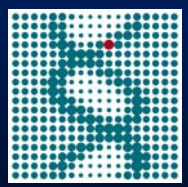
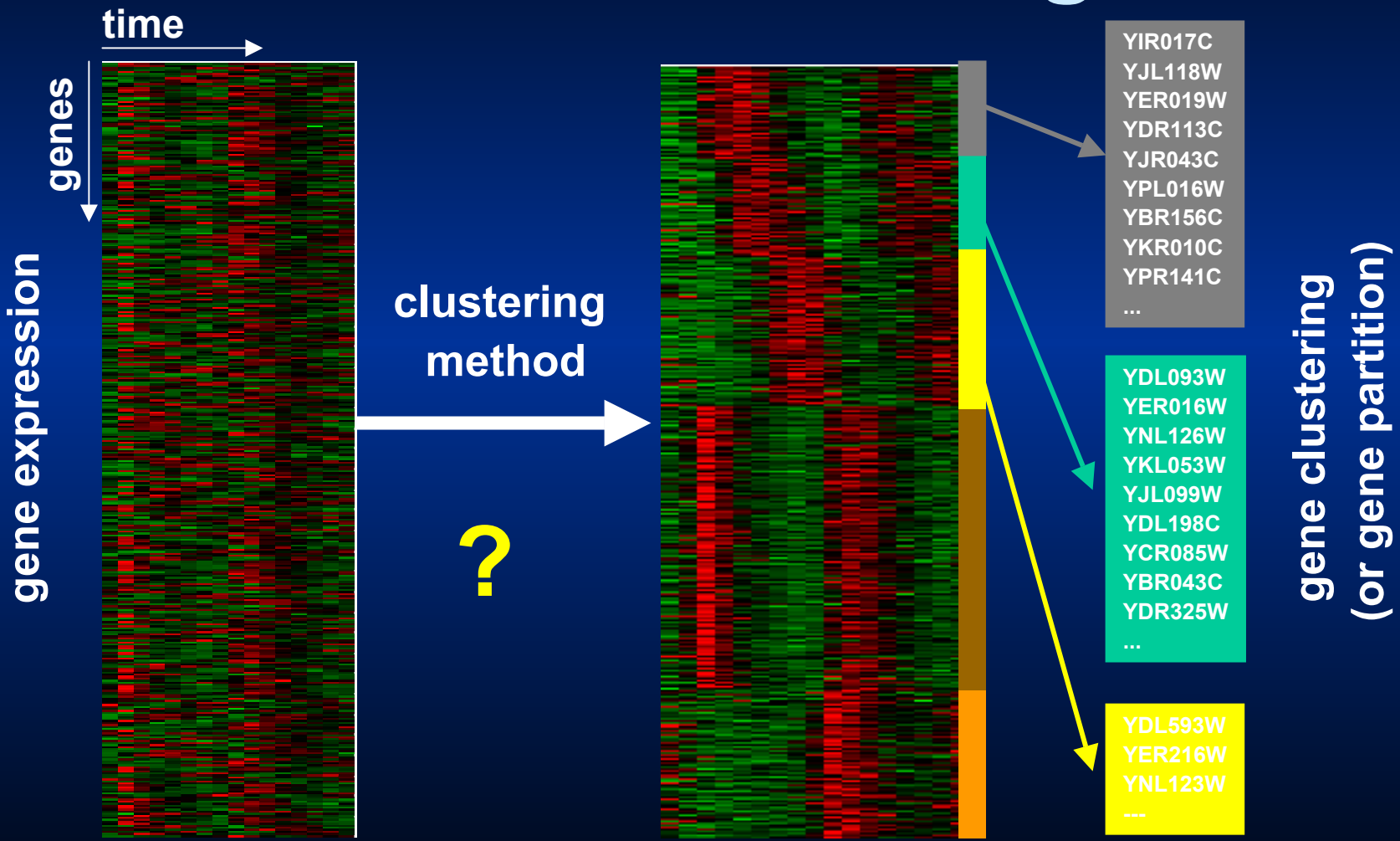
Validating Gene Clusterings by Selecting Informative Gene Ontology Terms with Mutual Information

Ivan G. Costa Filho
Marcilio C. P. de Souto
Alexander Schliep

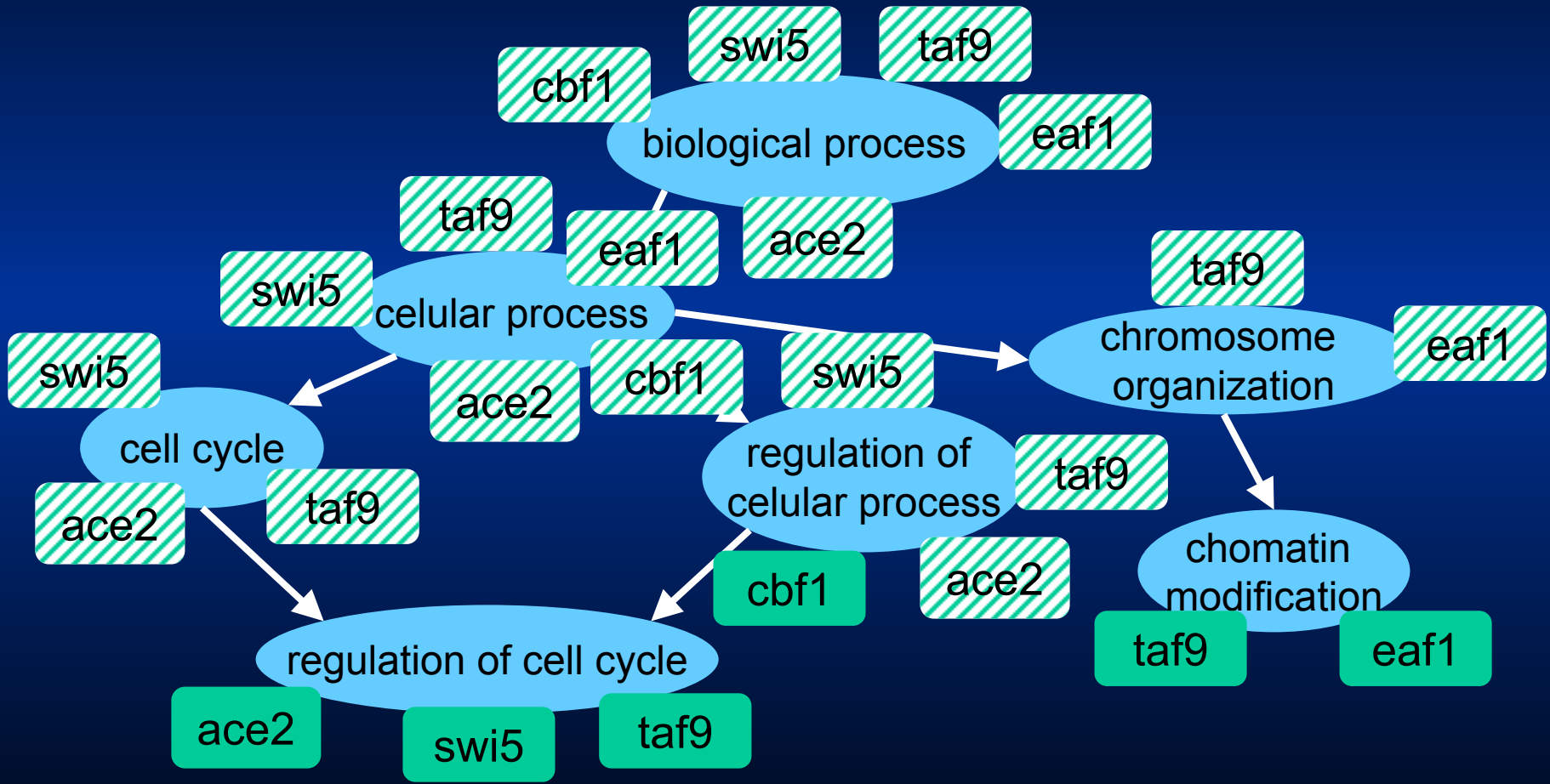


Computational Biology Department
Max Planck Institute for Molecular Genetics, Berlin

Gene Clustering



Gene Ontology (GO)



term-parent inheritance → **GO Term *redundancy***

Previous Work

- GO Term by Term enrichment (Beissbarth 2004, …)
  - statistical test to check how high is the count of genes for a GO Term
    + cluster interpretability
    - no cluster/clustering validation index
- GO Parent-Term enrichment (Grossman 2006)
  - refines GO Term by Term enrichment by correcting the test to parent-inheritance property
    ++ cluster interpretability
- ClusterJudge (Gibbons, *et. al* 2002)
  - uses a approximation of mutual Information to compare clustering with the complete GO annotation
    + gives a global fitness measure of a gene grouping
    - uses a heuristic parameter to reduce *redundancy* on GO Terms
    - gives no cluster *interpretability*

Goal

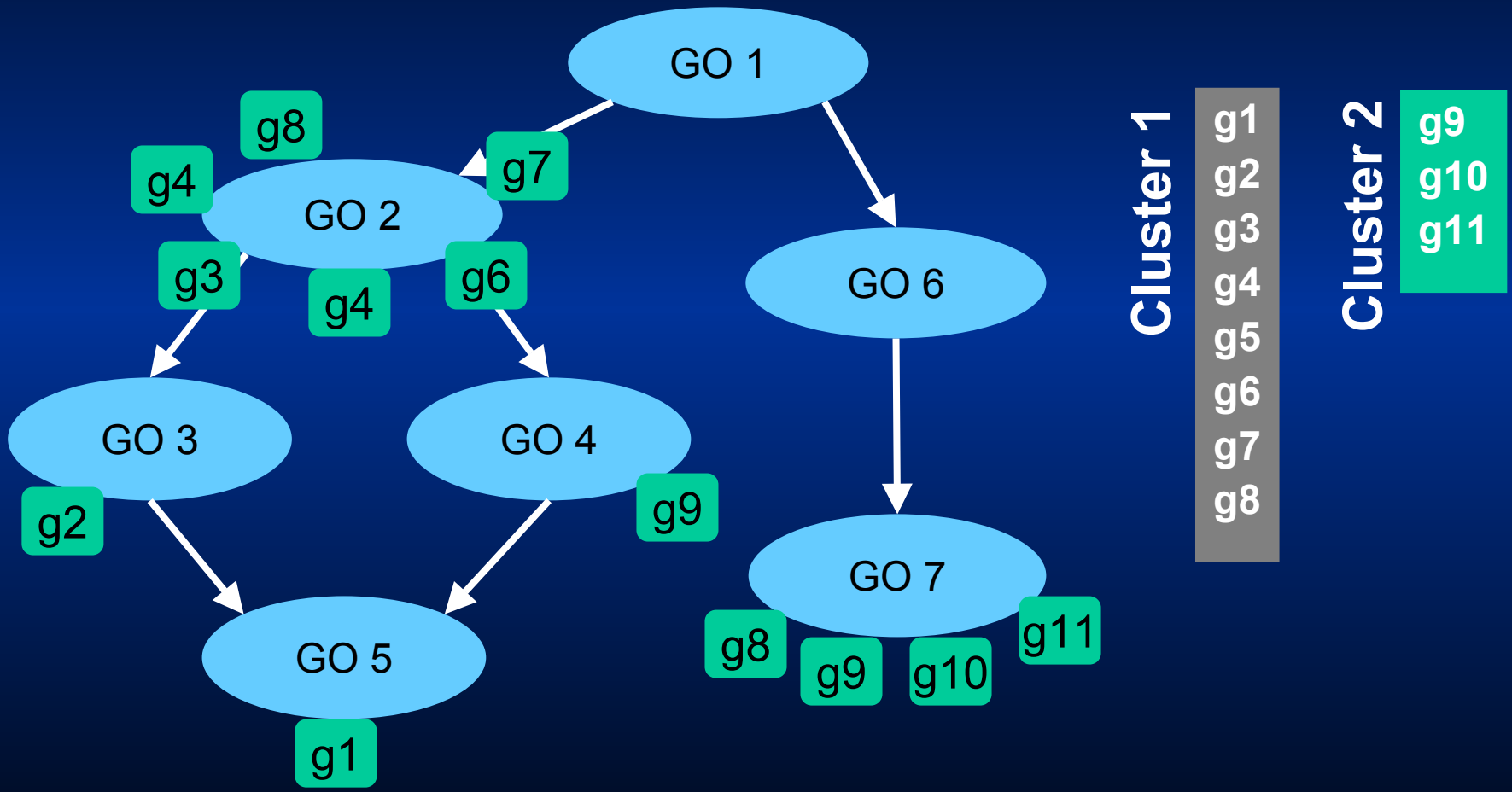
A index for global validation of gene clusterings by selection of *informative* and *non-redundant* GO Terms yielding individual cluster *interpretability*

Method

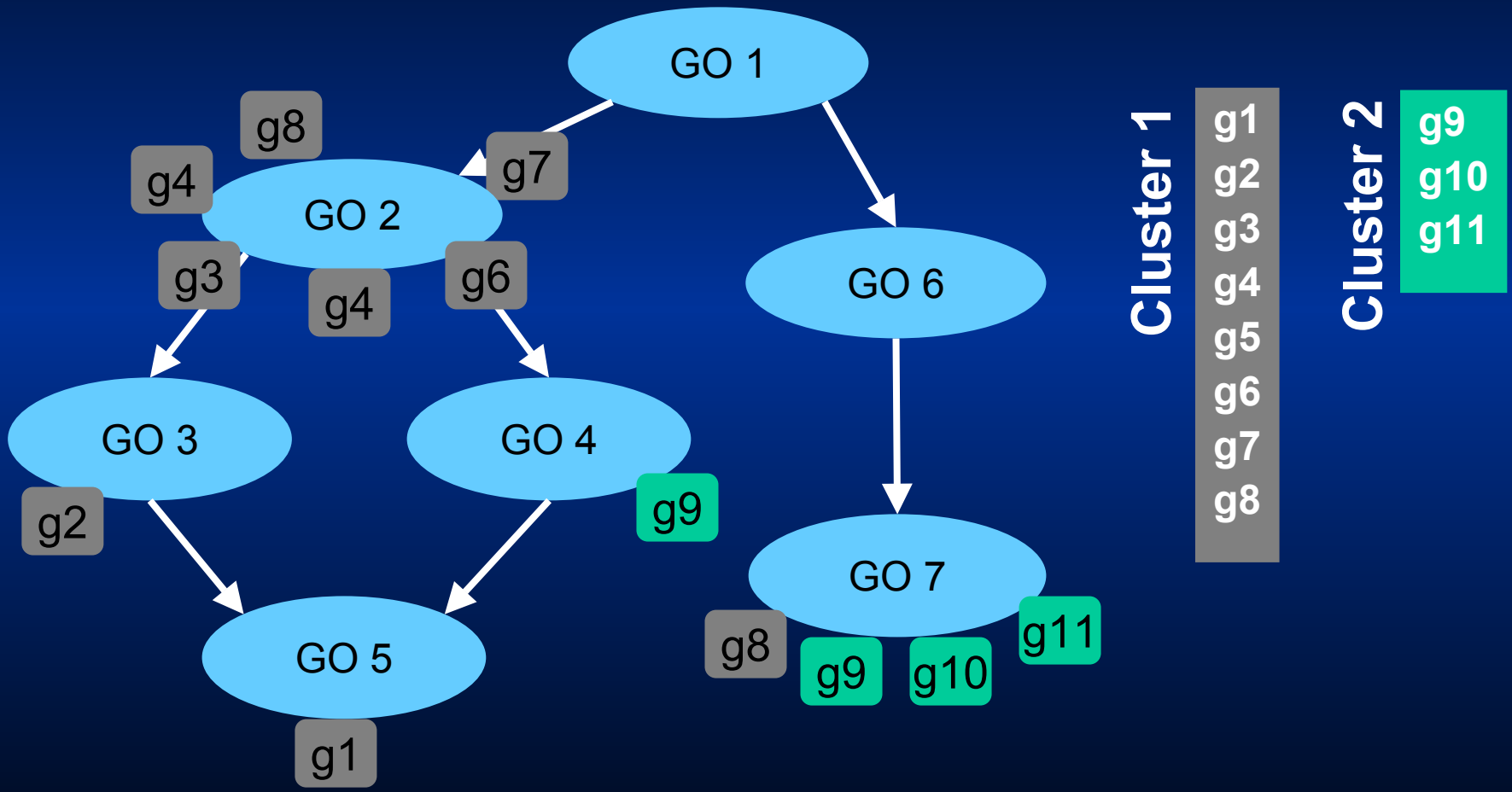
MutSel Outline

- Use the mutual information as a measure for selection of *informative* GO terms
- Take parent-child relation into account by exploring the DAG from leaves to the root to find *redundant* GO terms
- Calculate a global index (z-score) as in ClusterJudge using only *informative* and *non-redundant* GO Terms
- Use mutual information to relate *informative* and *non-redundant* GO Terms to clusters helping *interpretation*

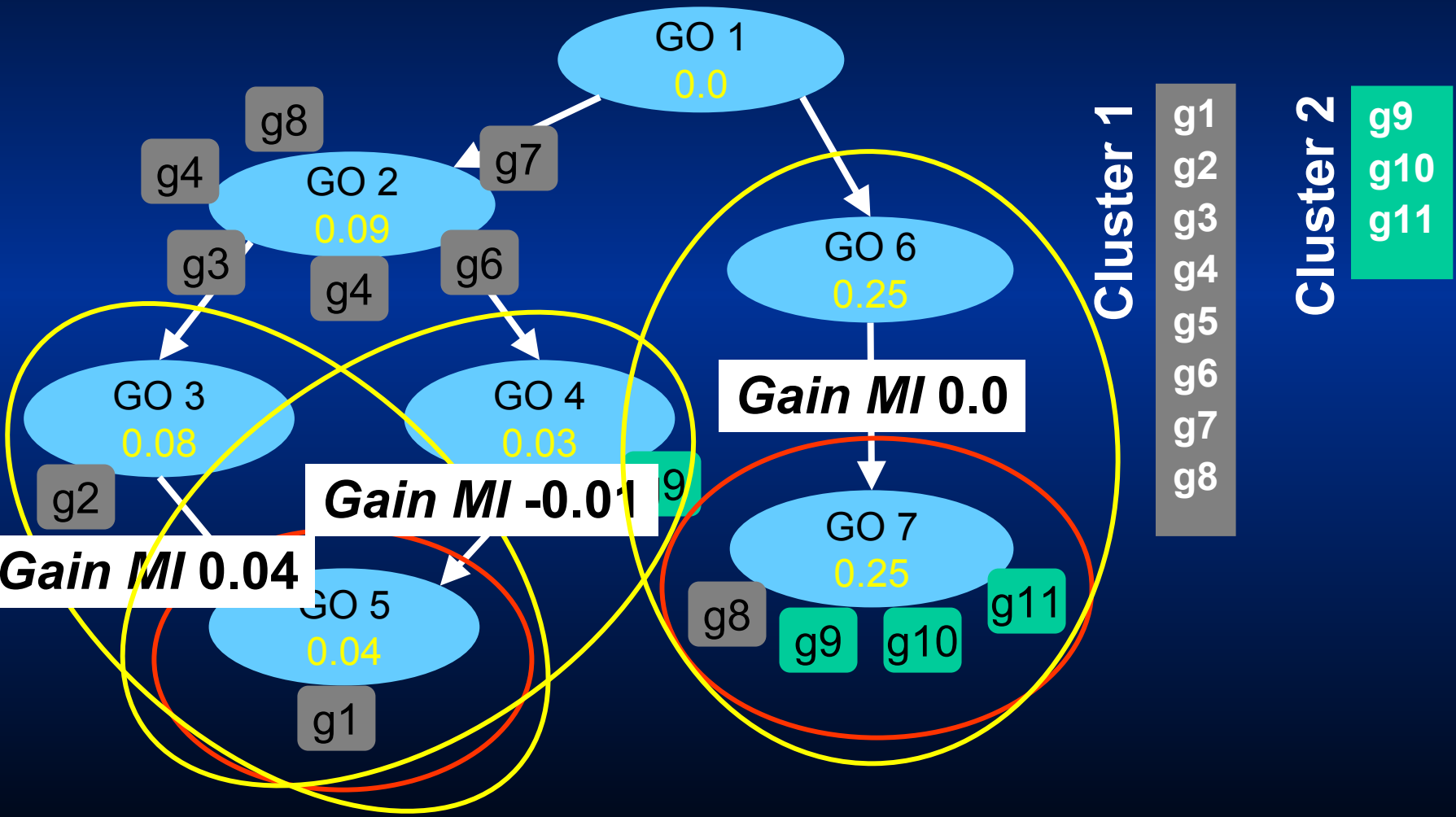
MutSel - Method Illustration



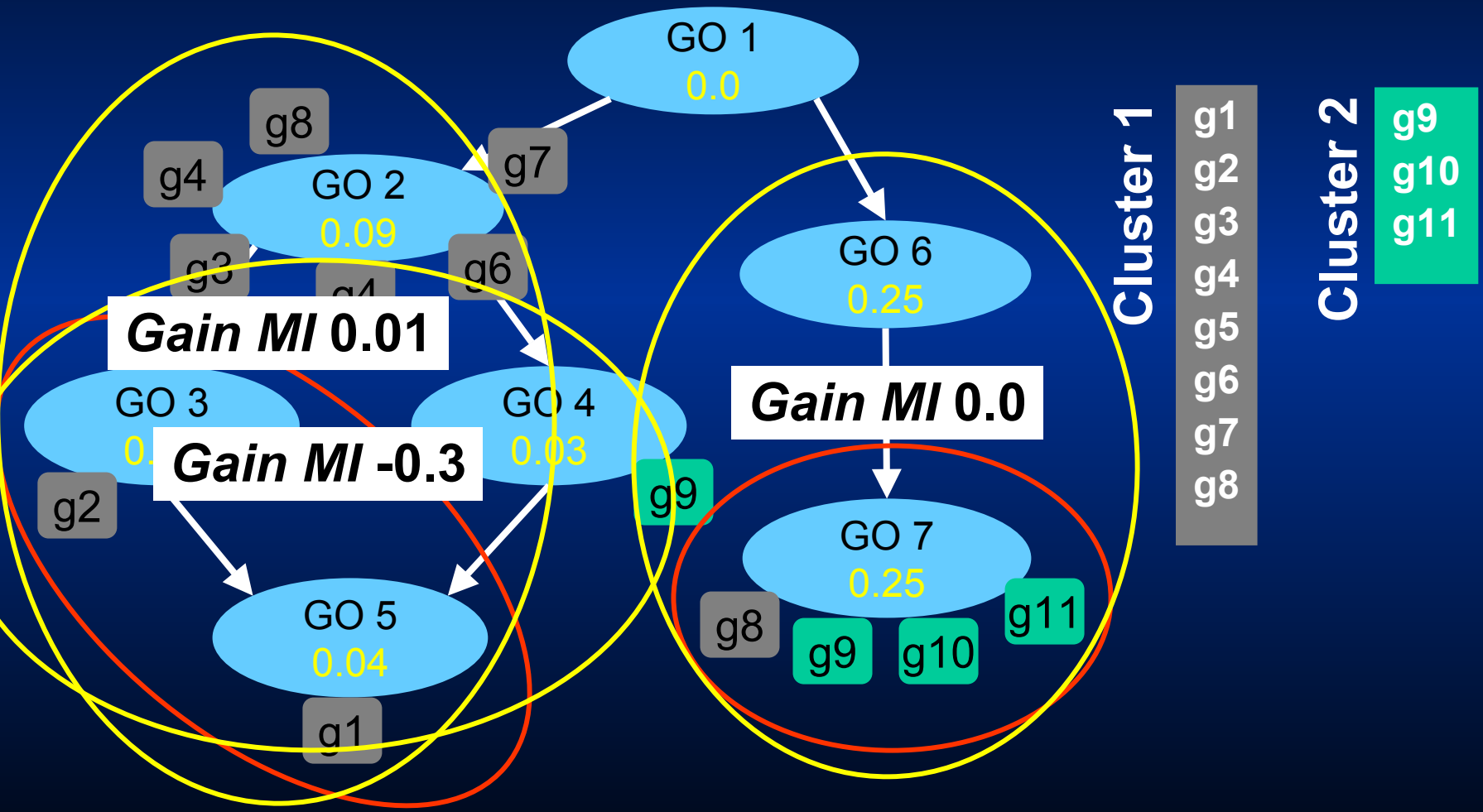
MutSel - Method Illustration



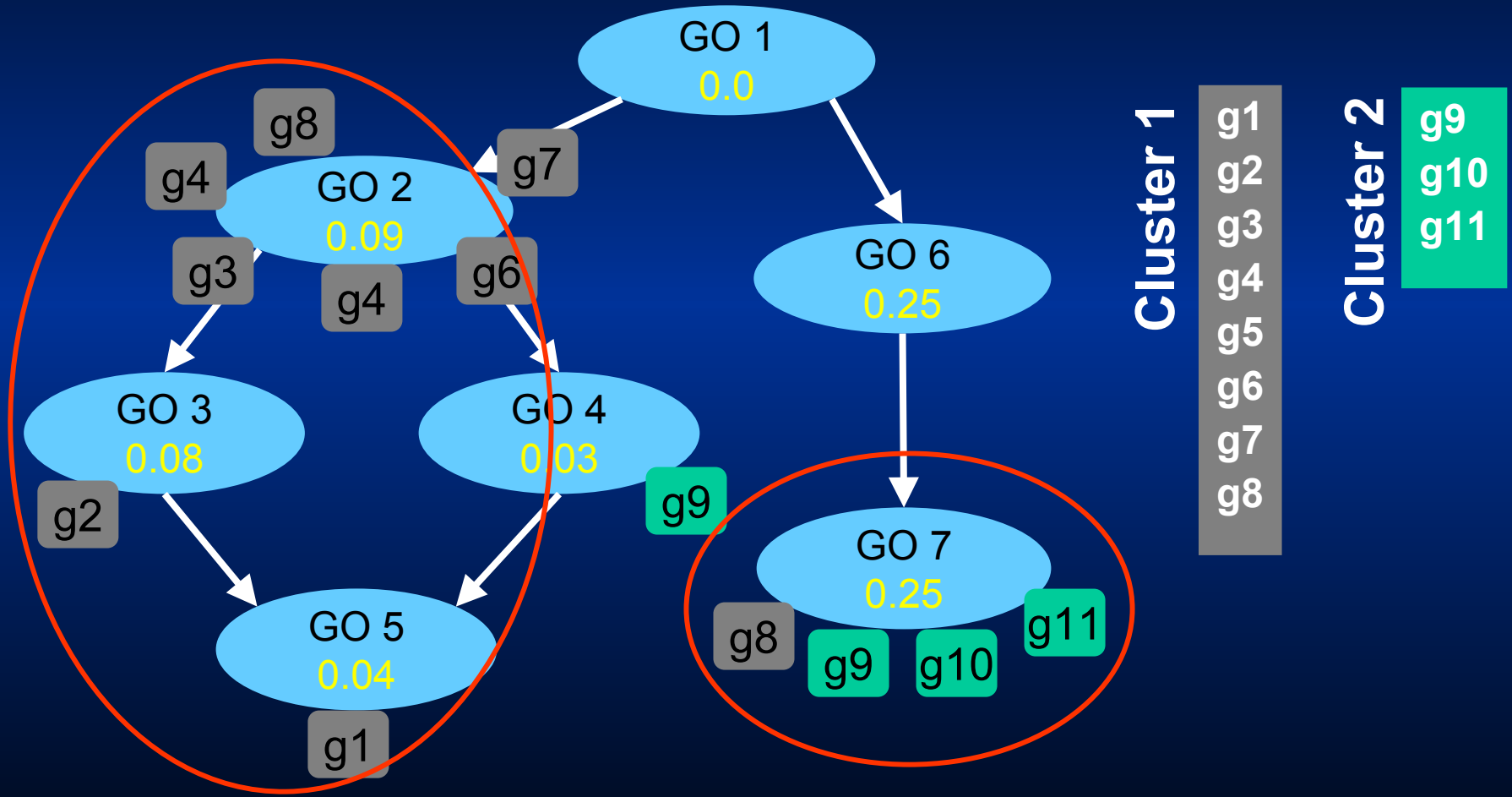
MutSel - Method Illustration



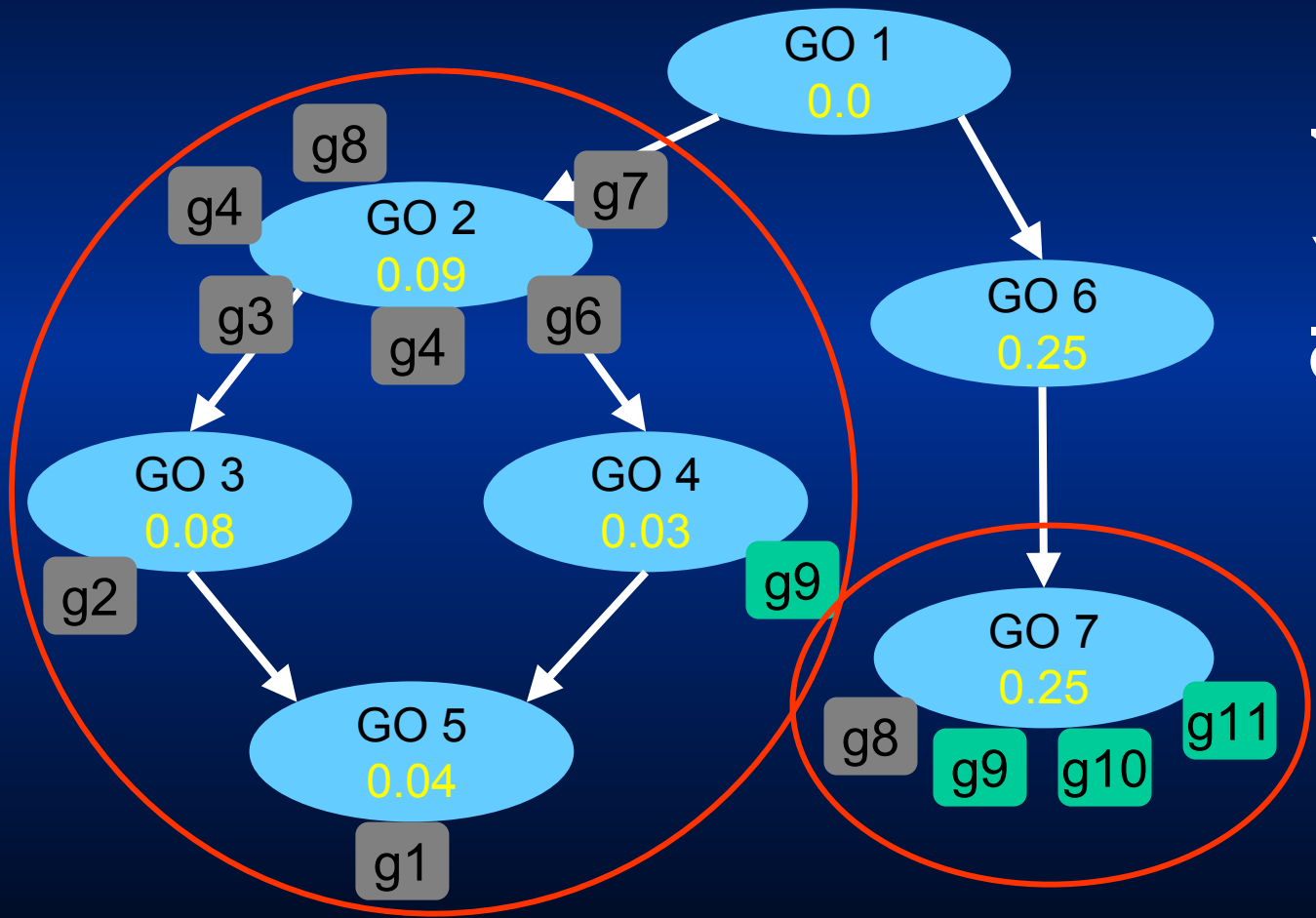
MutSel - Method Illustration



MutSel - Method Illustration



MutSel - Method Illustration



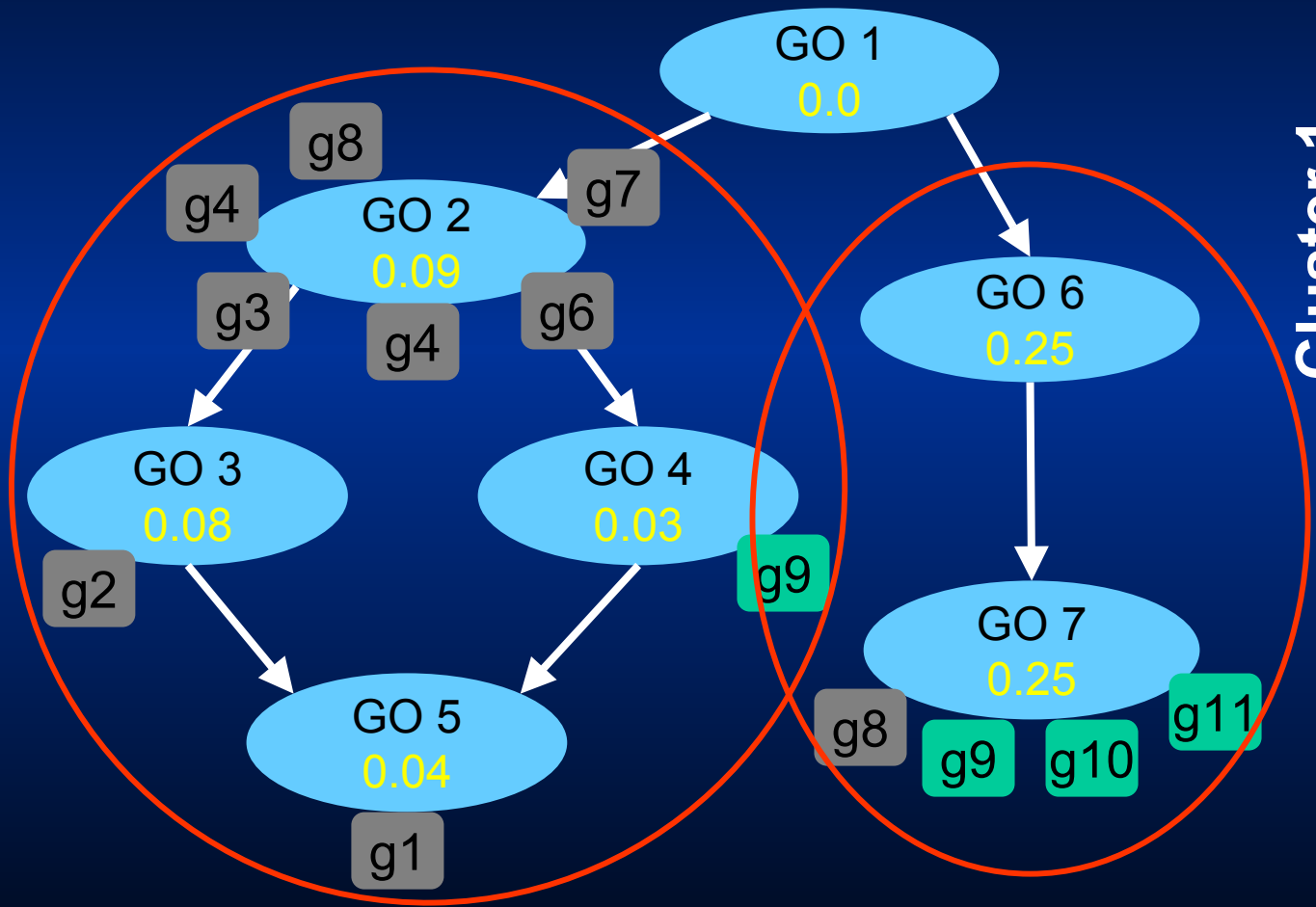
Cluster 1

g1
g2
g3
g4
g5
g6
g7
g8

Cluster 2

g9
g10
g11

MutSel - Method Illustration



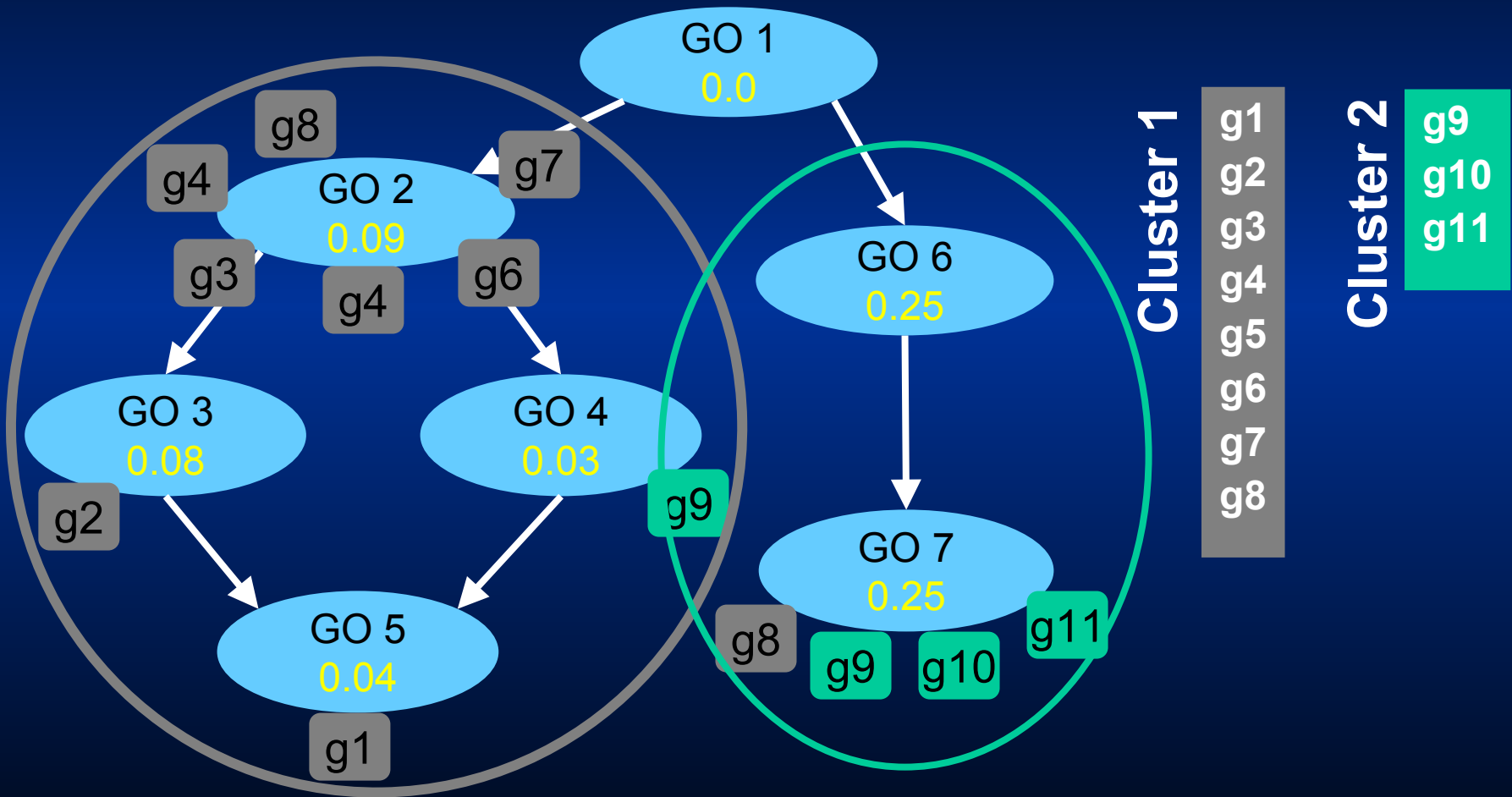
Cluster 1

g1
g2
g3
g4
g5
g6
g7
g8

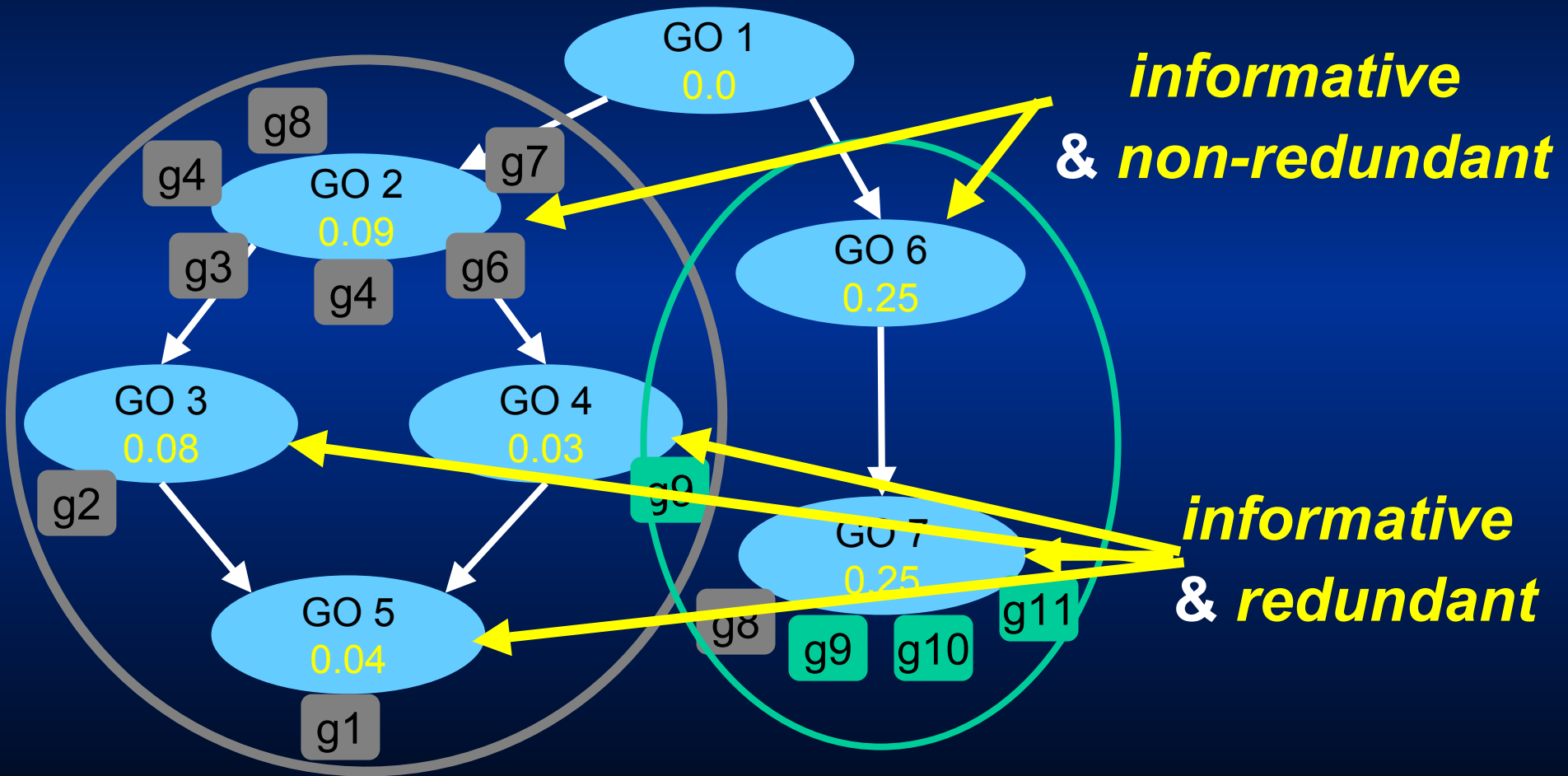
Cluster 2

g9
g10
g11

MutSel - Method Illustration



MutSel - Method Illustration



Experiments

Experiment
Biological Interpretability

- Yeast Treatment SM (Jie *et. al*, 2002)
  - Time course after treatment with a inhibitor of nucleic acid synthesis
  - 241 induced genes and 121 repressed genes
  - Simple 2 cluster problems in a well characterized biological scenario

Results MutSel
Biological Interpretability

Induced Genes

– amino acid and nitrogen metabolism *(Jie et al.)*

GO Term	*#I*	*#R*	*MI*
nitrogen compound metabolic process	68	14	0.022
vitamin biosynthetic process	14	0	0.022
amino acid and derivative metabolic process	60	13	0.018
transferase activity, transferring nitrogenous groups	11	0	0.017

Repressed Genes

– carbohydrate and lipid biosynthesis, ribossomes, cell cycle *(Jie et al.)*

macromolecule biosynthetic process	16	36	0.051
cell division	3	10	0.019
lipid biosynthetic process	4	11	0.018
ribonucleoprotein complex biogenesis and assembly	2	7	0.014

Methods Comparison Biological Interpretability

Go Term Sel. Method	*Induced*	*Repressed*
Parent-Term Enrichment	41	79
MutSel	13	22
MutSel ∩ Parent-Term	11 (85%)	18 (81%)

Term by Term Enrichment	79	159
MutSelAll (redundant Terms)	39	80
MutSelAll ∩ Term-Term	33 (84%)	80 (100%)

MutSel selects a sub-set of GO terms from established enrichment methods

Experiment - Clustering Method Comparison

- Yeast Cell cycle (Cho, 1998)
  - time-courses of 384 genes during mitotic cell division in Yeast
  - expert classification into five cell-cycle phases
- Clustering Methods:
  - hierarchical, *k*-means, mixture of Gausians and mixtures of HMMs
- Validation indices:
  - MutSel
  - ClusterJudge with several GO Term redundancy (*U*) parameter choices
  - Corrected Rand (using expert classification)

Results – Yeast Cell Cycle

Indices	*Rank 1*	*Rank 2*	*Rank 3*	*Rank 4*
ClusterJudge U=0.8	*k*-means	MixHMM	Hier.	MixGaus
ClusterJudge U=0.4	*k*-means	Hier.	MixHMM	MixGaus
ClusterJudge U=0.2	*k*-means	MixHMM	MixGaus	Hier.
ClusterJudge U=0.1	*k*-means	MixGaus	MixHMM	Hier.
ClusterJudge U=0.01	*k*-means	MixGaus	Hier.	MixHMM
MutSel	*k*-means	MixGaus	MixHMM	Hier.
Corrected Rand	*k*-means	Hier.	MixGaus	MixHMM

Results – Yeast Cell Cycle

indicates random solutions!

Indices	*Rank 1*	*Rank 2*	*Rank 3*	*Rank 4*
ClusterJudge U=0.1	*k*-means 0.86	MixGaus 0.37	MixHMM 0.36	Hier. -0.17
ClusterJudge U=0.01	*k*-means 1.4	MixGaus 0.83	Hier. 0.64	MixHMM -0.1
MutSel	*k*-means 1115.3	MixGaus 1034.0	MixHMM 791.9	Hier. 616.3
Corrected Rand	*k*-means 0.50	Hier. 0.46	MixGaus 0.43	MixHMM 0.39

MutSel stops at leaves at random solutions!

Conclusions

- MutSel was successful in selecting *informative* and *non-redundant* GO terms for global validation and individual cluster *interpretation*.
- MutSel make the selection of GO terms in a deterministic and principled way improving ClusterJudge

Outlook

- Large scale clustering evaluation with several data sets and clustering methods
- Accomplishment of a web tool.

Thanks.

- Joachin Selbig (MPI for Molecular Plant Physiology) and Steffen Grossman (MPI for Molecular Genetics) for helpful discussions.
- CNPq (Brasil) and DAAD (Germany) for funding.

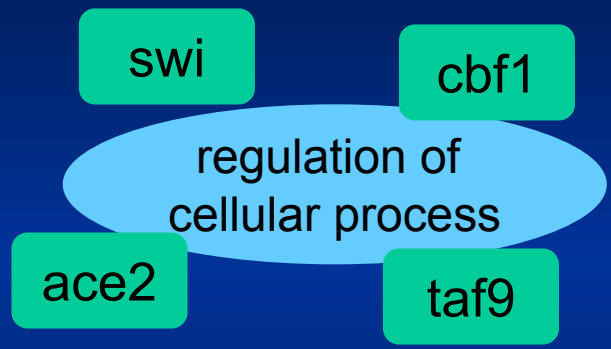
Previous Work
GO Term by Term Enrichment

Gene Cluster

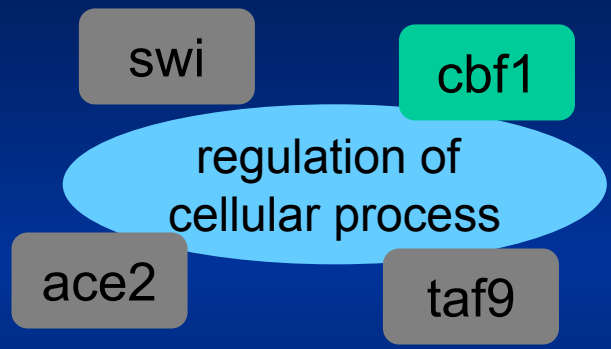
SWI
ACE2
CBF1
YJL099W
YDL198C
YCR085W
YCR043C
YDR825C

Reference set

YDL093W
YER016W
YNL126W
YKL053W
YJL099W
YDL198C
YCR085W
YBR043C
YDR325W
YCR085W
YBR043C
...



Previous Work
GO Term by Term Enrichment



8 out of 40 genes **3 out of 4 genes**

Perform a hypergeometric test to check odds of this event

\+ helps cluster *interpretability*

\- assumes independence on GO Terms

\- no suitable for groupings

\- multiple testing problems

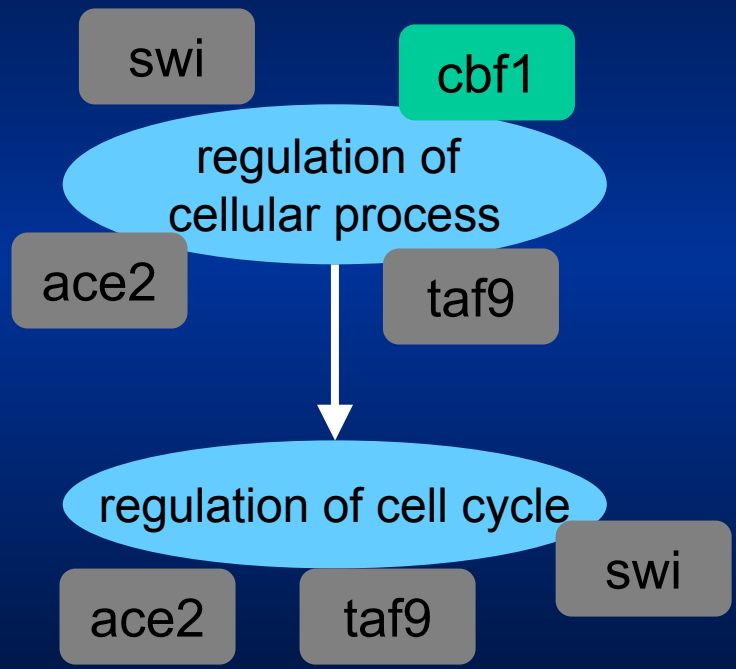
Previous Work
GO Parent-Term Enrichment

Gene Cluster

SWI
ACE2
CBF1
YJL099W
YDL198C
YCR085W
YCR043C
YDR825C

Reference set

YDL093W
YER016W
YNL126W
YKL053W
YJL099W
YDL198C
YCR085W
YBR043C
YDR325W
YCR085W
YBR043C
...



correction of hypergeometric test given the inheritance property (Grossmann, 2006)

++ helps cluster *interpretability*

- no suitable for groupings
- multiple testing problems

Previous Work
Global Indices

ClusterJudge (Gibbons, *et. al* 2002)

Uses a approximation of mutual Information to compare a clustering with the complete GO annotation

+ index based on a information-theoretic concept
+ gives a global fitness measure of a gene grouping
- assumes independence of GO terms
- uses a heuristic parameter to reduce *redundancy* on GO Terms
- gives no cluster *interpretability*

Method - Validation Index

Given Y, where $y_i=k$ means that the ith gene belongs to cluster k

X^p, where $x_i^p=1$ means GO term p annotates the ith gene

The mutual information is defined as

$$MI(X^P,Y)=\sum_{x=0}^{1}\sum_{y=1}^{K}P[X^P=x,Y=y]\log\frac{P[X^i=x,Y=y]}{P[X^i=x]P[Y=y]}$$

For a set of GO terms $X=\{X,\ldots,X^P\}$, the global index can be calculated as

$$MI^{app}(X,Y)=\sum_{p=1}^{P}MI(X^P,Y)$$

To quantify deviation from randomness, we calculate z scores

$$z^{MI^{app}(X,Y)}=\frac{MI^{app}(X,Y)-Mean(MI^{app}(X,Y^{random}))}{\sqrt{Var(MI^{app}(X,Y^{random}))}}$$