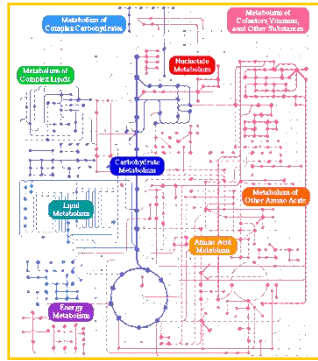


Intégration de données hétérogènes Enrichissement

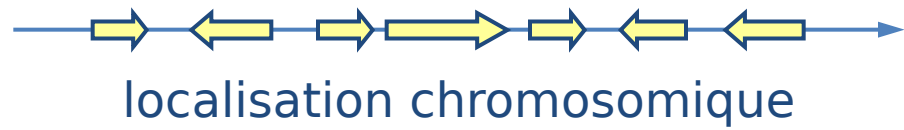
Master 2

Bioinformatique et Biologie des Systèmes

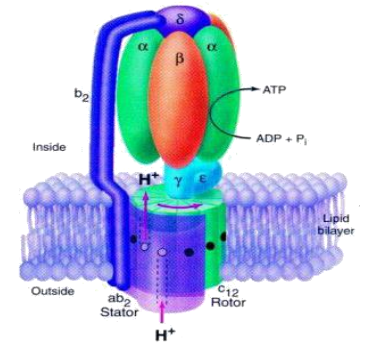
Recoupement de voisinages : approche ensembliste



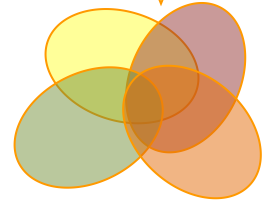
voies métaboliques



localisation chromosomique



complexes protéiques



ensembles de gènes

Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 2008

Nucleic Acids Research Advance Access published May 28, 2008
Nucleic Acids Research, 2008, 1-8
 doi:10.1093/nar/gkn235

ENDEAVOUR update: a web resource for gene prioritization in multiple species

Léon-Charles Tranchevent¹, Roland Barriot¹, Shi Yu¹, Steven Van Vooren¹, Peter Van Loo^{1,2,3}, Bert Coessens¹, Bart De Moor¹, Stein Aerts^{3,4} and Yves Moreau^{1,*}

¹Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, ²Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB Leuven, ³Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and ⁴Laboratory of Neurogenetics, Department of Molecular and Developmental Genetics, VIB, Leuven (Belgium)

Received February 7, 2008; Revised April 30, 2008; Accepted May 7, 2008

ABSTRACT
 Endeavour (<http://www.esat.kuleuven.be/endeavour> web; this web site is free and open to all users and there is no login requirement) is a web resource for the prioritization of candidate genes. Using a training set of genes known to be involved in a biological process of interest, our approach consists of (i) inferring several models based on various genomic data sources, (ii) applying each model to the candidate genes to rank those candidates against the profile of the known genes and (iii) merging the several rankings into a global ranking of the candidate genes. In the present

BACKGROUND
 With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, covering genomic data into biological knowledge to identify genes involved in a particular process or disease remains a major challenge. Nevertheless, there is much evidence to suggest that functionally related genes often cause similar phenotypes (1-3). To identify which genes are responsible for which phenotype, association studies and linkage analyses are often used, resulting in large lists of candidate genes. In

co-citation



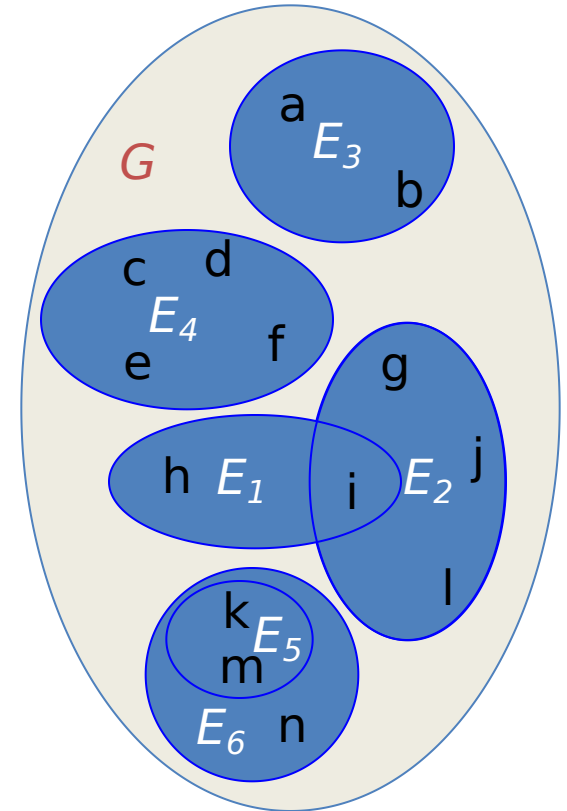
domaines protéiques



Gene Ontology

Définitions

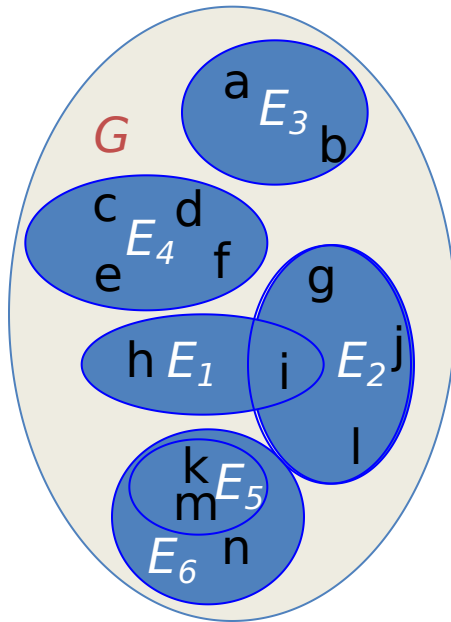
- (Identifiants de) gène → ARNm → protéine
- G : ensemble des gènes d'un organisme
- *Fonction de regroupement* : relation entre gènes basée sur un indice de similarité.
- *Ensemble de (gènes) voisins* : ensemble de gènes $E \subseteq G$ regroupés par une fonction de regroupement.
- *Voisinage* : sous-ensemble de $P(G)$ formant un ensemble d'ensembles de voisins, $V \subseteq P(G)$, regroupés par une même fonction de regroupement.



$$V = \{E_1, E_2, E_3, E_4, E_5, E_6\} \subseteq P(G)$$

Représentation d'un voisinage : ordre partiel (poset)

- Un voisinage est un ensemble (d'ensembles de voisins) ordonné par la relation d'inclusion \subseteq



$$V = \{E_1, E_2, E_3, E_4, E_5, E_6\}$$

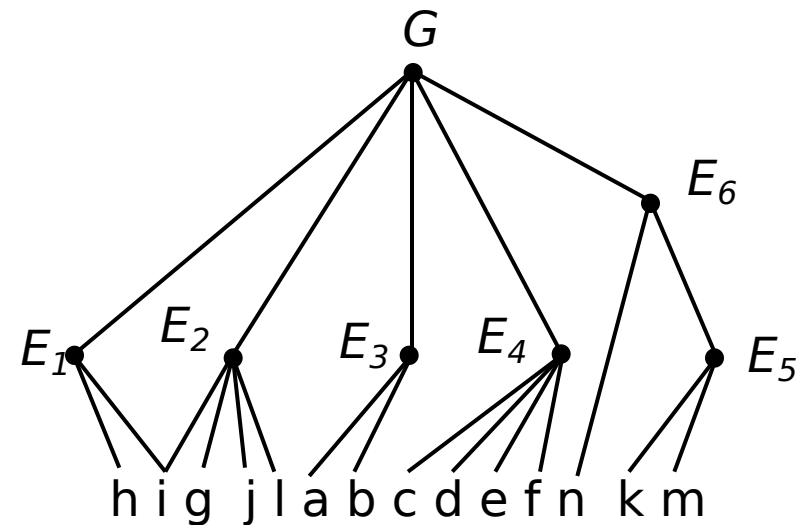
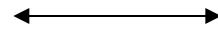
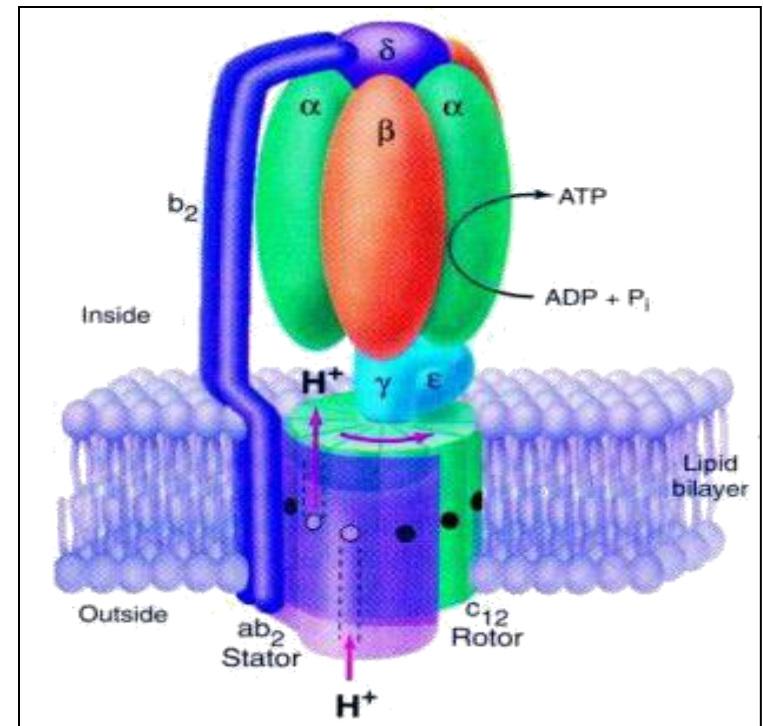
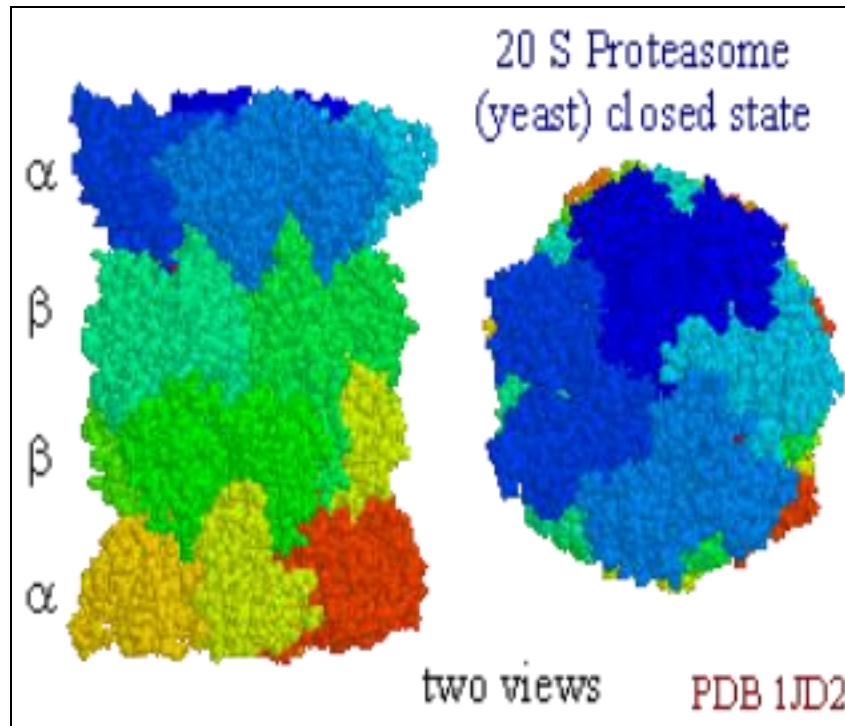


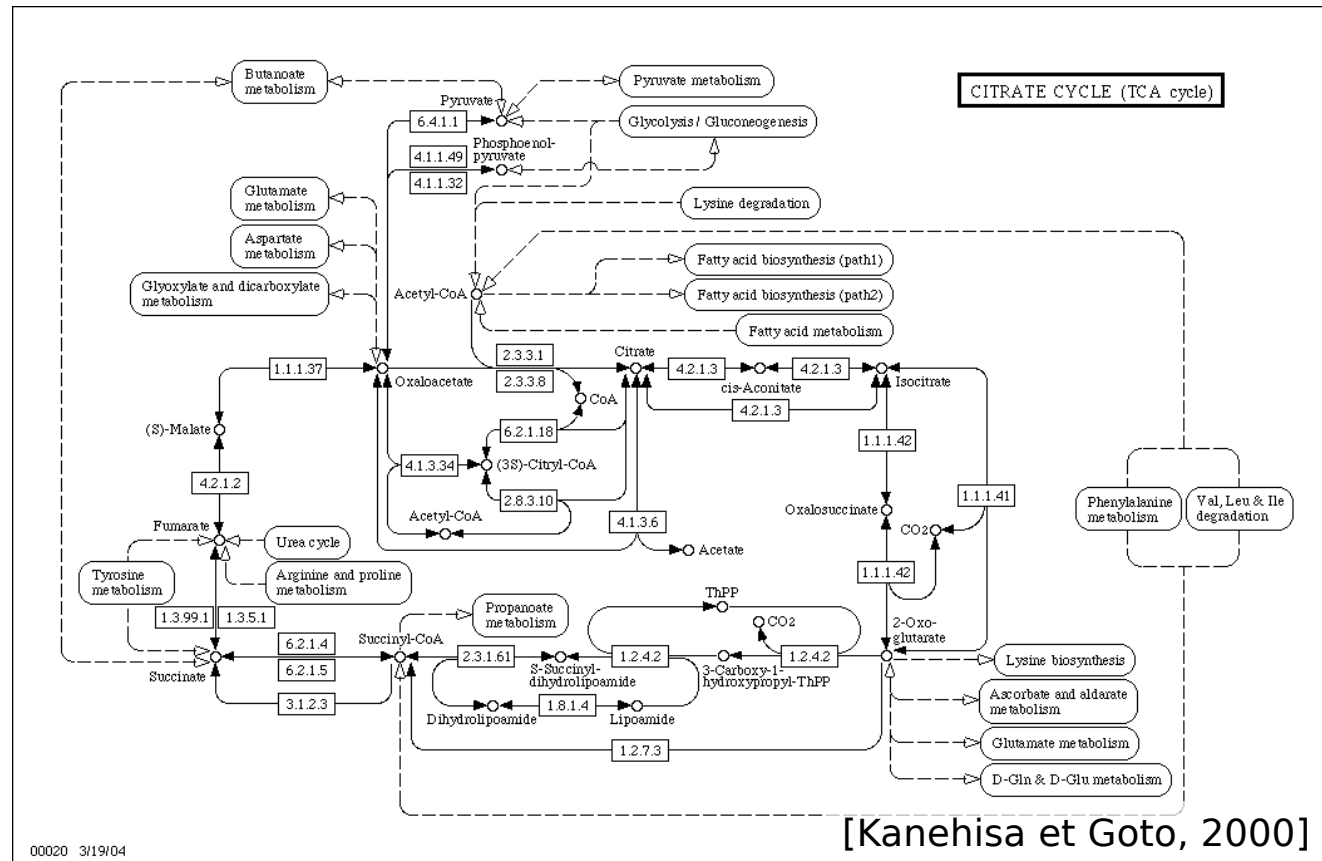
diagramme de Hasse de V

Exemple de fonction de regroupement : complexes protéiques



un complexe → un ensemble de protéines

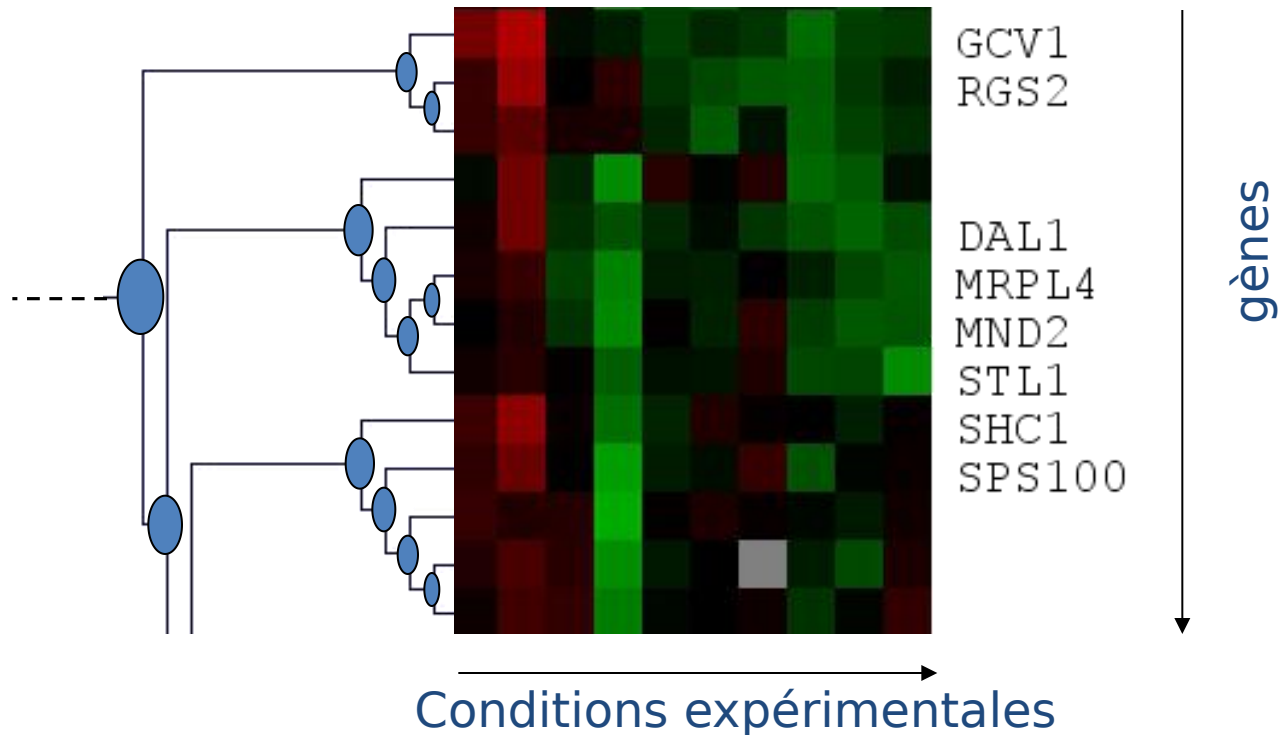
Exemple de critère de regroupement : voies métaboliques



une voie métabolique → un ensemble de protéines

Exemple de critère de regroupement : données d'expression

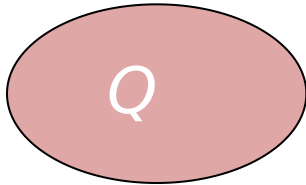
clustering hiérarchique
des profils



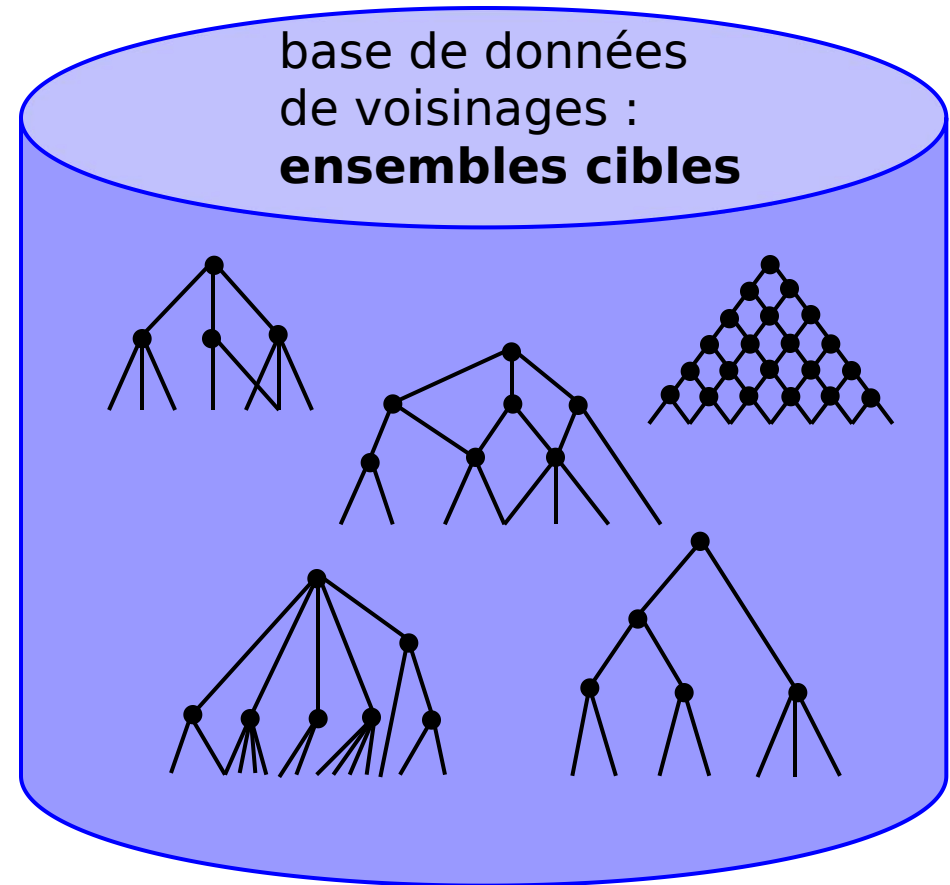
un cluster → un ensemble de gènes

- Recherche d'ensembles similaires

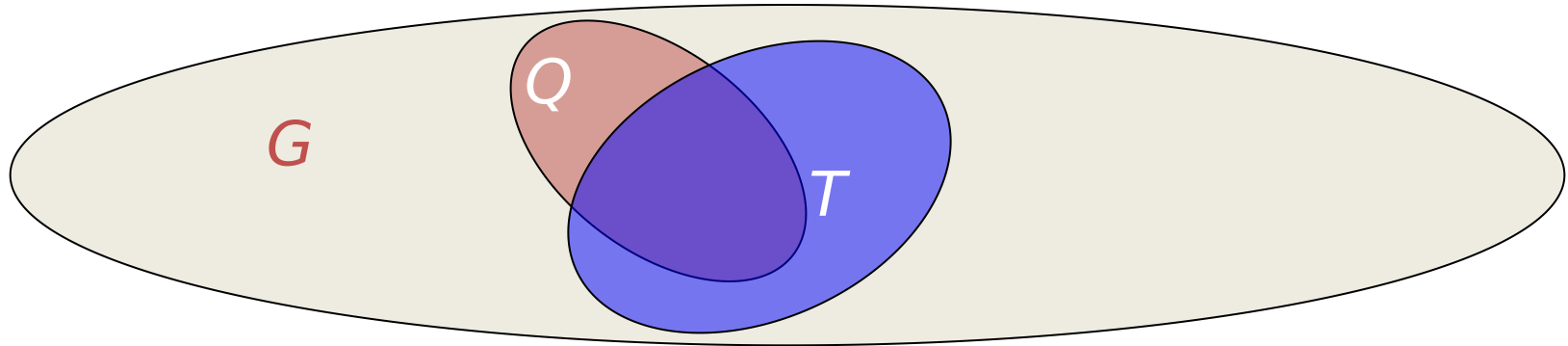
ensemble requête
 $Q \subseteq G$



**Quels sont les
ensembles cibles
qui lui sont similaires ?**



Mesure de (dis)similarité



- Loi hypergéométrique : probabilité d'avoir au moins le nombre d'éléments communs observé entre 2 échantillons issus d'une même population

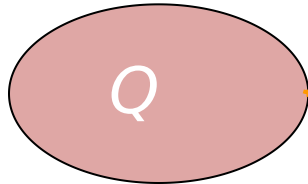
$$p\text{-valeur}(c, t, q, g) = \sum_{k=c}^{\min(q, t)} \frac{\binom{t}{k} \binom{g-t}{q-k}}{\binom{g}{q}}$$

avec

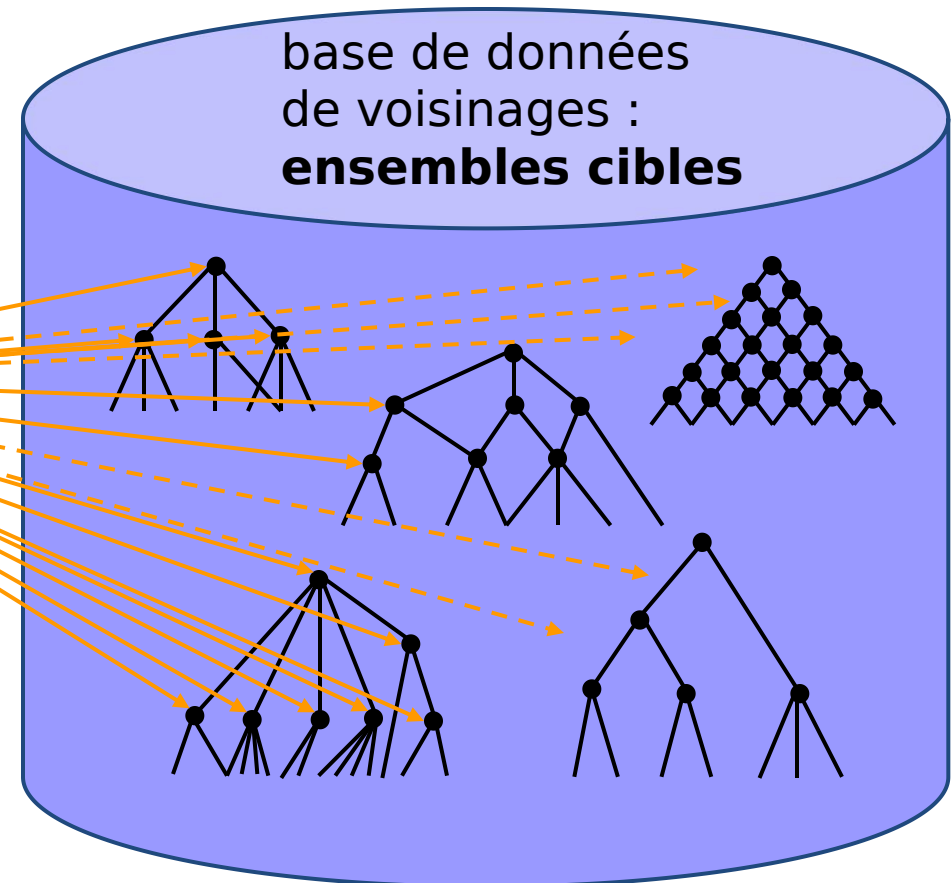
- $g = |G|$: taille de la population
 - $q = |Q|$: taille de l'ensemble requête
 - $t = |T|$: taille de l'ensemble cible
 - $c = |Q \cap T|$: nombre d'éléments communs
- Autres mesures :
 - Loi binomiale
 - χ^2
 - ratio, pourcentage

- Recherche d'ensembles similaires

ensemble requête
 $Q \subseteq G$



Quels sont les
ensembles cibles
qui lui sont similaires ?



Tests multiples et significativité des p-valeurs

- Probabilité d'obtenir une p-valeur aussi faible par hasard : fonction de répartition des p-valeurs minimales
- Simulations

RandomSet_1, minPi = M1

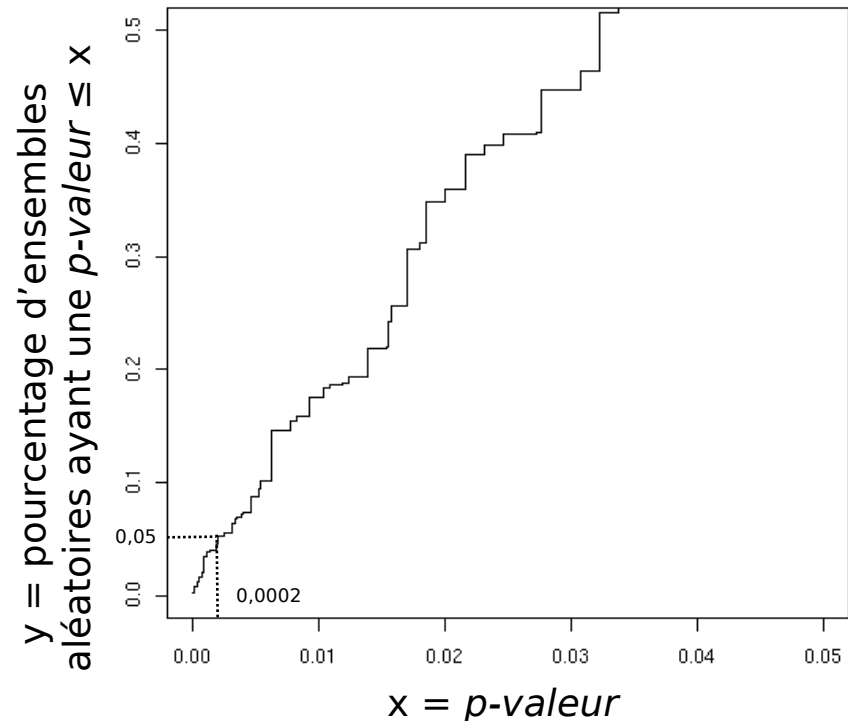
RandomSet_2, minPi = M2

.

.

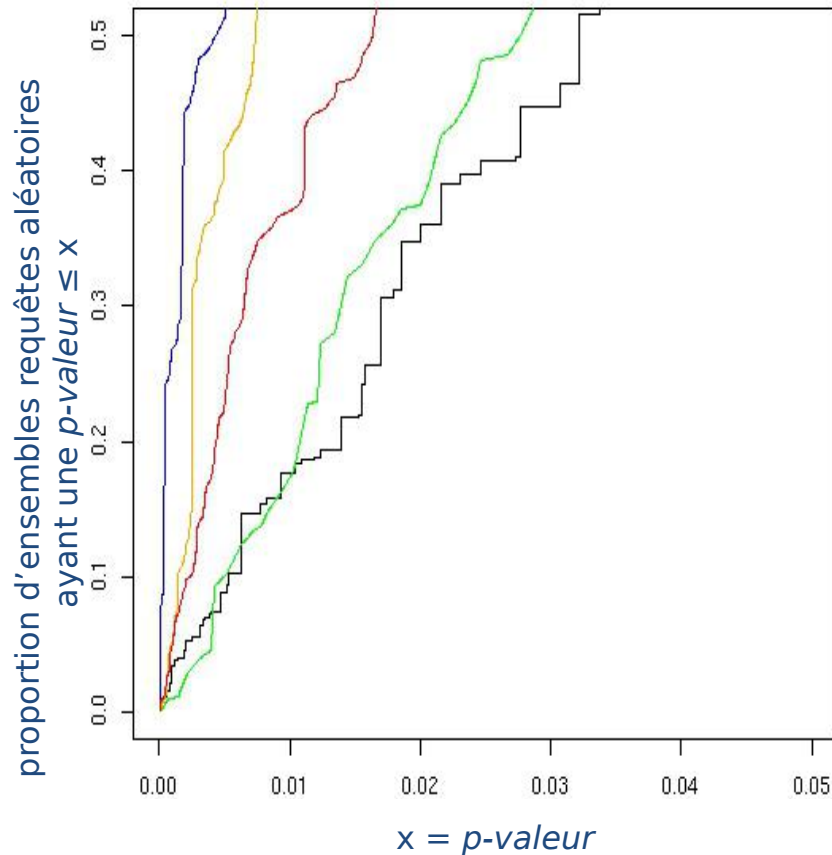
RandomSet_n, minPi = Mn

Étant donnée une p-valeur p
Combien ont un meilleur score ?

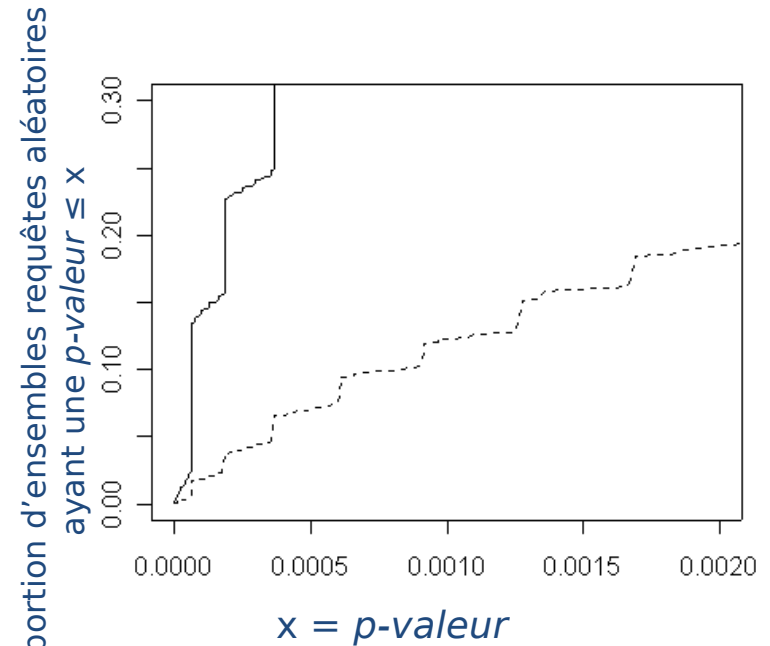


levure *Saccharomyces cerevisiae*
n=500, q=9, g=5786, KEGG Pathways

Significativité des p-valeurs obtenues



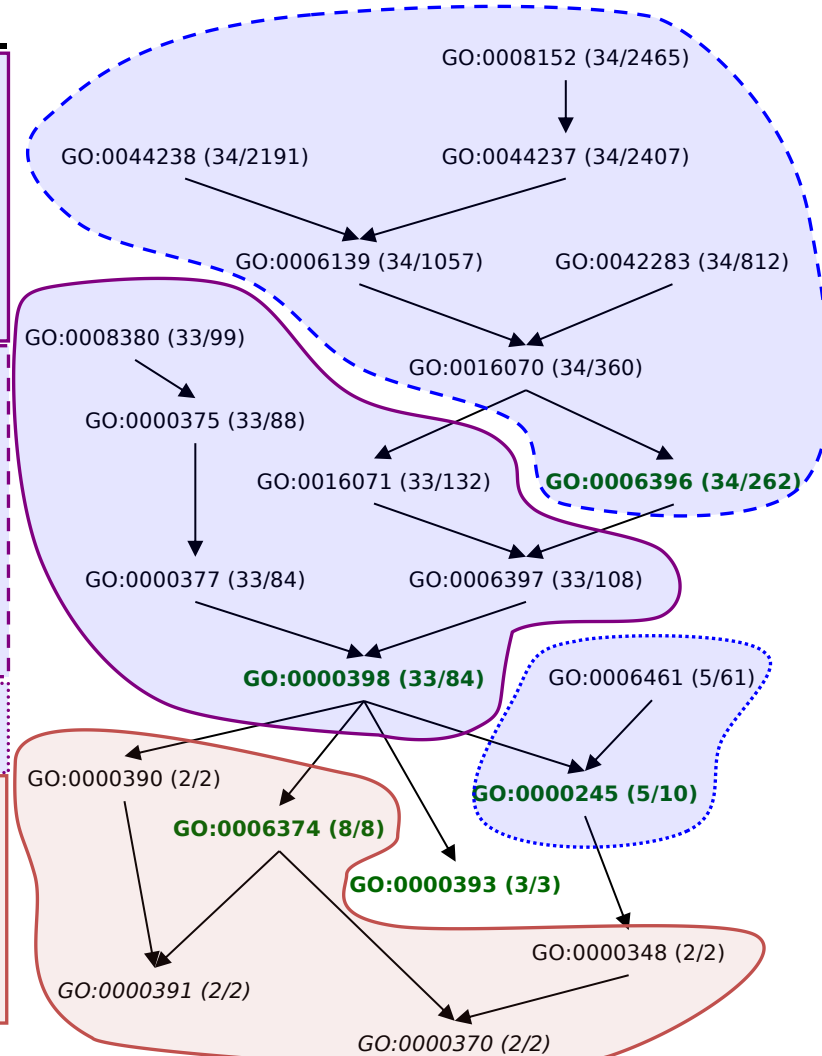
Saccharomyces cerevisiae
 $n=500$, $q=6-9-200-500-1000$,
 $g=5786$, KEGG Pathways



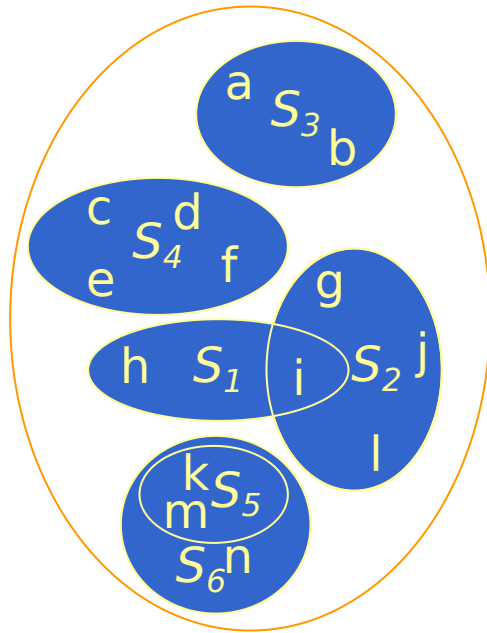
Saccharomyces cerevisiae
 $n=500$, $q=50$, $g=5786$,
 — GO molecular function,
 - - - Ferea *et al.*, 1999

Complex 440.30.10 mRNA splicing

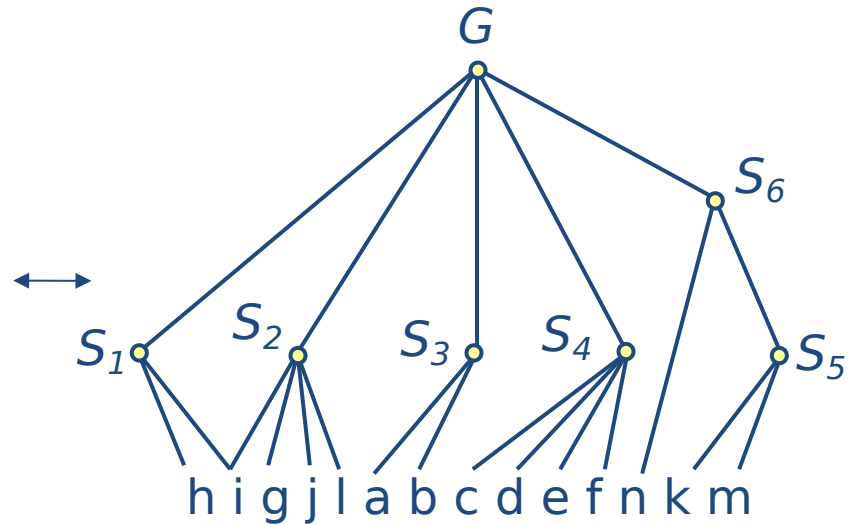
GO Term	Description	Target size	Common elements
GO:0000398	nuclear mRNA splicing, via spliceosome	84	33
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	84	33
GO:0000375	RNA splicing, via transesterification reactions	88	33
GO:0008380	RNA splicing	99	33
GO:0006397	mRNA processing	108	33
GO:0016071	mRNA metabolism	132	33
GO:0006396	RNA processing	262	34
GO:0016070	RNA metabolism	360	34
GO:0043283	biopolymer metabolism	812	34
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1057	34
GO:0044238	primary metabolism	2191	34
GO:0044237	cellular metabolism	2407	34
GO:0008152	metabolism	2465	34
GO:0000245	spliceosome assembly	10	5
GO:0006461	protein complex assembly	61	5
GO:0006374	nuclear mRNA splicing via U2-type spliceosome	8	8
GO:0000391	<i>U2-type spliceosome disassembly</i>	2	2
GO:0000390	spliceosome disassembly	2	2
GO:0000370	<i>U2-type nuclear mRNA branch site recognition</i>	2	2
GO:0000348	nuclear mRNA branch site recognition	2	2
GO:0000393	spliceosomal conformational changes to generate catalytic conformation	3	3



Optimisations



$$N = \{S_1, S_2, S_3, S_4, S_5, S_6\}$$



Hasse diagram of N

a target set T is **pertinent** if

- $Q \cap T \neq \emptyset$
- and
- $\nexists T' \in N$ such that $T' \subset T$ and $T' \cap Q = T \cap Q$
- and
- $\nexists T' \in N$ such that $T \subset T'$ and $T' - Q = T - Q$

Pertinence definition

- Q a non empty query set
- N a neighborhood
- a target set $T \in N$
- T pertinent if

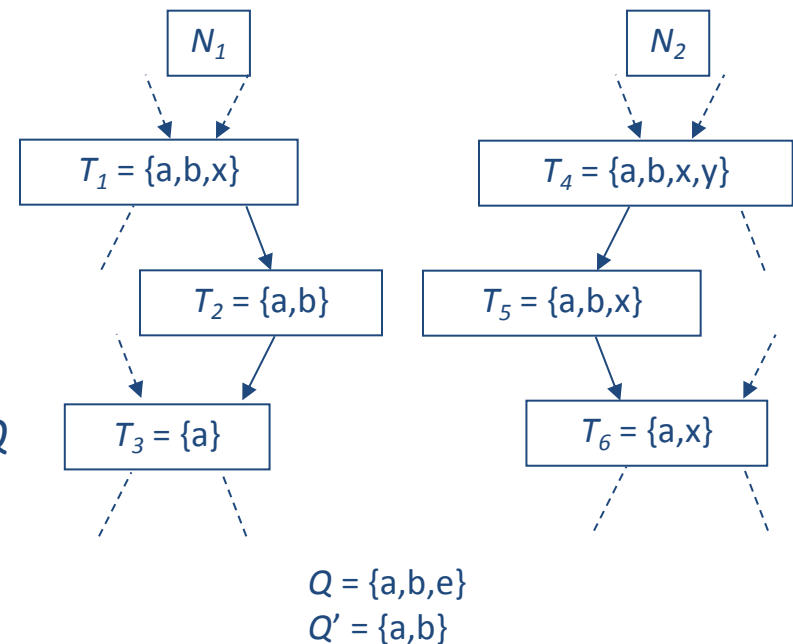
$$Q \cap T \neq \emptyset$$

and

$\nexists T' \in N$ such that $T' \subset T$ and $T' \cap Q = T \cap Q$

and

$\nexists T' \in N$ such that $T \subset T'$ and $T' - Q = T - Q$

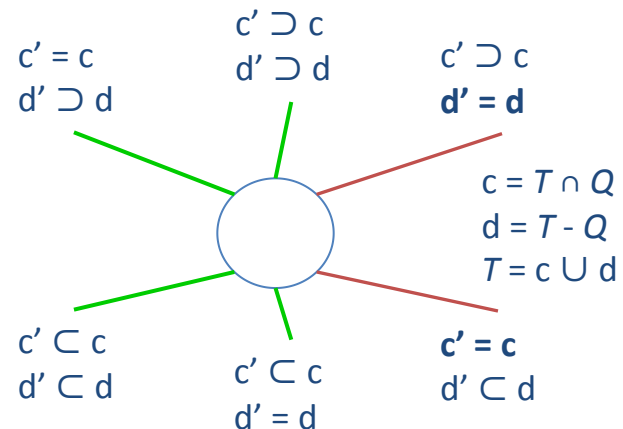


Local decision

$$|c| > 0$$

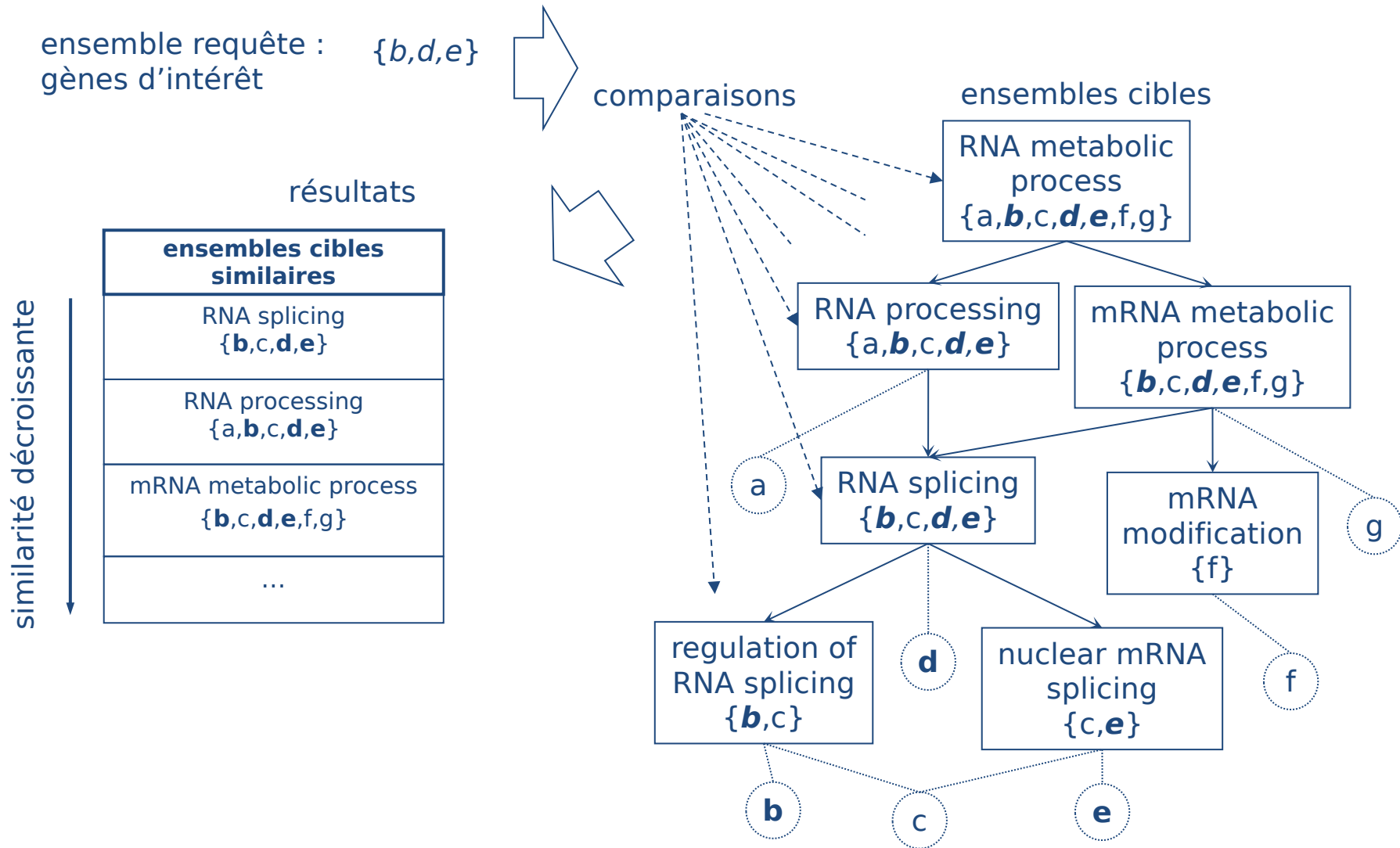
$$|d| < \min(\{d_{\text{parents}}\})$$

$$|c| > \max(\{c_{\text{children}}\})$$

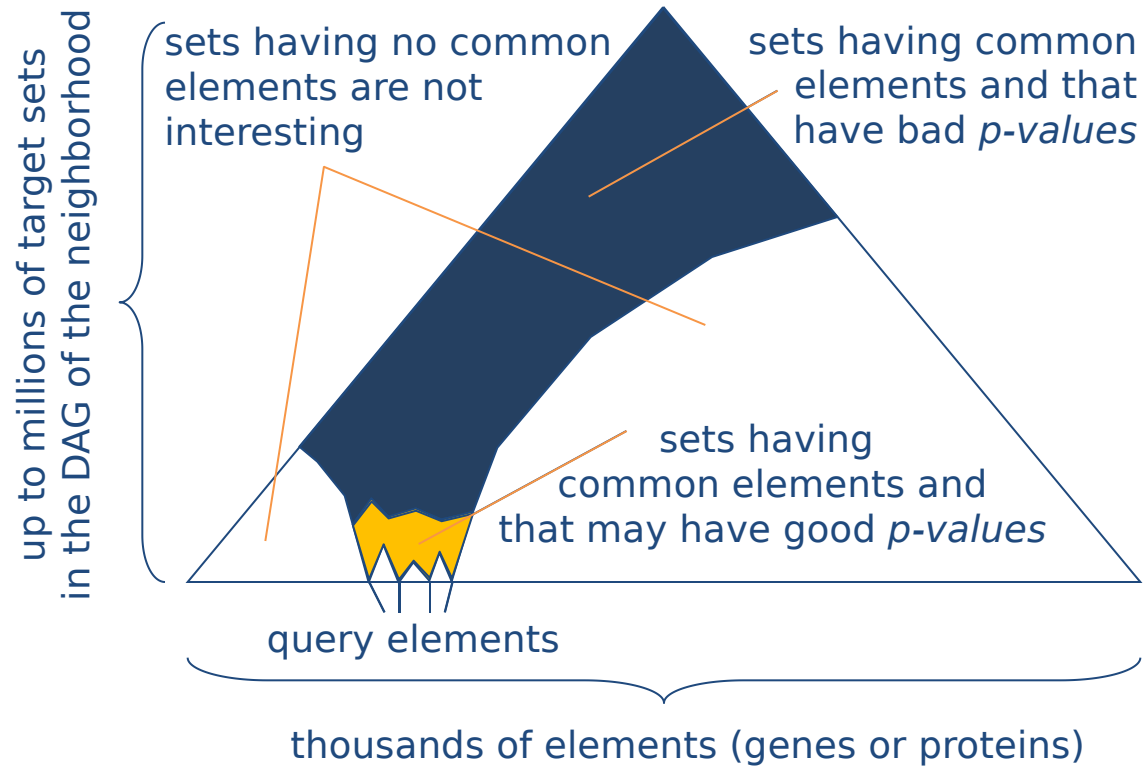


Illustration

- Pertinence des comparaisons & redondance des résultats

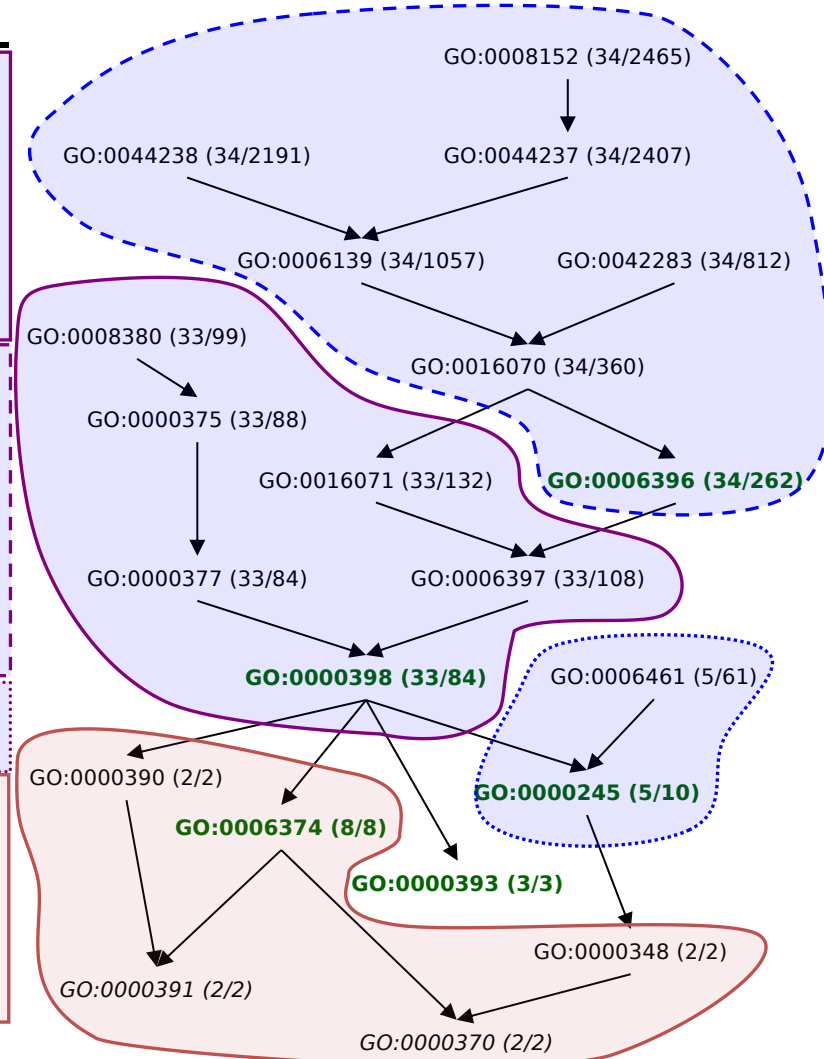


A small portion of the DAG is searched



Complex 440.30.10 mRNA splicing

GO Term	Description	Target size	Common elements
GO:0000398	nuclear mRNA splicing, via spliceosome	84	33
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	84	33
GO:0000375	RNA splicing, via transesterification reactions	88	33
GO:0008380	RNA splicing	99	33
GO:0006397	mRNA processing	108	33
GO:0016071	mRNA metabolism	132	33
GO:0006396	RNA processing	262	34
GO:0016070	RNA metabolism	360	34
GO:0043283	biopolymer metabolism	812	34
GO:0006139	nucleobase, nucleoside, nucleotide and nucleic acid metabolism	1057	34
GO:0044238	primary metabolism	2191	34
GO:0044237	cellular metabolism	2407	34
GO:0008152	metabolism	2465	34
GO:0000245	spliceosome assembly	10	5
GO:0006461	protein complex assembly	61	5
GO:0006374	nuclear mRNA splicing via U2-type spliceosome	8	8
GO:0000391	<i>U2-type spliceosome disassembly</i>	2	2
GO:0000390	spliceosome disassembly	2	2
GO:0000370	<i>U2-type nuclear mRNA branch site recognition</i>	2	2
GO:0000348	nuclear mRNA branch site recognition	2	2
GO:0000393	spliceosomal conformational changes to generate catalytic conformation	3	3



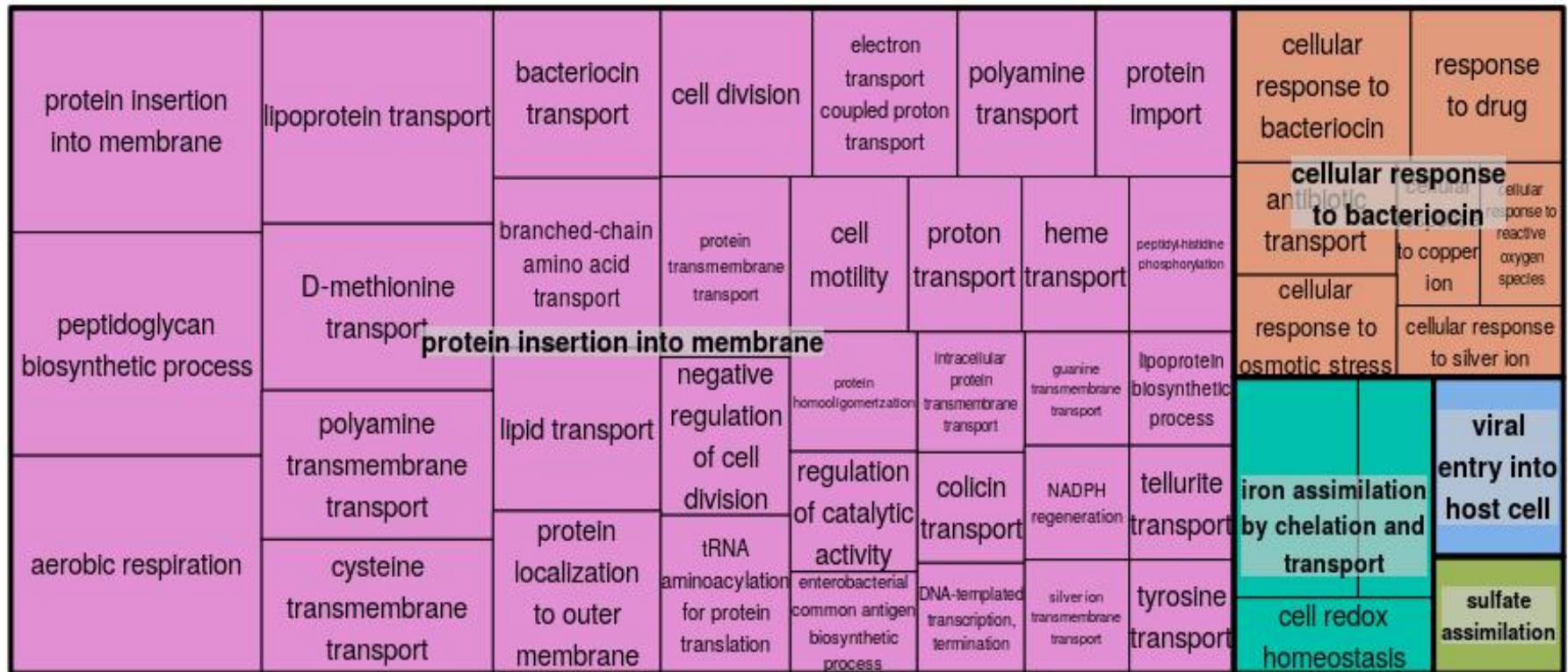
Autre approche pour la visualisation et l'interprétation des résultats

- 134 GO Terms, $p\text{-value} < 0.1$ (no FDR)
- **63 BP**, 19 CC, 52 MF

GO Term	p-value		# common genes	# genes with this annotation in the genome
GO:0009060	2.40863488950467E-05	BP: aerobic respiration	8	22
GO:0003333	0.066449205854714	BP: <u>amino acid transmembrane</u> transport	4	29
GO:0006865	0.052296588628331	BP: <u>amino acid</u> transport	2	7
GO:0042891	0.028532161108969	BP: antibiotic transport	2	5
GO:0015986	0.099101977694913	BP: ATP synthesis coupled proton transport	1	2
GO:0043213	0.003995472333092	BP: bacteriocin transport	3	6
GO:0015803	0.003995472333092	BP: branched-chain <u>amino acid</u> transport	3	6
GO:0051301	0.007038015937826	BP: cell division	6	32
GO:0048870	0.028532161108969	BP: cell motility	2	5
GO:0045454	0.052296588628331	BP: cell redox homeostasis	2	7
GO:0071237	0.005053643900752	BP: cellular response to <u>bacteriocin</u>	2	2
GO:0071280	0.099101977694913	BP: cellular response to copper ion	1	2
GO:0071470	0.039728901461296	BP: cellular response to osmotic stress	2	6
GO:0034614	0.099101977694913	BP: cellular response to reactive oxygen species	1	2
GO:0071292	0.099101977694913	BP: cellular response to silver ion	1	2
GO:0009992	0.039728901461296	BP: cellular water homeostasis	2	6
GO:0042914	0.099101977694913	BP: <u>colicin</u> transport	1	2
GO:1903712	0.002401639427418	BP: <u>cysteine transmembrane</u> transport	3	5
GO:0042883	0.001277327190964	BP: <u>cysteine</u> transport	3	4
GO:0048473	0.000559823205814	BP: <u>D-methionine</u> transport	3	3
GO:0006353	0.099101977694913	BP: <u>DNA-templated</u> transcription, termination	1	2
GO:0015990	0.008780962317314	BP: electron transport coupled proton transport	3	8
GO:0009246	0.080891352640116	BP: <u>enterobacterial common antigen biosynthetic</u> process	2	9
GO:1903716	0.099101977694913	BP: <u>guanine transmembrane</u> transport	1	2
GO:0015886	0.039728901461296	BP: heme transport	2	6
GO:0025244	0.00101077694913	BP: <u>hypoxanthine</u> transport	1	2

Treemap visualisation

- 134 GO Terms, $p\text{-value} < 0.1$ (no FDR)
- **63 BP**, 19 CC, 52 MF

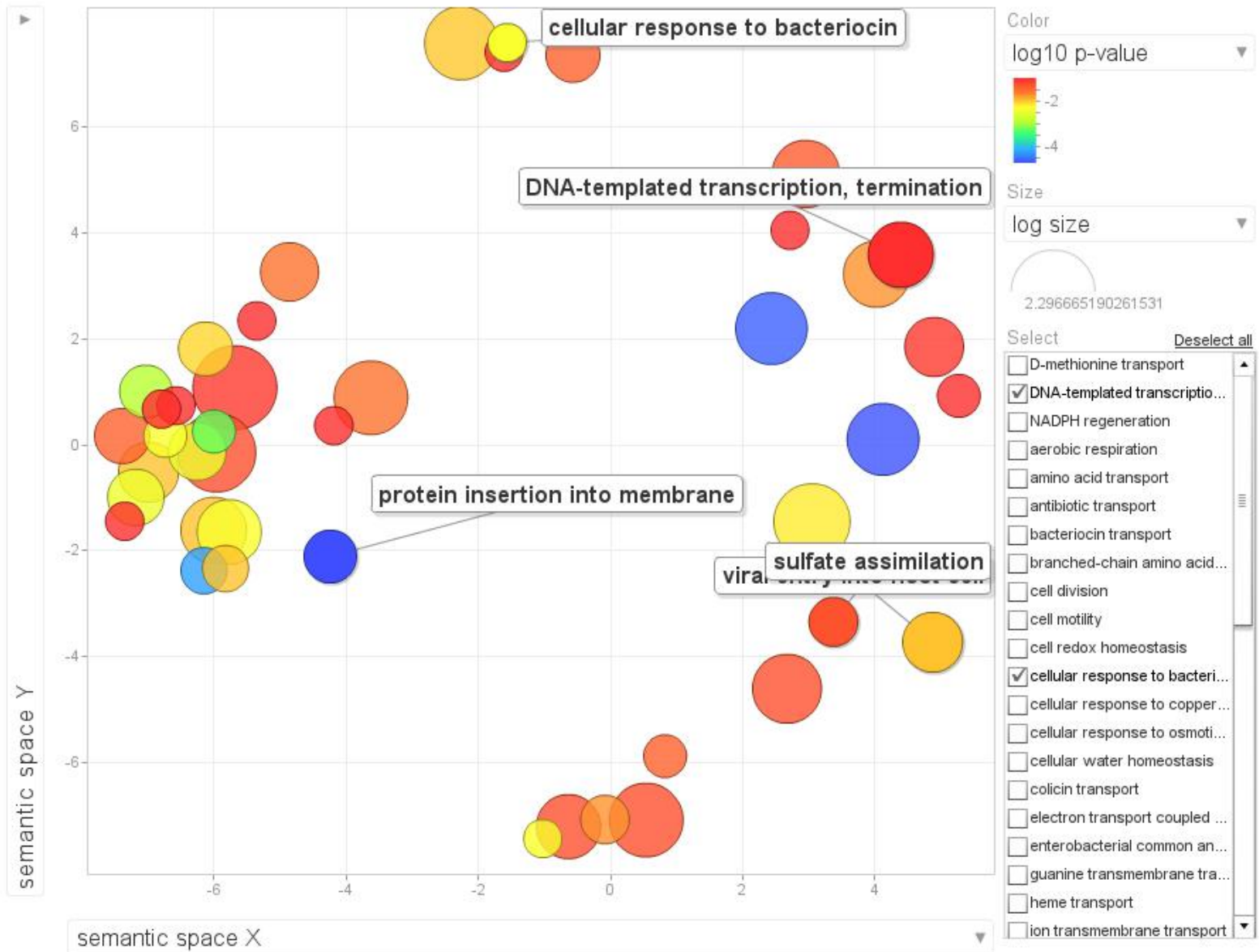


REVIGO scatterplot

Scatterplot & Table

Interactive Graph

TreeMap



click empty space and and drag to zoom

interpretation of the coordinate axes?

- Node/term information content

$$IC(term) = -\log p(term) \quad \text{with } p(term) = \text{freq}(term)$$

- $MICA(t_1, t_2)$: Maximum Information Common Ancestor

$$MICA(t_1, t_2) = \arg \max IC(t_i), t_i \in \text{ancestors}(t_1, t_2)$$

- $sim_{res}(t_1, t_2) = IC(MICA(t_1, t_2))$ [Resnik, 1995]

- $sim_{lin}(t_1, t_2) = IC(MICA(t_1, t_2)) / (IC(t_1) + IC(t_2))$ [Lin, 1998]

- $sim_{gic}(t_1, t_2) = \frac{\sum_{t \in \{GO(t_1) \cap GO(t_2)\}} IC(t)}{\sum_{t \in \{GO(t_1) \cup GO(t_2)\}} IC(t)}$ [Pesquita et al., 2008]

REVIGO interactive graph

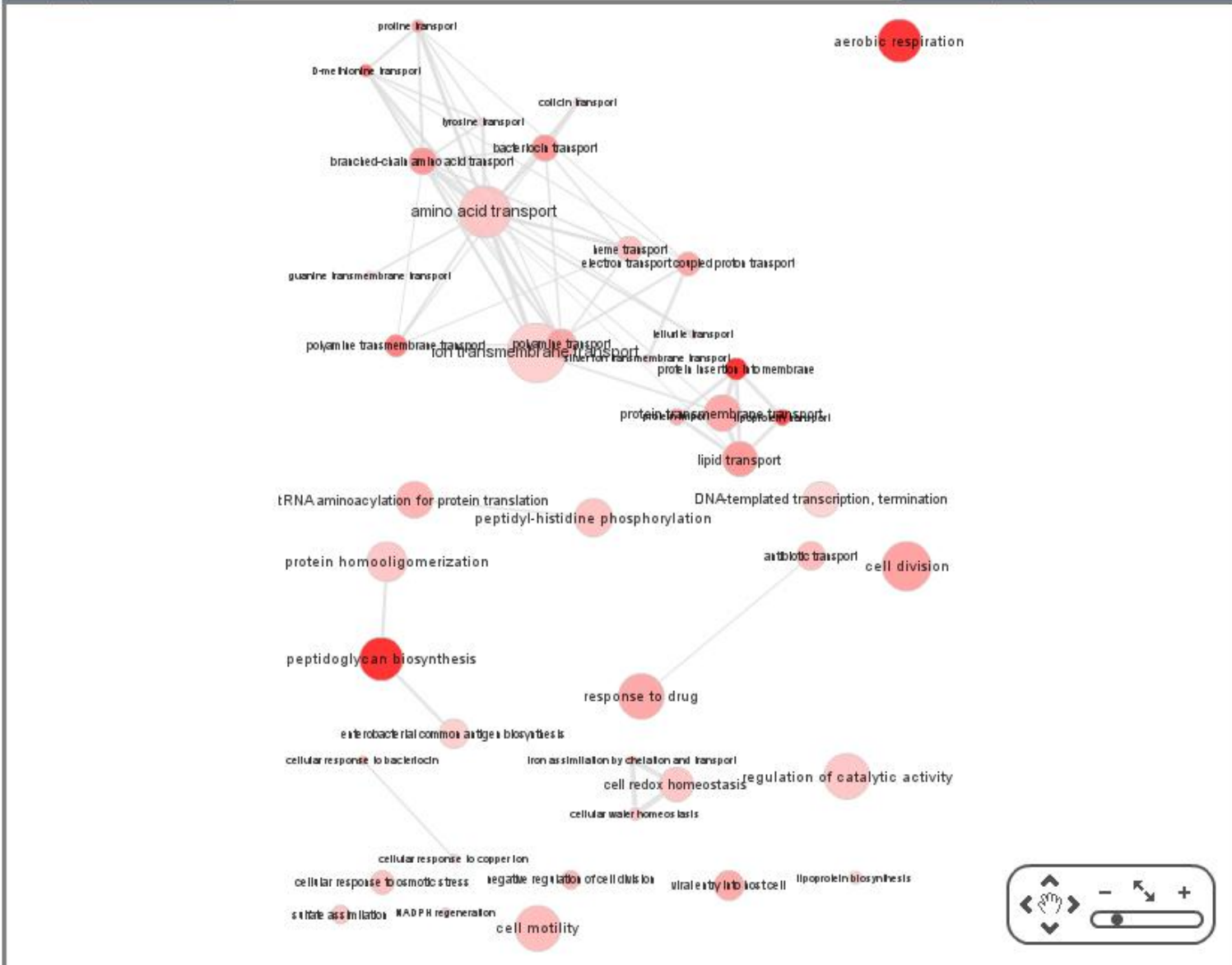
Scatterplot & Table

Interactive Graph

TreeMap

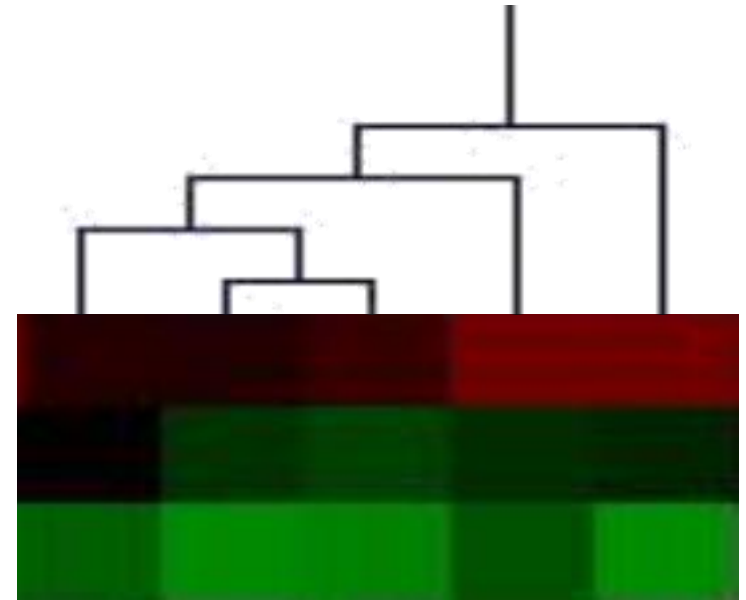
[Run Cytoscape in Java web start](#)

[Download Cytoscape XGMML file for offline use](#)



Further optimizations (1/2)

- each node has only 1 parent
- Algorithm
 - parses the input with a stack of stacks at the time it is loaded
 - $O(|G|)$ time

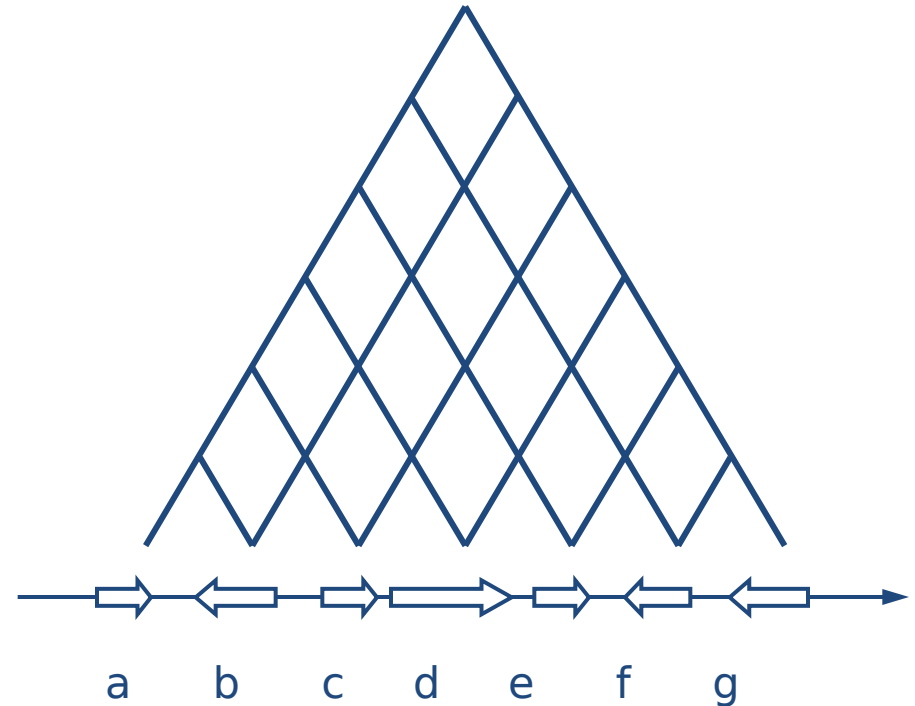


(((a (b c)) d) e)

tree

Further optimizations (2/2)

- DAG is implicit, e.g. adjacent genes on the chromosome:
 - ♦ store the genes order
 - ♦ $\Theta(|G|)$ space instead of $\Theta(|G|^2)$
 - ♦ each pair of genes defines an interval which defines a set
- requires a specific algorithm
 - ♦ $O(|Q|^2)$ time

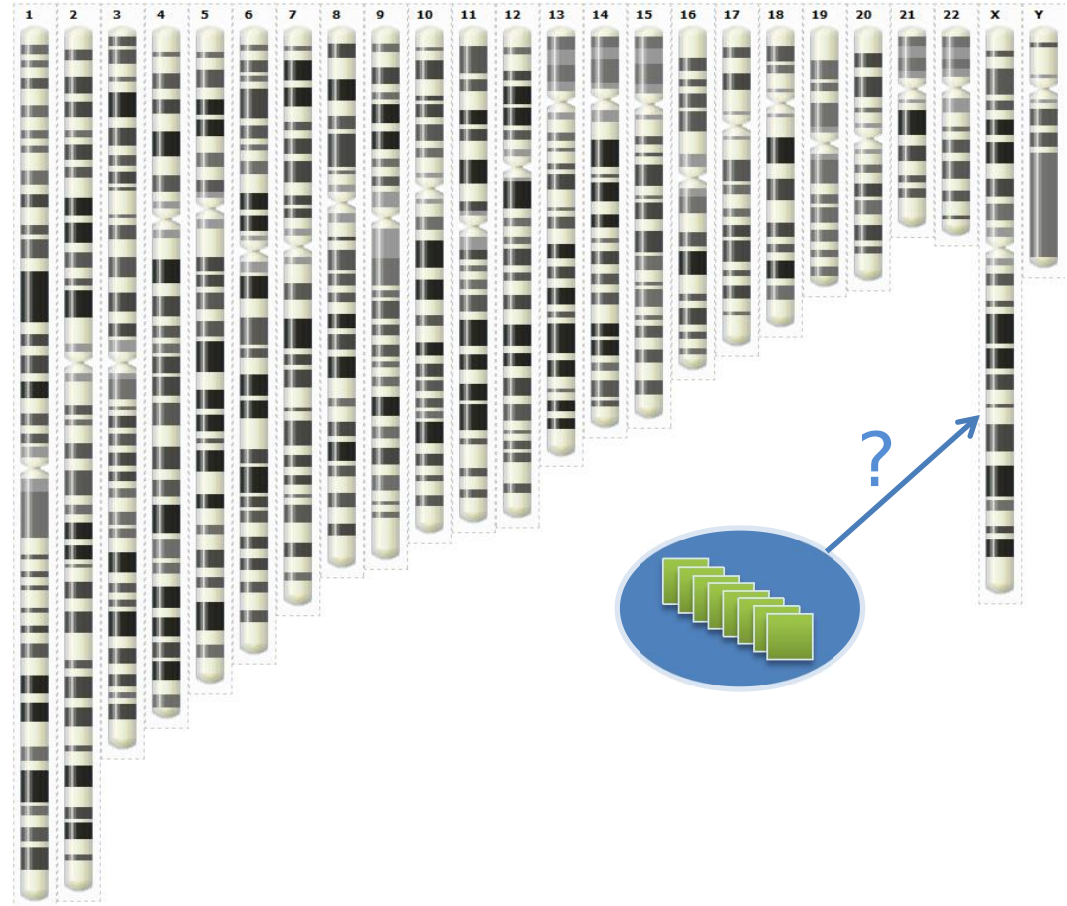


implicit

Set of genes of interest

Examples

- ◆ Differentially expressed genes
- ◆ Co-expressed genes
- ◆ Tissue specific genes
- ◆ Partners of a protein complex
- ◆ Imprinted genes
- ◆ ...



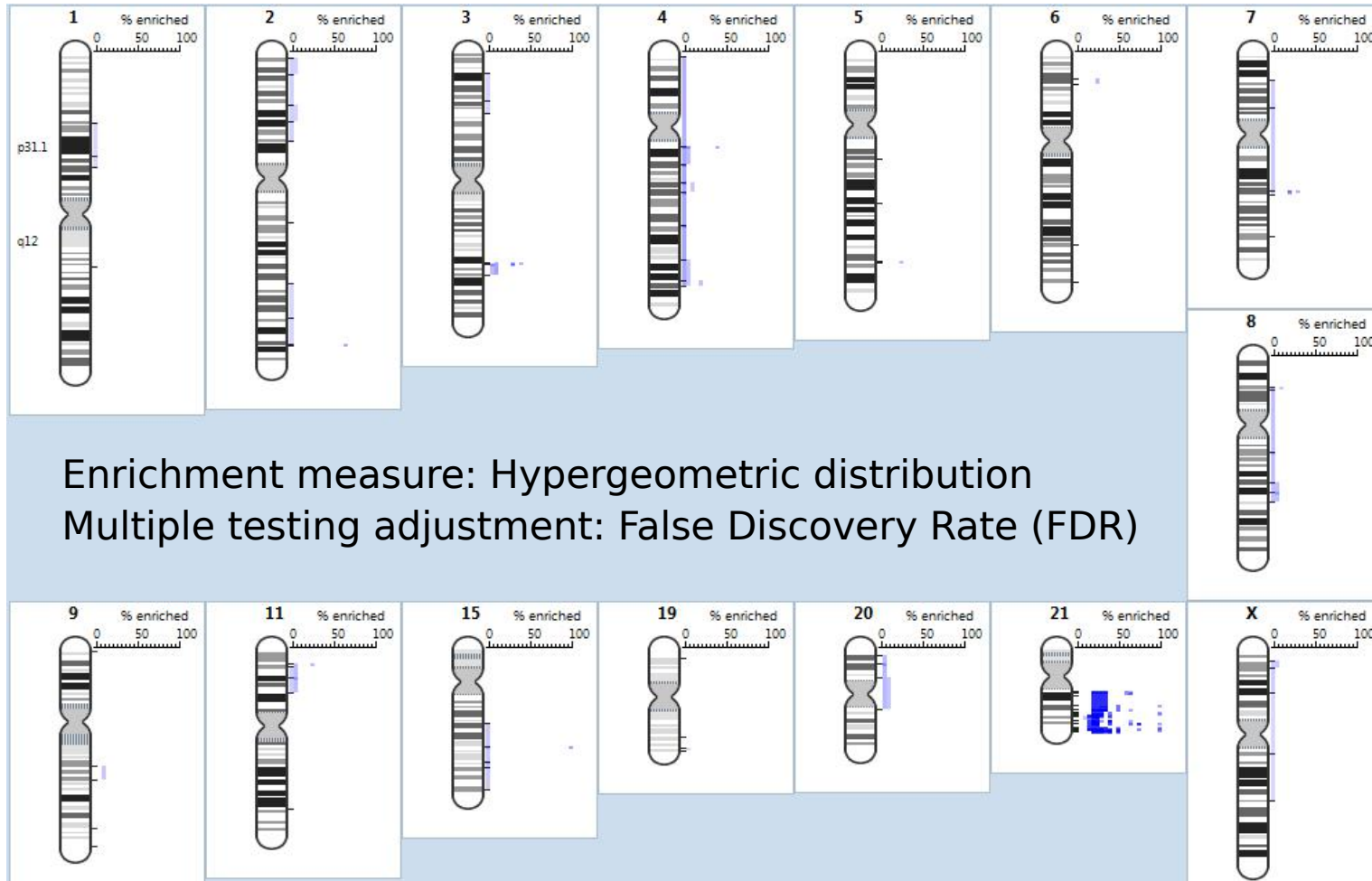
→ Question: Do those genes surprisingly cluster in the genome?

Goal: consider every possible region for enrichment

Down Syndrome differentially expressed genes

Experiment:

Published list of **differentially expressed genes** in **Down syndrome patients** from Mao, R., C.L. Zielke, H.R. Zielke, and J. Pevsner, Global up-regulation of chromosome 21 gene expression in the developing Down syndrome brain (2003) *Genomics* **81**: 457-467.



Issues:

- Number of regions to test
- False positives
- Redundancy

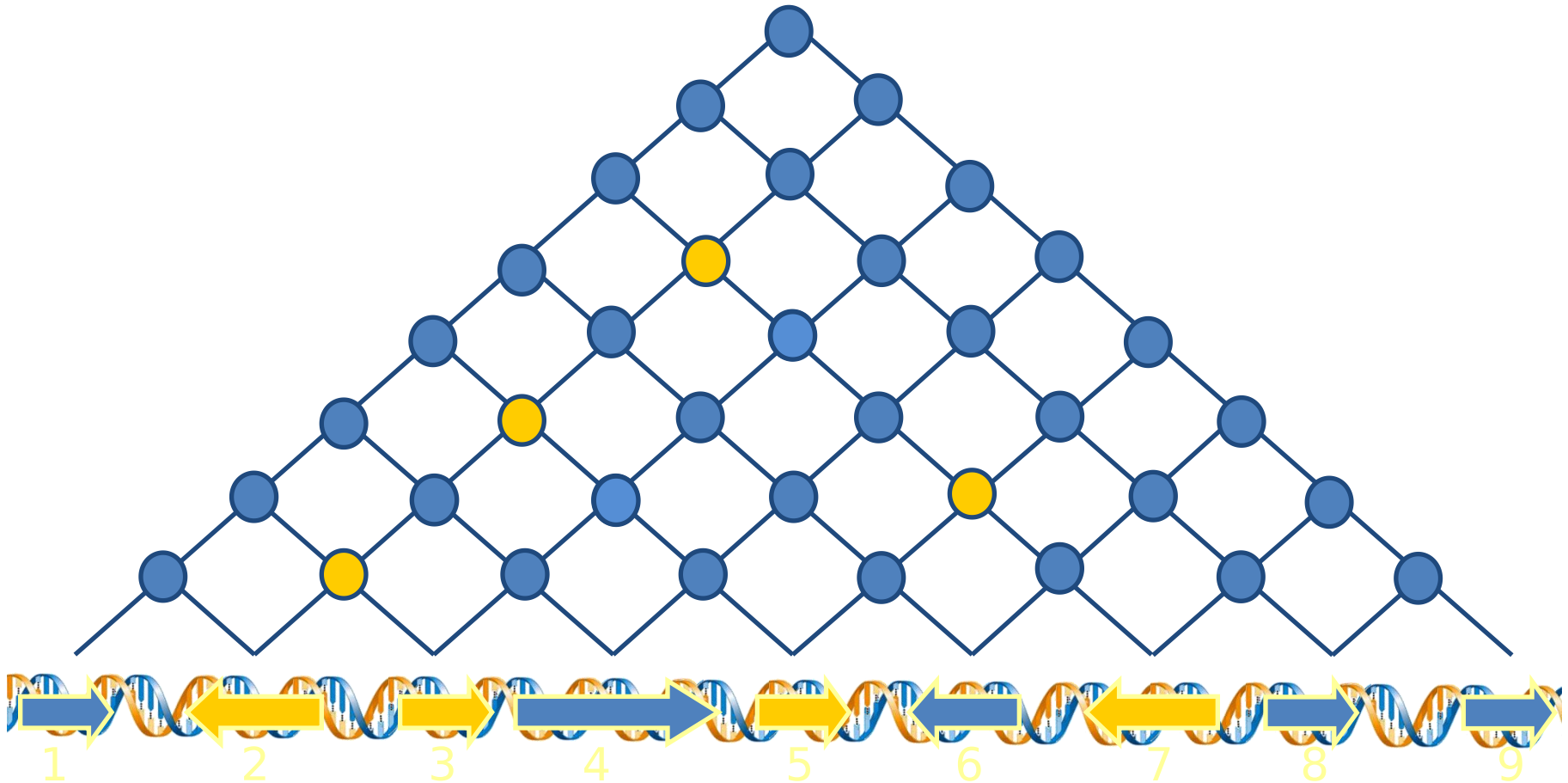
Enrichment measure: Hypergeometric distribution

Multiple testing adjustment: False Discovery Rate (FDR)

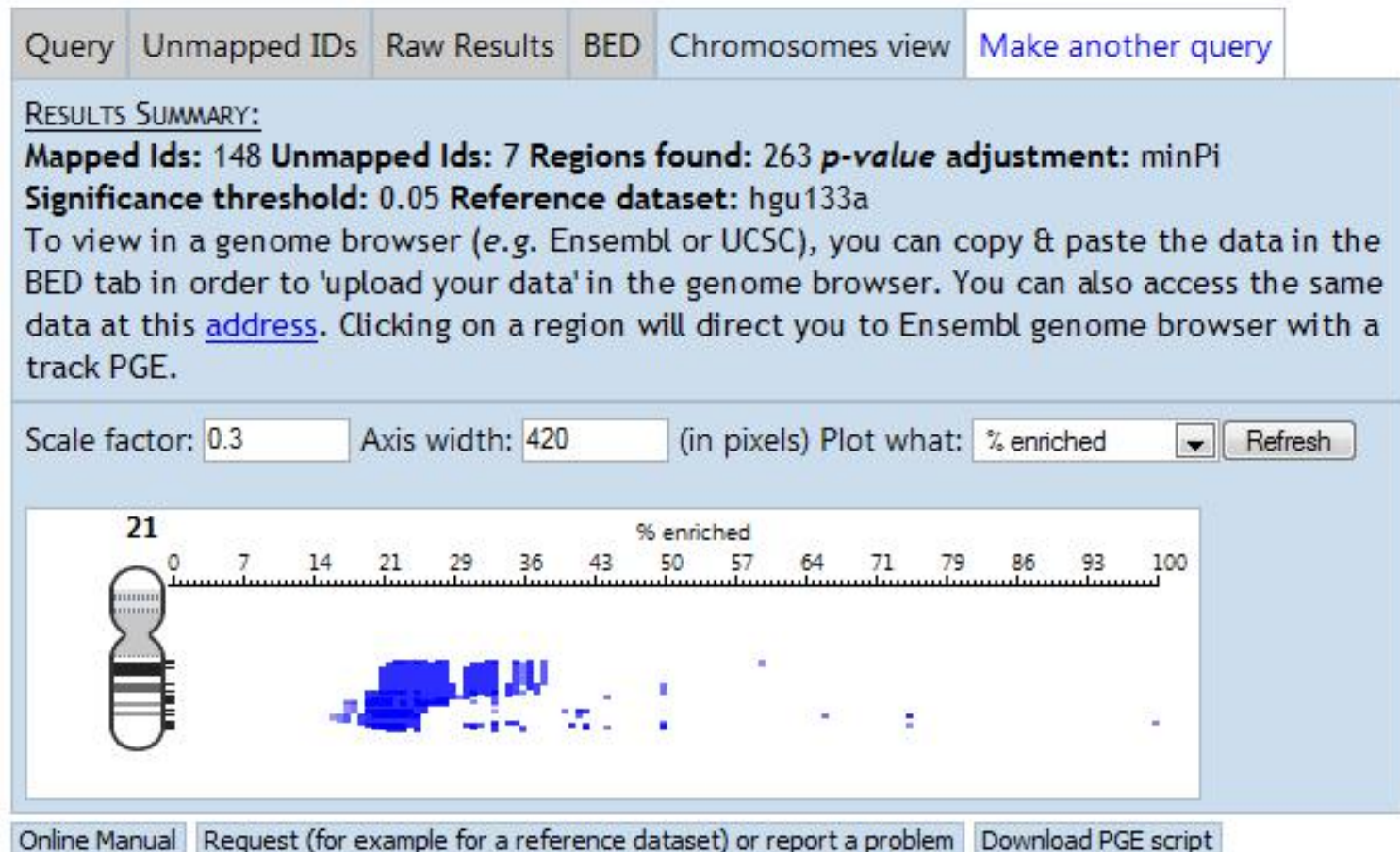
Pertinent Regions

A region is pertinent if it is:

- bounded by genes of interests
- the largest, when genes of interest are consecutive



Down Syndrome ($\min P_i$)



Large regions tend to have smaller *p-values* while small regions tend to have higher percentage of enrichment

→ A smaller region included in a more significant one is pertinent if it has a much higher percentage of genes of interests (>50%)

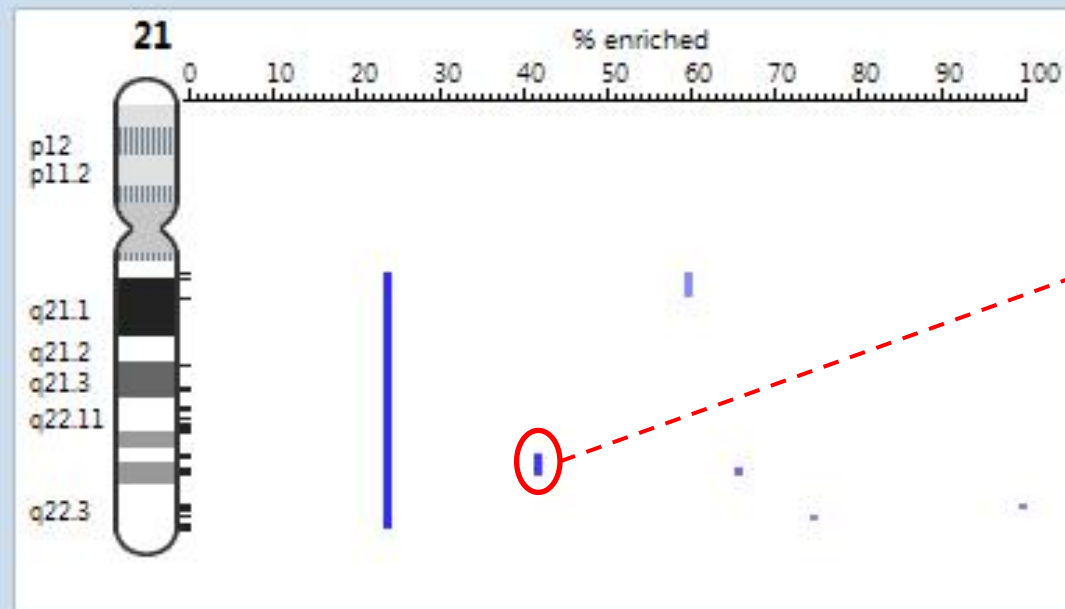
Down Syndrome d.e.g. Final Results

Query Unmapped IDs Raw Results BED Chromosomes view [Make another query](#)

RESULTS SUMMARY:

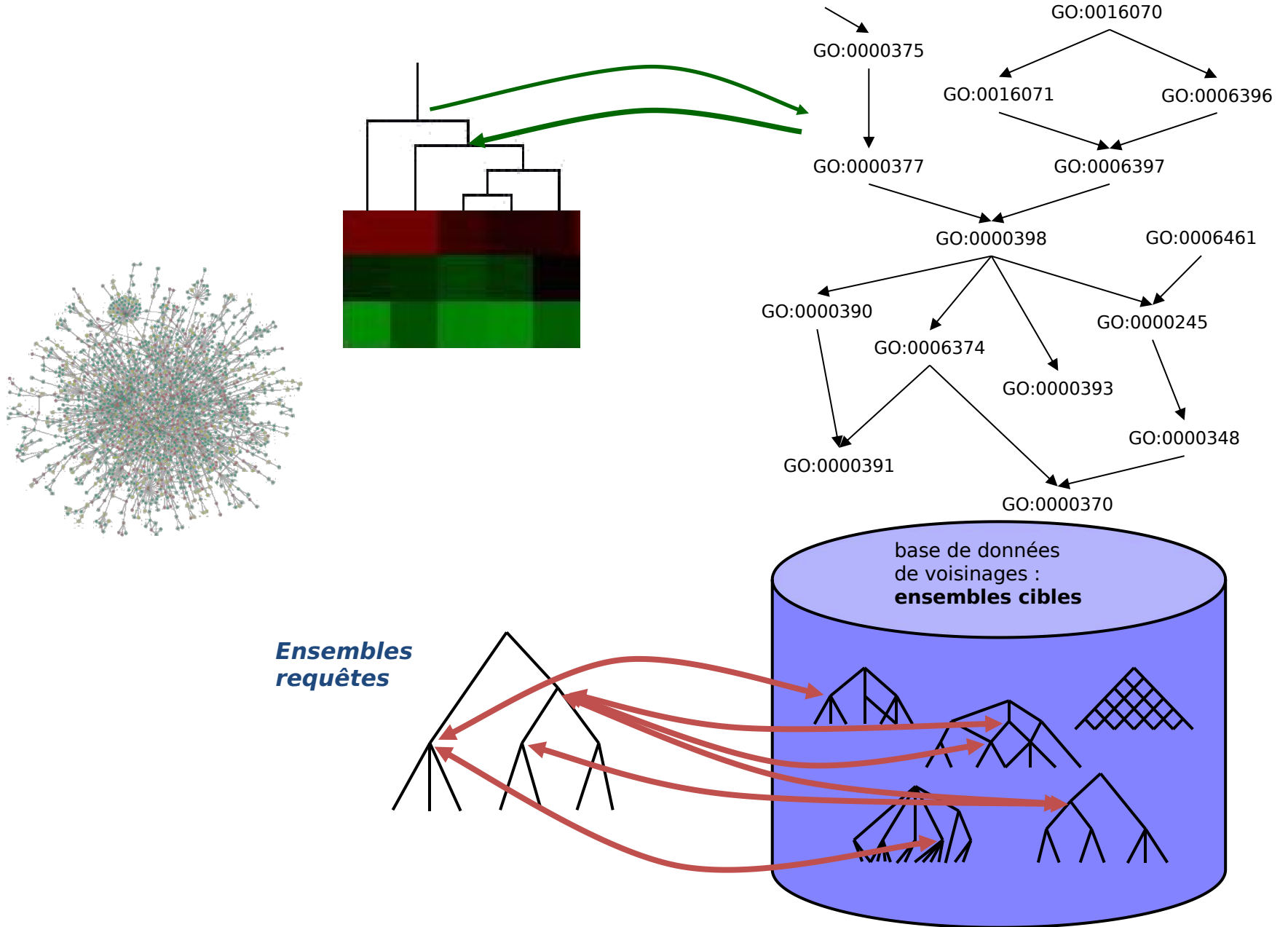
**Mapped Ids: 148 Unmapped Ids: 7 Regions found: 6 *p*-value adjustment: minPi
Significance threshold: 0.05 Reference dataset: hgu133a**

Scale factor: Axis width: (in pixels) Plot what:

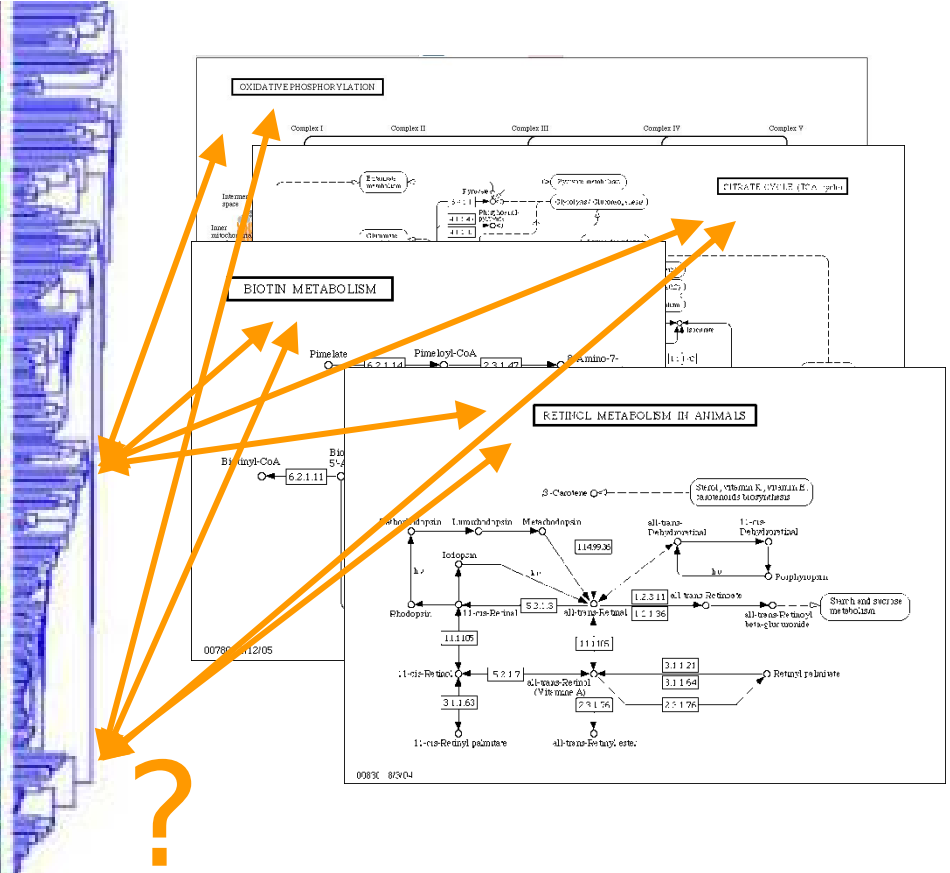
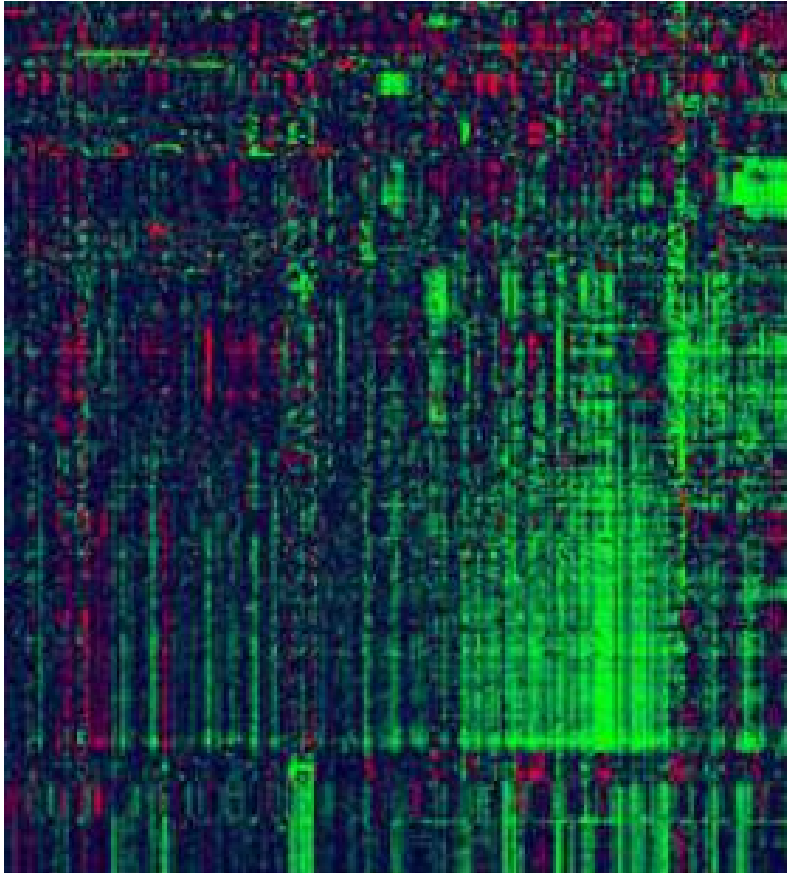


p-value: 7.87E-10
p-value_{adj}: <0.002
 Score: 91.039
 Score_{adj}: INF
 Common elements: 6 / 14 (43%)
 Overlapping regions: 1
 Start of region: 37 359 546 bp
 End of region: 40 223 183 bp
 Genes:
 DSCR2, DYRK1A, PCP4, PIGP, SH3BGR
 WRB

Défis actuels



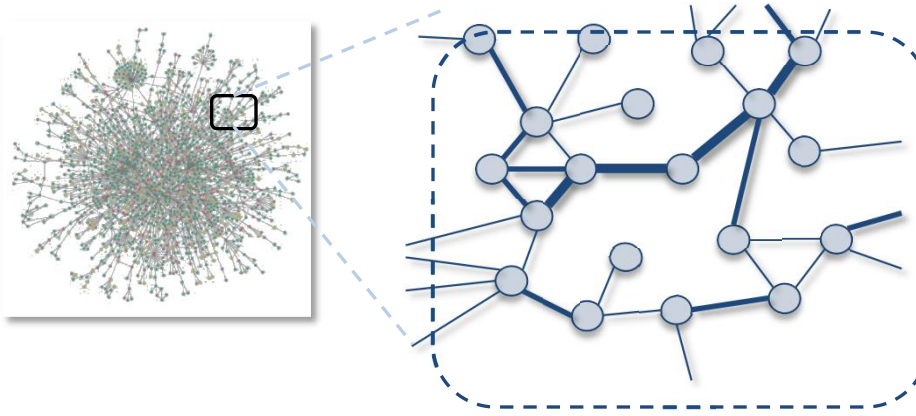
Analyse de données d'expression



[Ferea *et al.*, 1999]

[Kanehisa & Goto, 2000]

Extraction de sous-graphe pertinent & visualisation



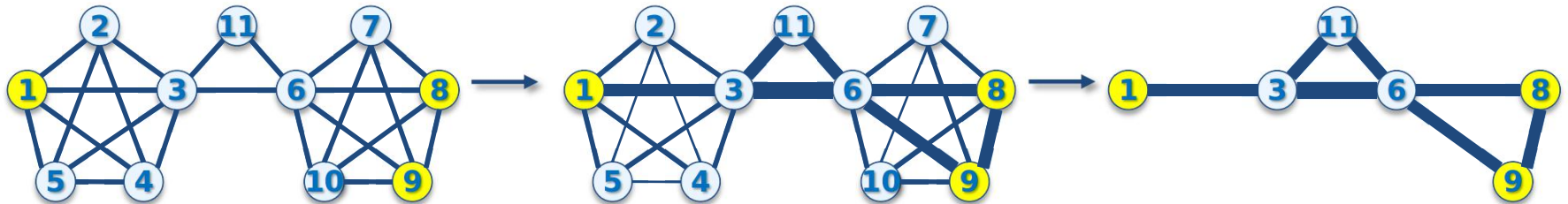
Idée :

- Grands graphes d'interactions physiques et/ou fonctionnelles
- Visualiser les relations entre gènes d'intérêt

Gènes ayant la même annotation
ex : interaction with host

Marche aléatoire :
pondération des arcs

Surreprésentation :
sous-graphe pertinent



Visualisation du sous graphe expliquant le mieux ce qui lie les gènes d'intérêt