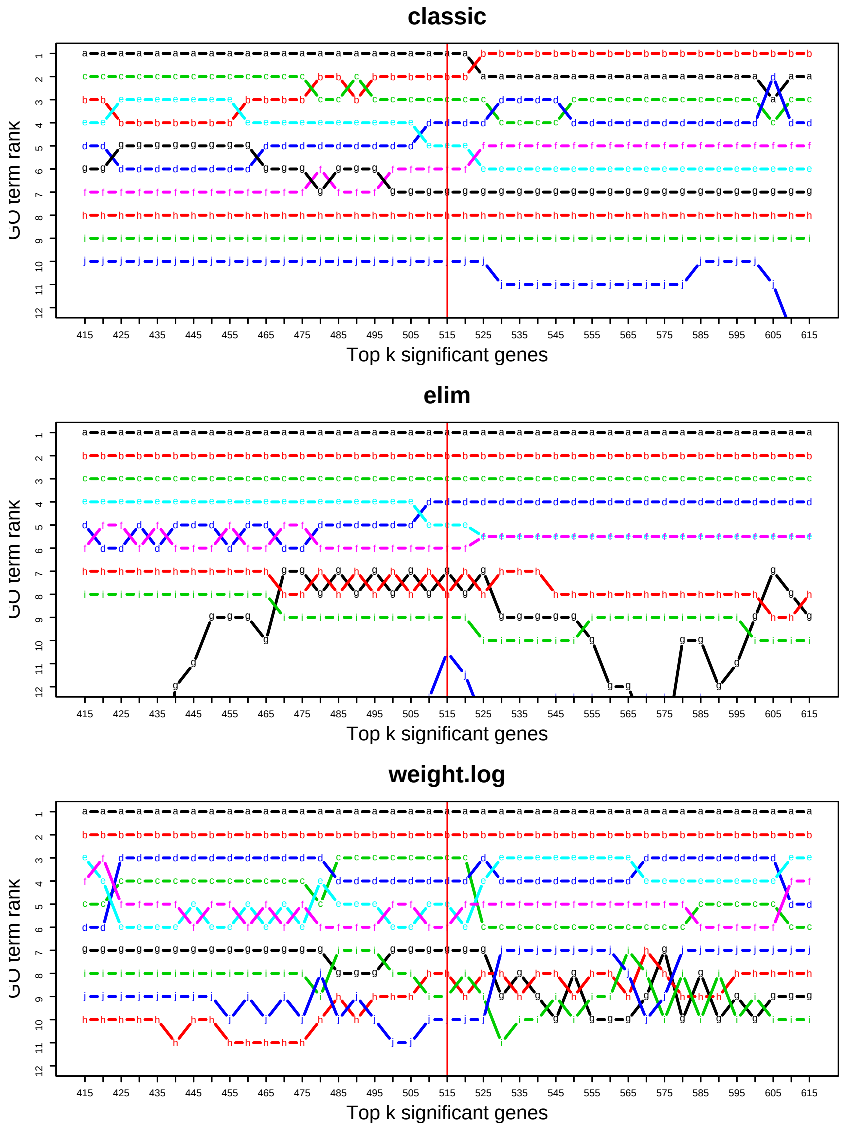
Influence of the p-values adjustment

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)

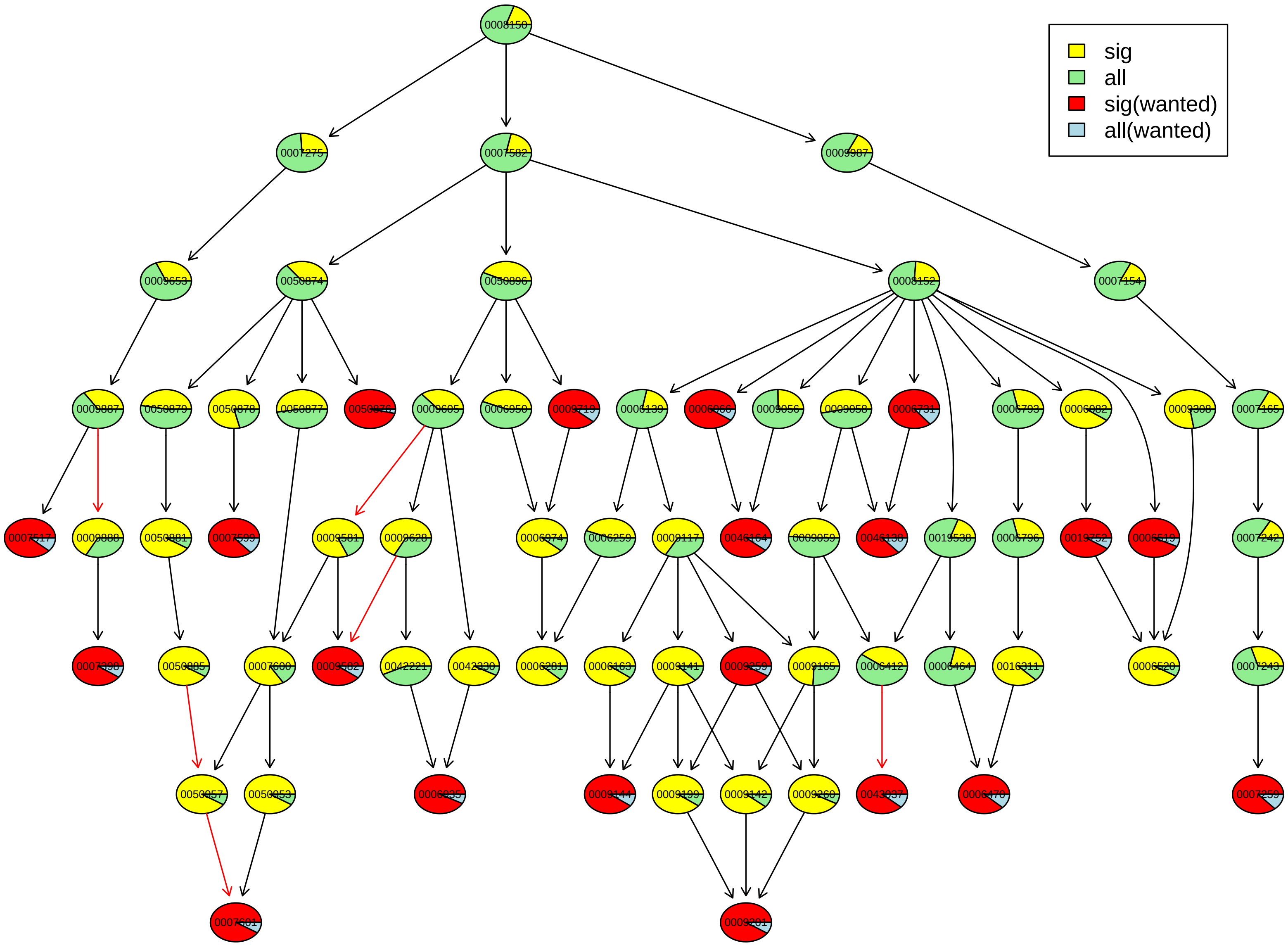


Evaluation on simulated data

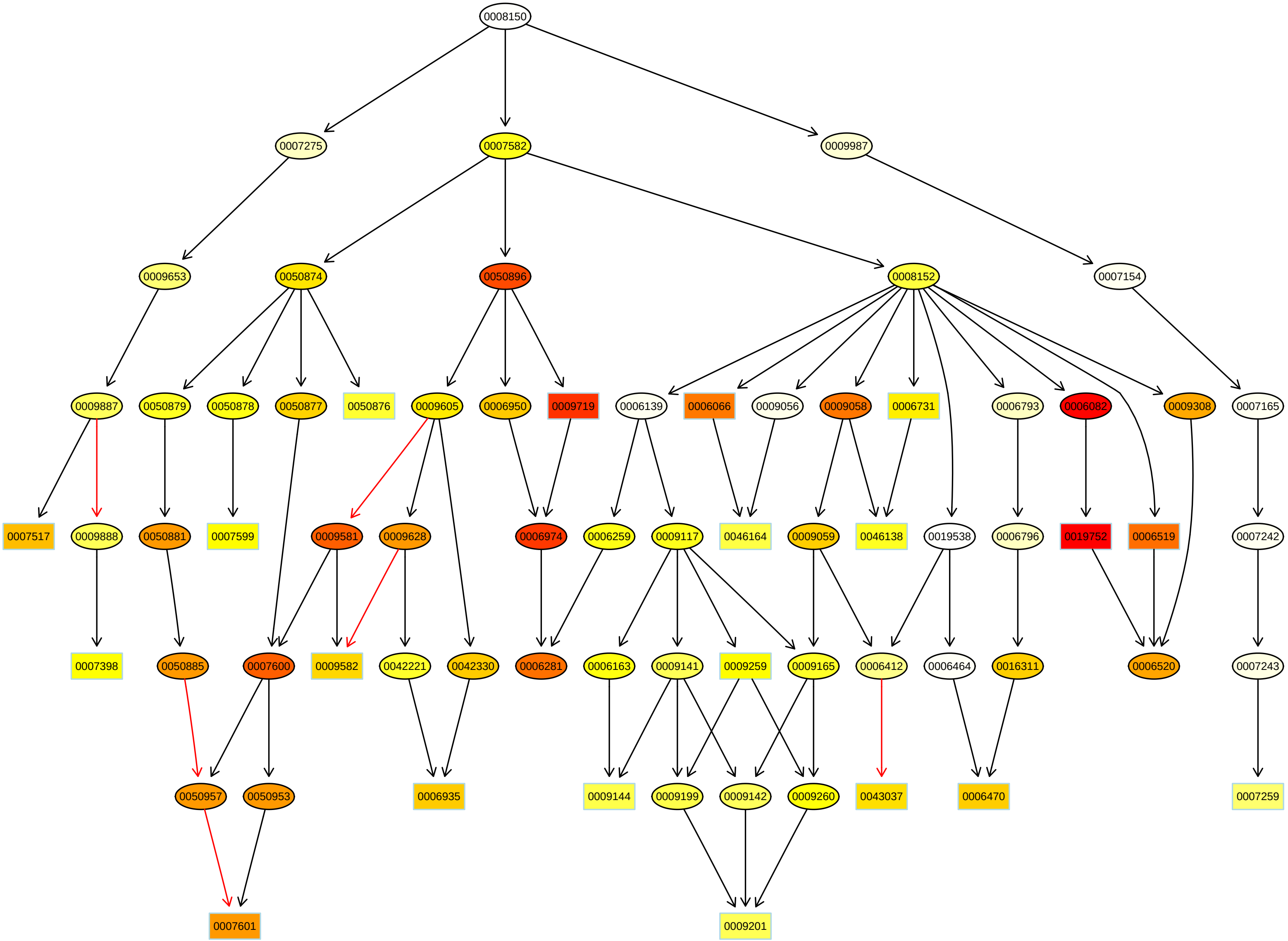
- We use the GO graph structure (2311 nodes), and all the genes from HGU95aV2 Affymetrix chip (9623 mapped to the GO graph)
- Select only the nodes that have the no. of mapped genes in some range $(10 \ldots 100)$
- Choose randomly a number of nodes (50 in our case) from the selected nodes. These nodes 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

- 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



Simulated dataset

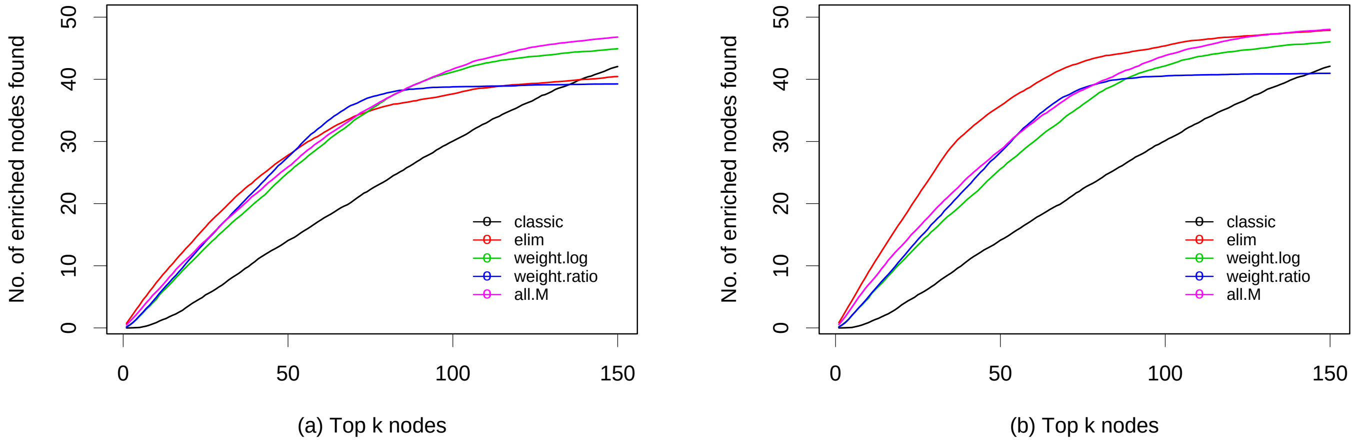


Quality of GO scoring methods

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

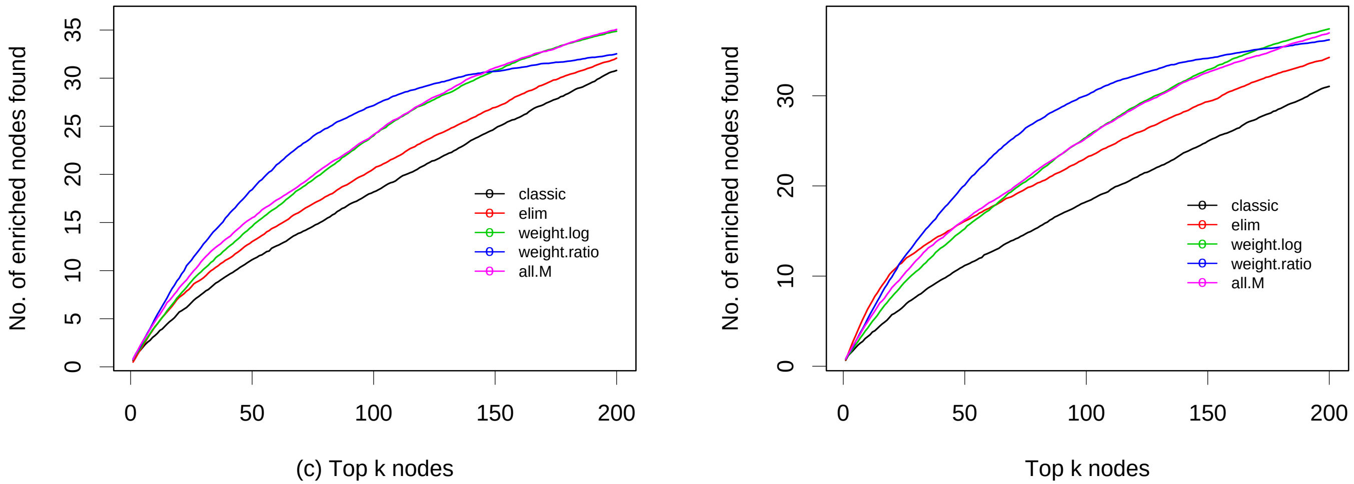
10% noise level.



10 to 1000 genes

annotated

40% noise level.



- Gene set enrichment
- Gene Ontology terms scoring
- Evaluation and stability of the methods
- **Conclusions**

max planck institut
informatik

Other proposed test statistics

- Local enrichment of GO terms [Grossmann et al., 2006]
- Goeman's global test [Goeman, J. J., *et al.*, 2004]
- ANCOVA approach [Mansmann and Meister, 2005]

Conclusions

- GO analysis performed on ALL data shows the methods are robust.
- Common biological processes to both studies, GO:0019884 and GO:0019886 underline the general importance of *antigen presentation* and *antigen processing* for ALL.
- Proposed methods perform better than current state-of-the-art methods even in more noisy conditions.
- The result of the methods is stable w.r.t. small variations of the cutoff, but a Kolmogorov-Smirnov like test is preferred.

[Cario, G., *et al.*, 2005] Cario, G., *et al.* (2005). Distinct gene expression profiles determine molecular treatment response in childhood acute lymphoblastic leukemia. *Blood*, 105(2):821–826.

[Chiaretti, S., *et al.*, 2004] Chiaretti, S., *et al.* (2004). Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival. *Blood*, 103(7):2771–2778.

[Goeman, J. J., *et al.*, 2004] Goeman, J. J., *et al.* (2004). A global test for groups of genes: testing association with a clinical outcome. *Bioinformatics*, 20(1):93–99.

[Grossmann et al., 2006] Grossmann, S., Bauer, S., Robinson, P. N., and Vingron, M. (2006). An improved statistic for detecting over-represented Gene Ontology annotations in gene sets. In *Proc. 10th Ann. Int. Conf. on Res. in Comput. Biol. (RECOMB '06)*, Venice.

[Khatri and Draghici, 2005] Khatri, P. and Draghici, S. (2005). Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, 21(18):3587–3595.

[Mansmann and Meister, 2005] Mansmann, U. and Meister, R. (2005). Testing differential gene expression in functional groups. *Methods Inf Med*, 3(44):449–453.

[Subramanian, A., *et al.*, 2005] Subramanian, A., *et al.* (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PNAS*, 102(43):15545–15550.