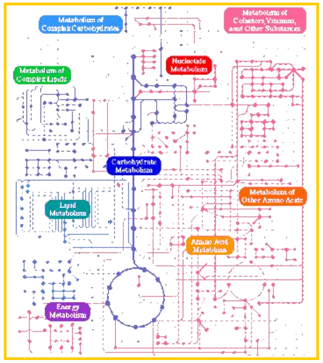
Intégration de données hétérogènes
Enrichissement

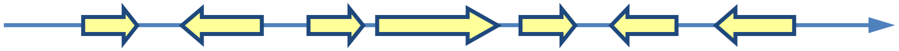
Master 2

Bioinformatique et Biologie des Systèmes

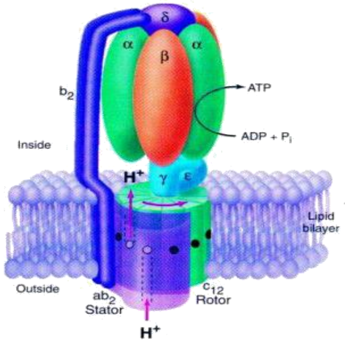
Recoupement de voisinages : approche ensembliste


Metabolism of Complex Carbohydrates
Metabolism of Cofactors, Vitamins, and Other Substances
Nucleotide Metabolism
Metabolism of Complex Lipids
Carbohydrate Metabolism
Metabolism of Other Amino Acids
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism

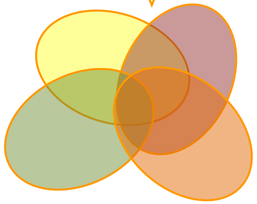
voies métaboliques



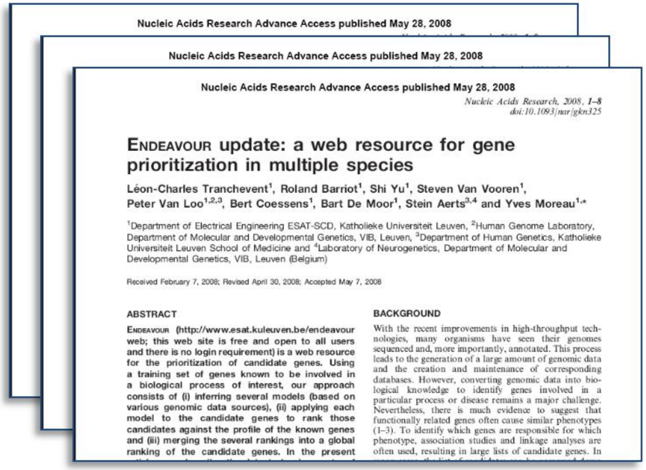
localisation chromosomique



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/gkn325

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

Léon-Charles Tranchevent[1], Roland Barriot[1], Shi Yu[1], Steven Van Vooren[1], Peter Van Loo[1,2,3], Bert Coessens[1], Bart De Moor[1], Stein Aerts[3,4] and Yves Moreau[1,*]

[1]Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, [2]Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, [3]Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and [4]Laboratory 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, converting 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

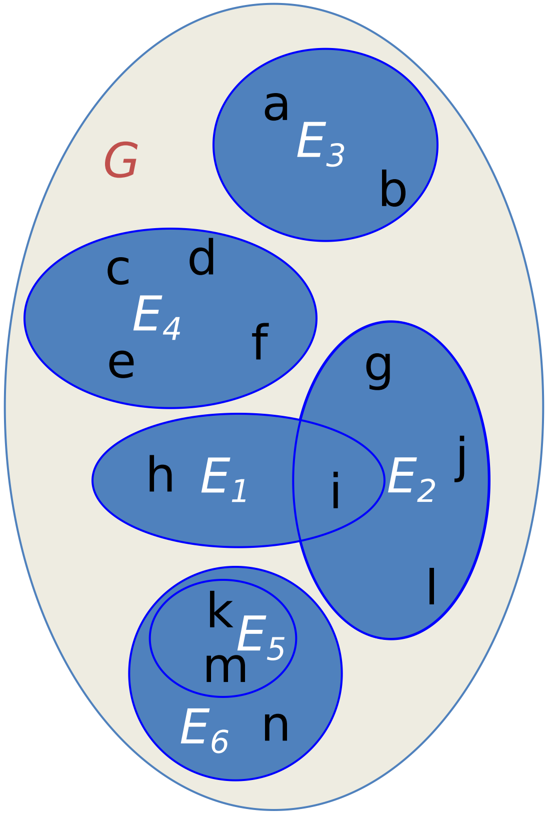
co-citation

domaines protéiques

Gene Ontology

Définitions

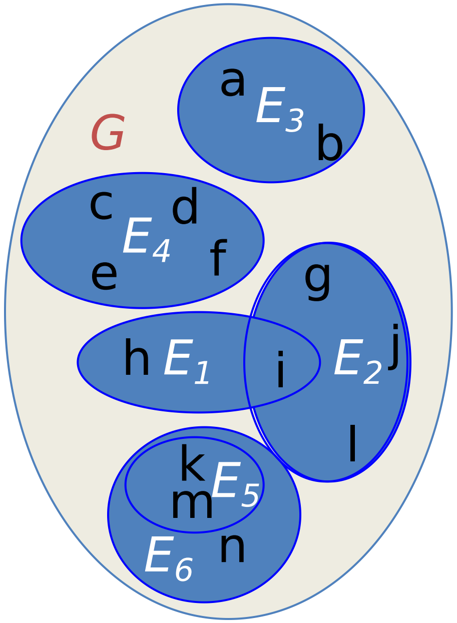
- (Identifiants de) gène → ARNm → protéine
- *G* : ensemble des gènes d'un organisme
- *Fonction de regroupement* : relation entre gènes basée sur un indice de similarité.
- *Ensemble de (gènes) voisins* : ensemble de gènes $E \subseteq G$ regroupés par une fonction de regroupement.
- *Voisinage* : sous-ensemble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



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

Représentation d’un voisinage : ordre partiel (poset)

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



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

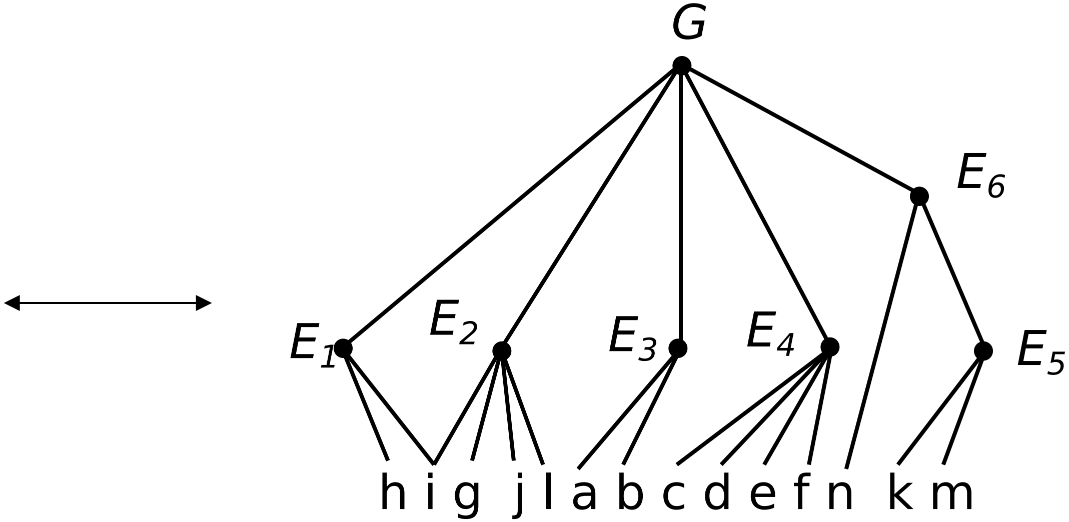
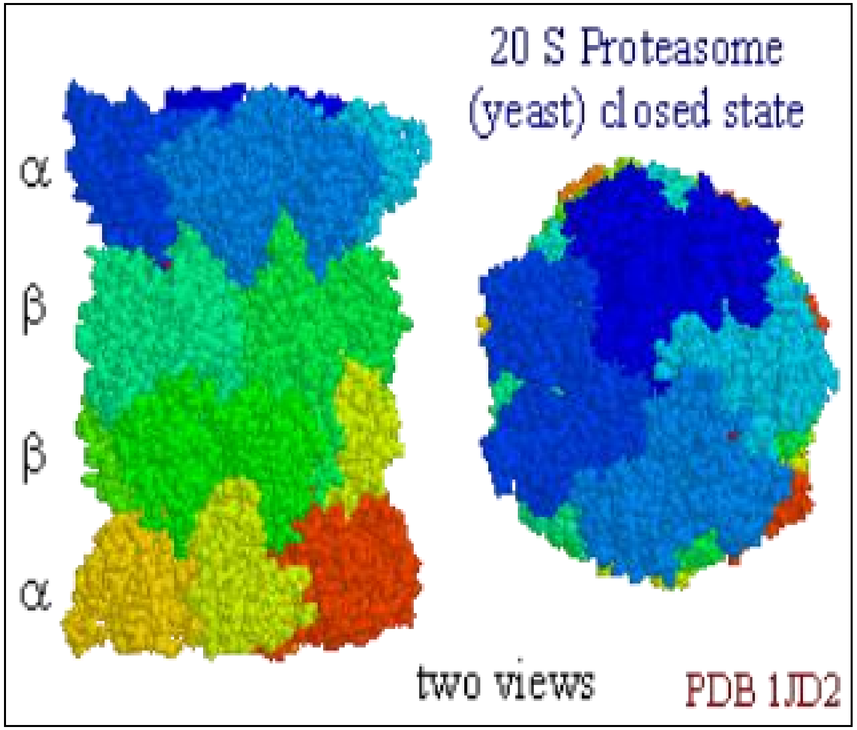
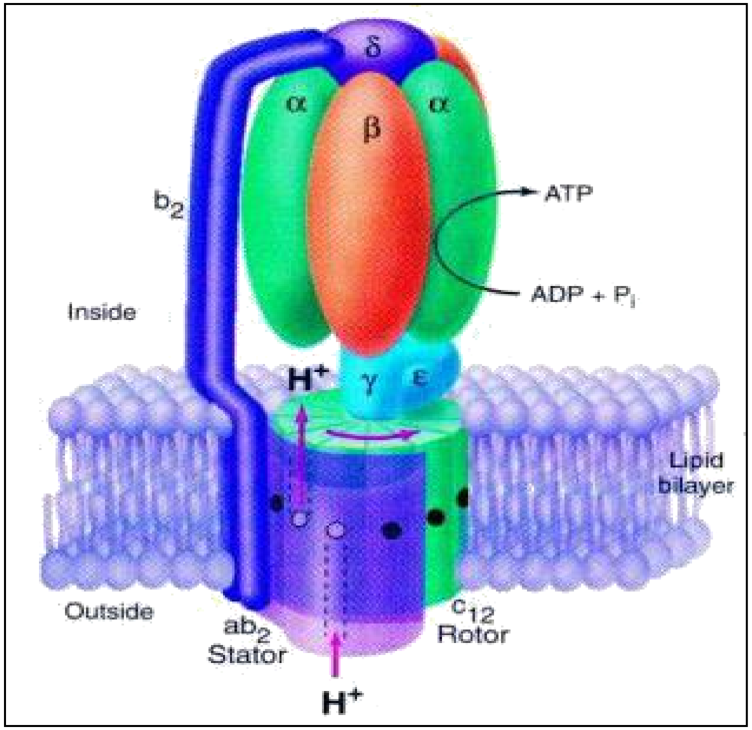


diagramme de Hasse de V

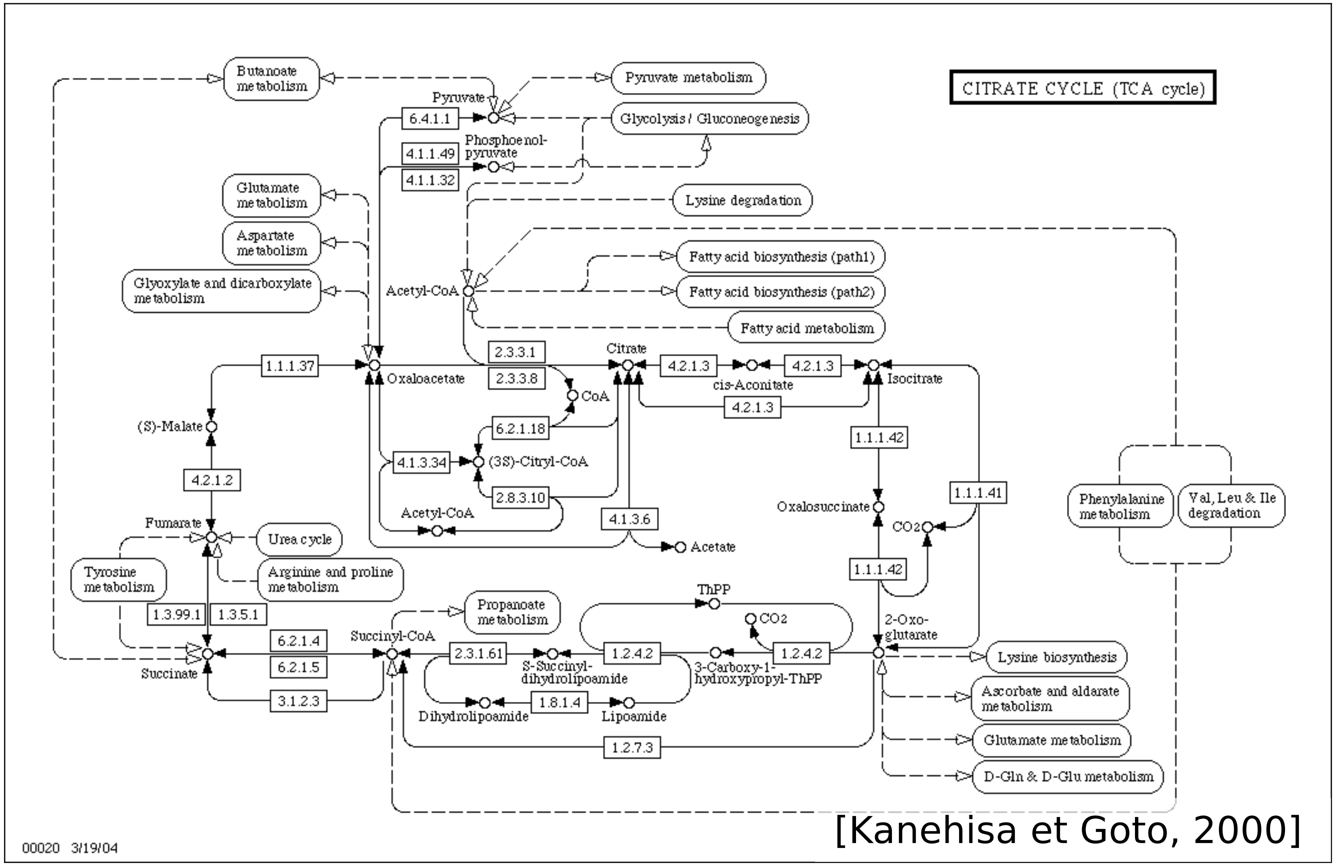
Exemple de fonction de regroupement : complexes protéiques





un complexe → un ensemble de protéines

Exemple de critère de regroupement : voies métaboliques

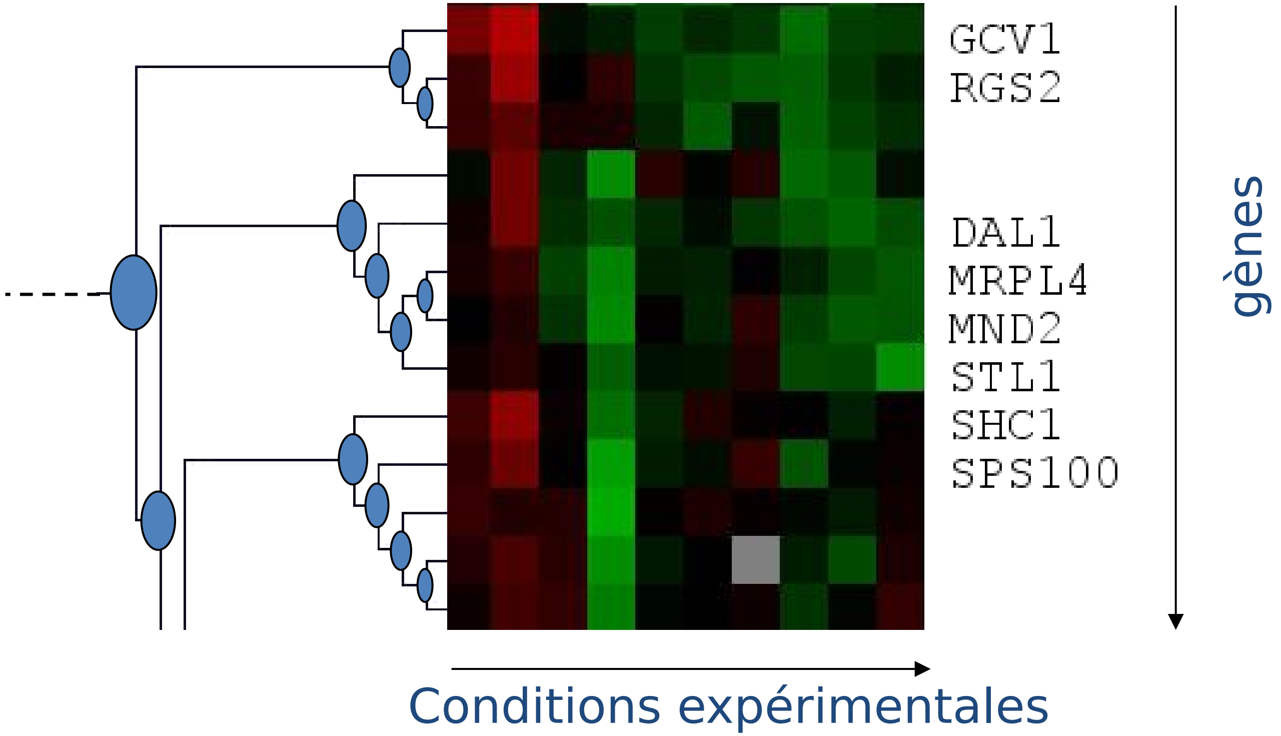


[Kanehisa et Goto, 2000]

une voie métabolique → un ensemble de protéines

Exemple de critère de regroupement : données d’expression

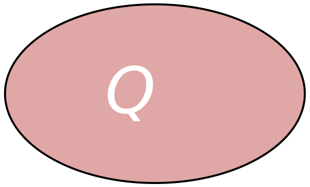
clustering hiérarchique
des profils



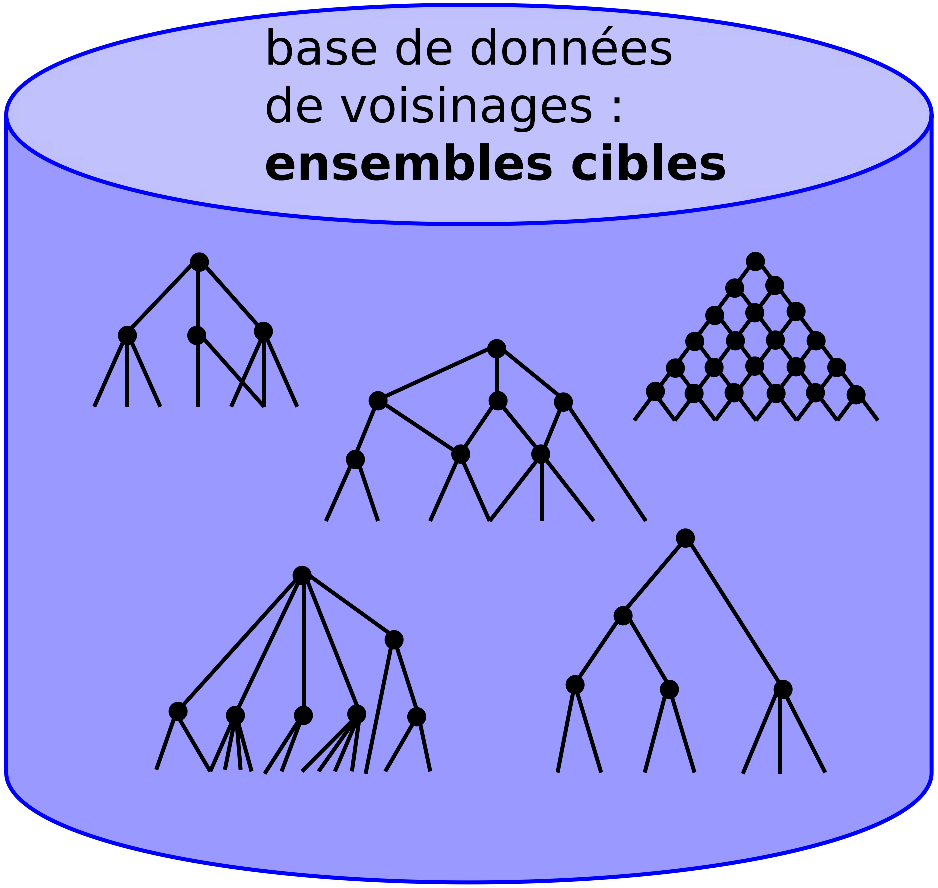
un cluster → un ensemble de gènes

- Recherche d’ensembles similaires

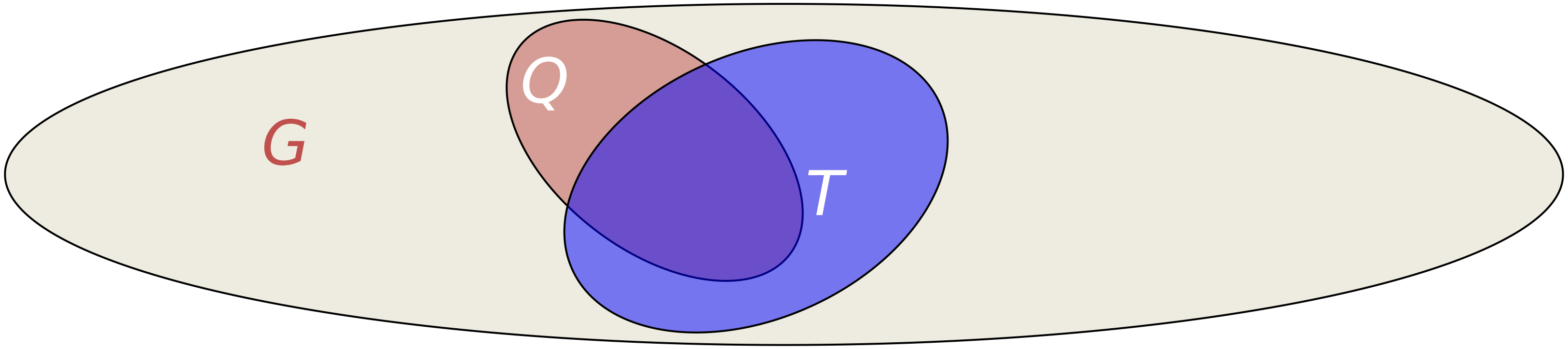
ensemble requête
$Q \subseteq G$



Quels sont les ensembles cibles qui lui sont similaires ?


base de données
de voisinages :
ensembles cibles

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

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

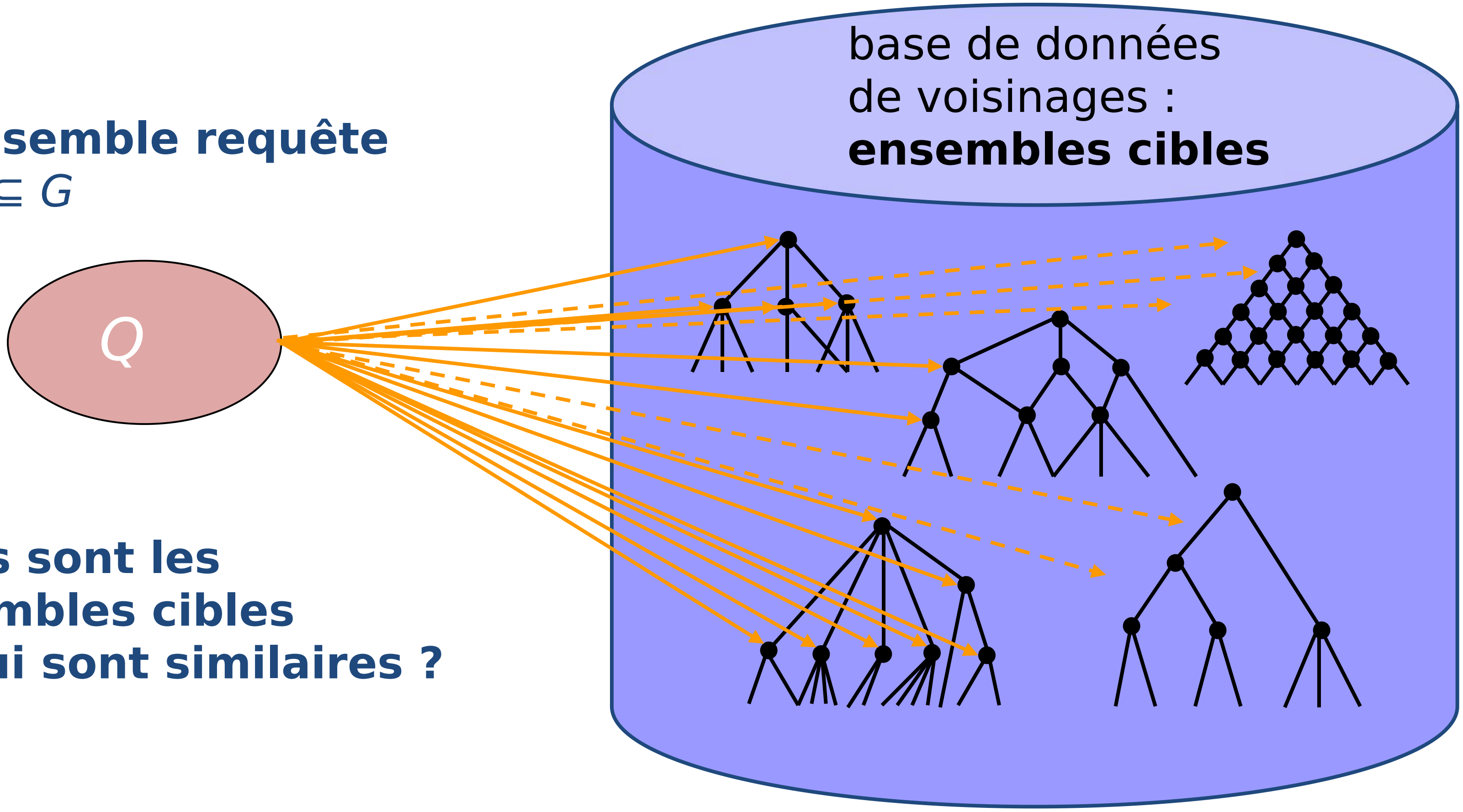
avec

- *g* = |*G*| : taille de la population
- *q* = |*Q*| : taille de l’ensemble requête
- *t* = |*T*| : taille de l’ensemble cible
- *c* = |*Q* ∩ *T*| : nombre d’éléments communs

- Autres mesures :
  - Loi binomiale
  - χ^2
  - ratio, pourcentage

- Recherche d’ensembles similaires

ensemble requête
Q ⊆ *G*



Q

base de données
de voisinages :
ensembles cibles

Quels sont les
ensembles cibles
qui lui sont similaires ?

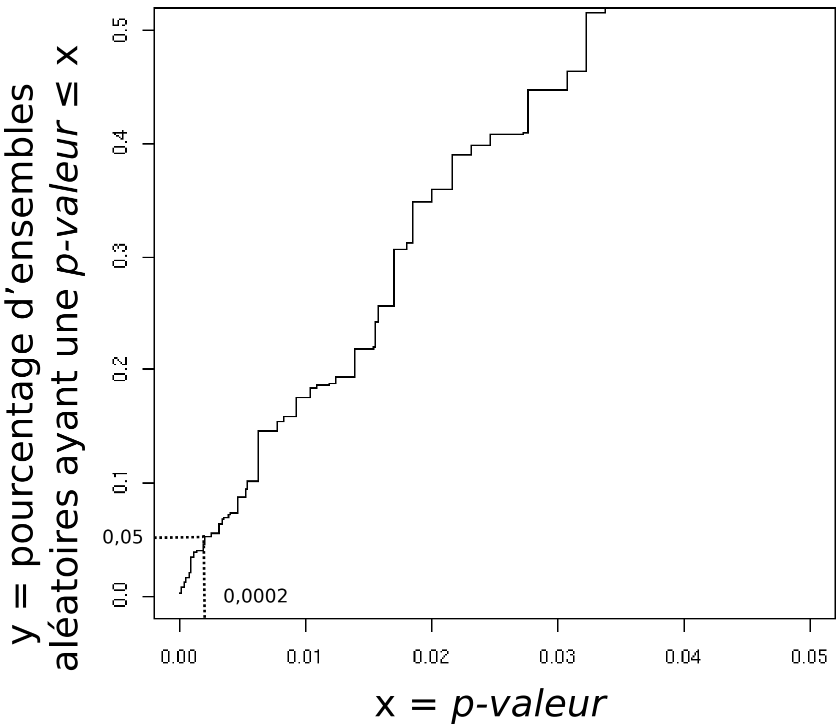
Tests multiples et significativité des p-valeurs

- 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 ?

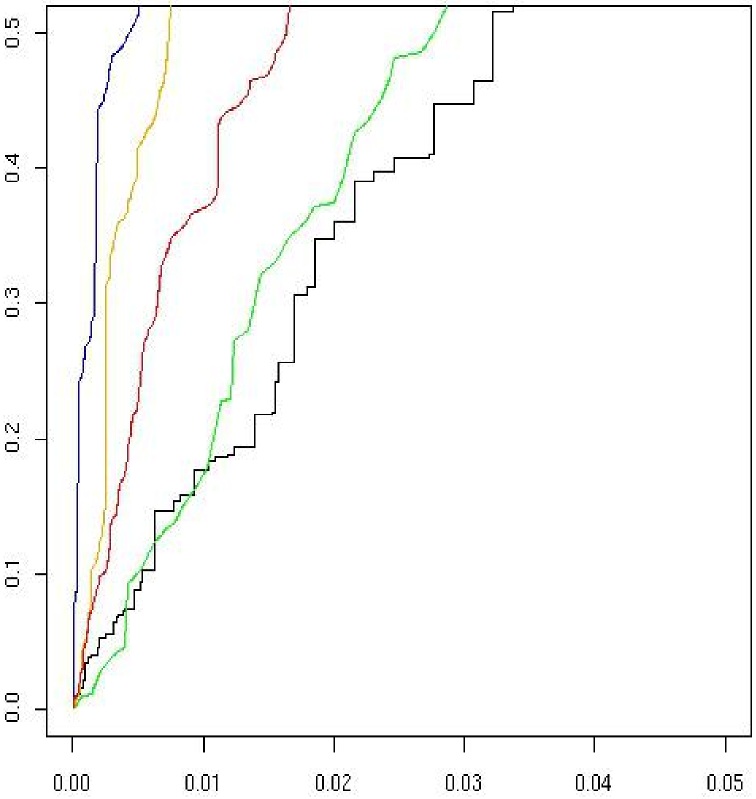


levure *Saccharomyces cerevisiae*
n=500, q=9, g=5786, KEGG Pathways

[Barriot *et al.* 2004]

Significativité des p-valeurs obtenues

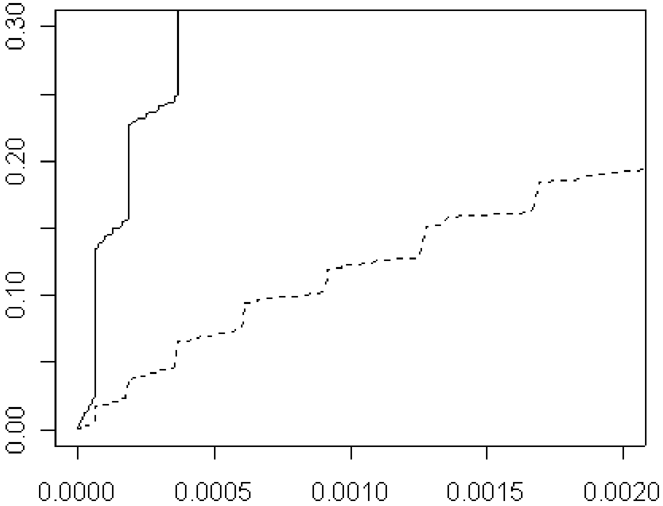
proportion d'ensembles requêtes aléatoires ayant une *p-valeur* ≤ x



x = *p-valeur*

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

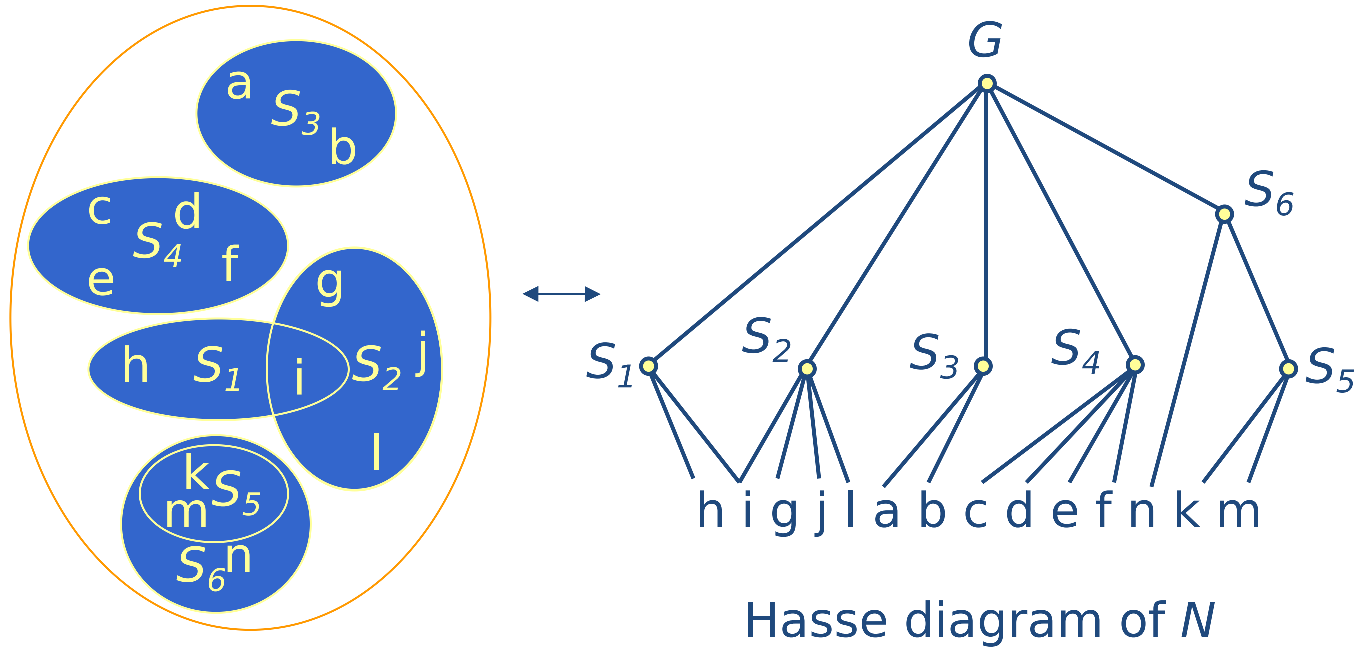
proportion d'ensembles requêtes aléatoires ayant une *p-valeur* ≤ x



x = *p-valeur*

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

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 ∈ N
- T pertinent if

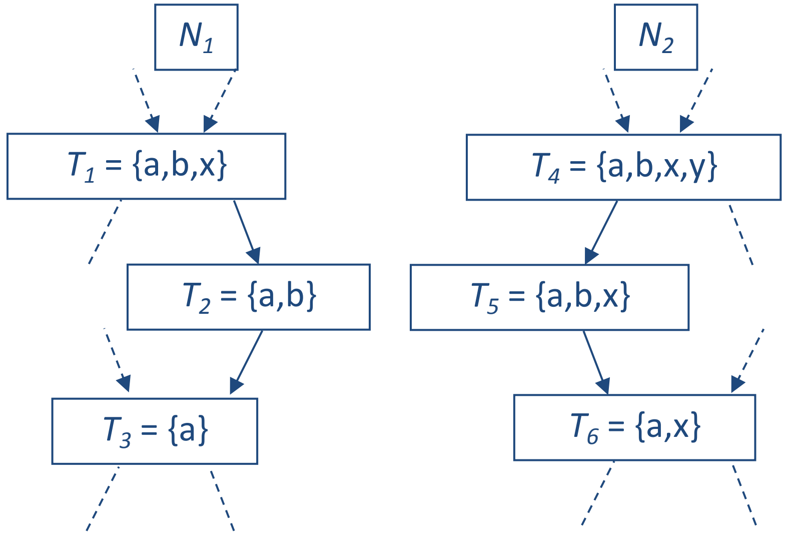
$$Q \cap T \neq \emptyset$$

and

$$\nexists\, T' \in N \text{ such that } T' \subset T \text{ and } T' \cap Q = T \cap Q$$

and

$$\nexists\, T' \in N \text{ such that } T \subset T' \text{ and } T' - Q = T - Q$$



N_1

$T_1 = \{a,b,x\}$

$T_2 = \{a,b\}$

$T_3 = \{a\}$

N_2

$T_4 = \{a,b,x,y\}$

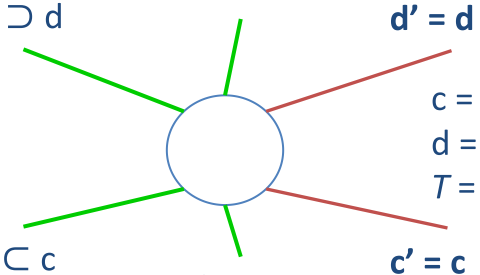
$T_5 = \{a,b,x\}$

$T_6 = \{a,x\}$

$Q = \{a,b,e\}$
$Q' = \{a,b\}$

Local decision

$|c| > 0$
$|d| < \min(\{d_{parents}\})$
$|c| > \max(\{c_{children}\})$



c' = c
d' ⊃ d

c' ⊃ c
d' ⊃ d

c' ⊃ c
d' = d

c = *T* ∩ *Q*
d = *T* - *Q*
T = c ∪ d

c' ⊂ c
d' ⊂ d

c' ⊂ c
d' = d

c' = c
d' ⊂ d

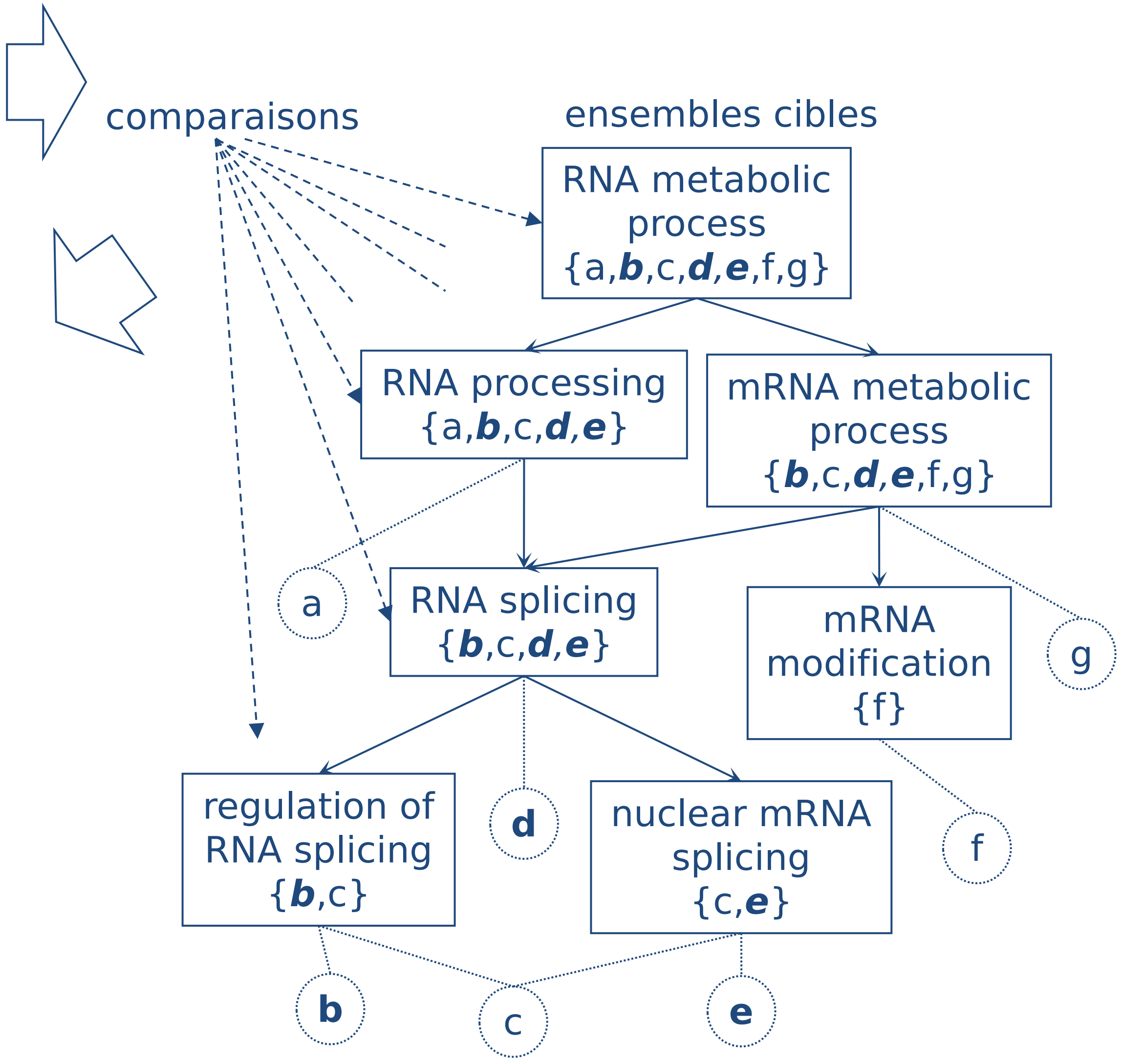
[Barriot, Dutour, Sherman, 2007, *BMC Bioinformatics*]

Illustration

- Pertinence des comparaisons & redondance des résultats

ensemble requête :
gènes d’intérêt

{b,d,e}

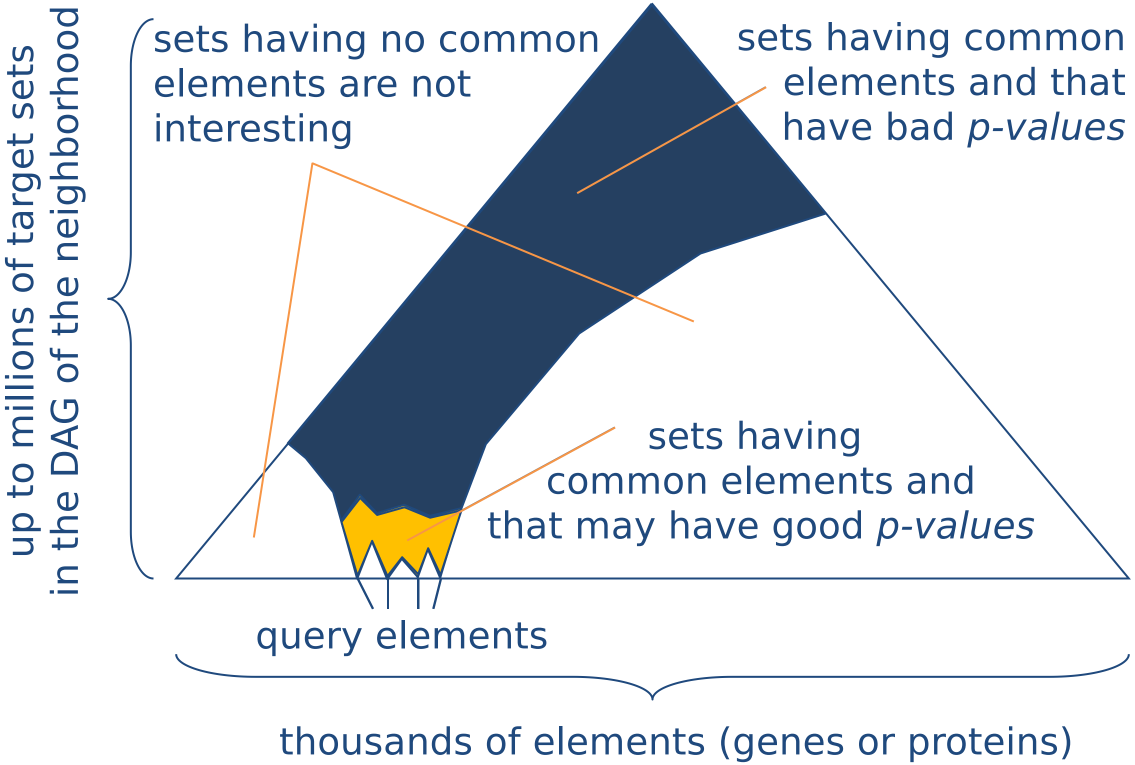

comparaisons
ensembles cibles
RNA metabolic process {a,***b***,c,***d***,***e***,f,g}
RNA processing {a,***b***,c,***d***,***e***}
mRNA metabolic process {***b***,c,***d***,***e***,f,g}
a
RNA splicing {***b***,c,***d***,***e***}
mRNA modification {f}
g
regulation of RNA splicing {***b***,c}
d
nuclear mRNA splicing {c,***e***}
f
b
c
e

résultats

similarité décroissante

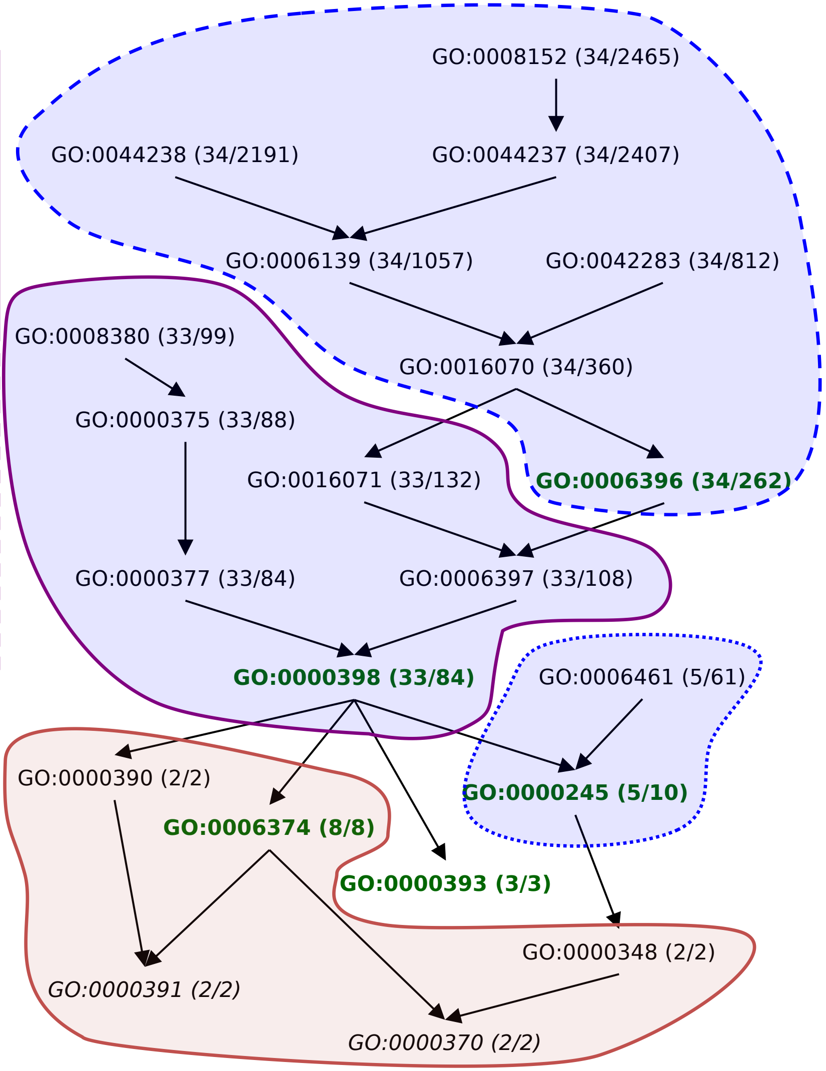
ensembles cibles similaires
RNA splicing {**b**,c,**d**,**e**}
RNA processing {a,**b**,c,**d**,**e**}
mRNA metabolic process {**b**,c,**d**,**e**,f,g}
…

A small portion of the DAG is searched


up to millions of target sets
in the DAG of the neighborhood
sets having no common elements are not interesting
sets having common elements and that have bad *p-values*
sets having common elements and that may have good *p-values*
query elements
thousands of elements (genes or proteins)

Complex 440.30.10 mRNA splicing

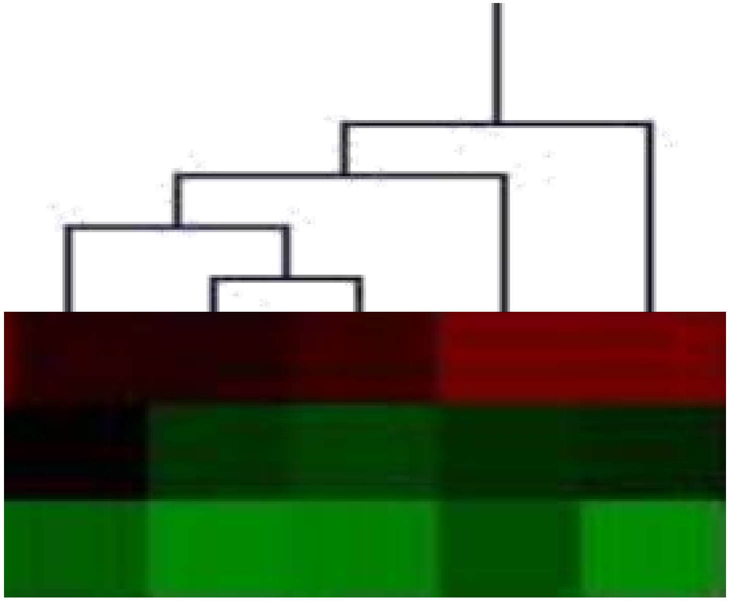
GO Term	Description	Target size	Common elements
GO:0000398	**nuclear mRNA splicing, via spliceosome**	**84**	**33**
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	84	33
GO:0000375	RNA splicing, via transesterification reactions	88	33
GO:0008380	RNA splicing	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	*U2-type spliceosome dissembly*	*2*	*2*
GO:0000390	spliceosome dissembly	2	2
GO:0000370	*U2-type nuclear mRNA branch site recognition*	*2*	*2*
GO:0000348	nuclear mRNA branch site recognition	2	2
GO:0000393	**spliceosomal conformational changes to generate catalytic conformation**	**3**	**3**



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

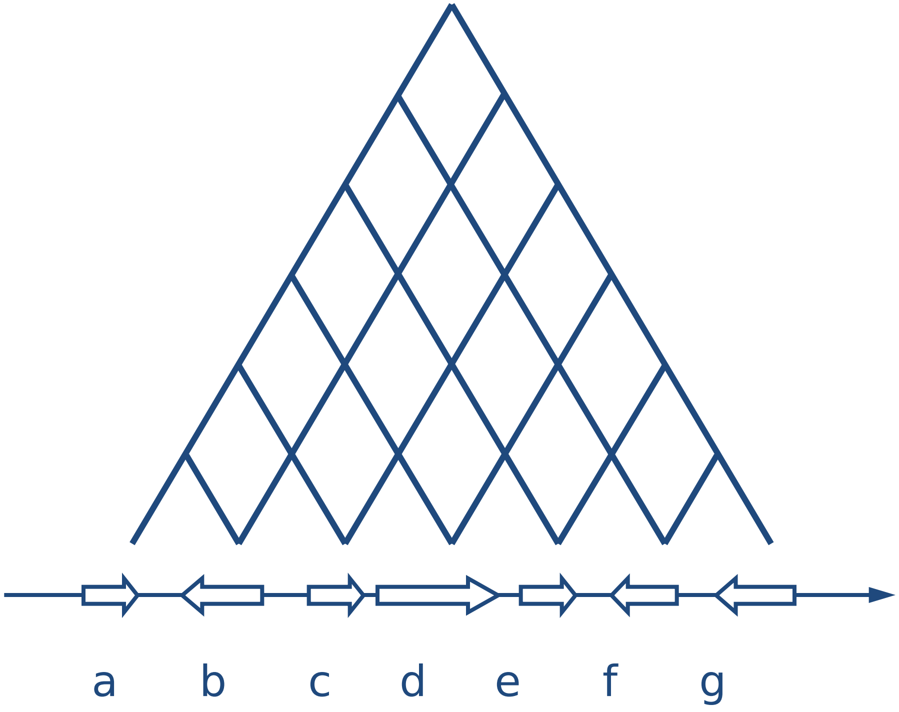


(((a (b c)) d) e)

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



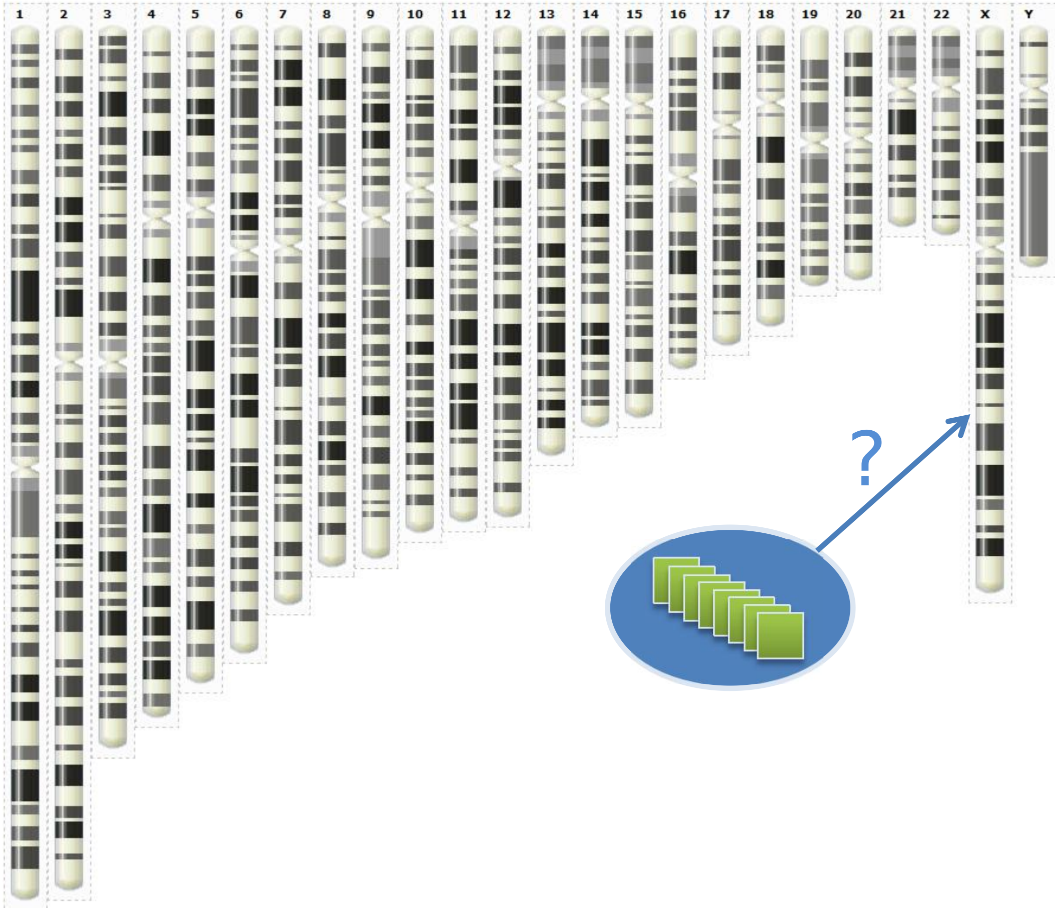
implicit

Context

Set of genes of interest

Examples

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



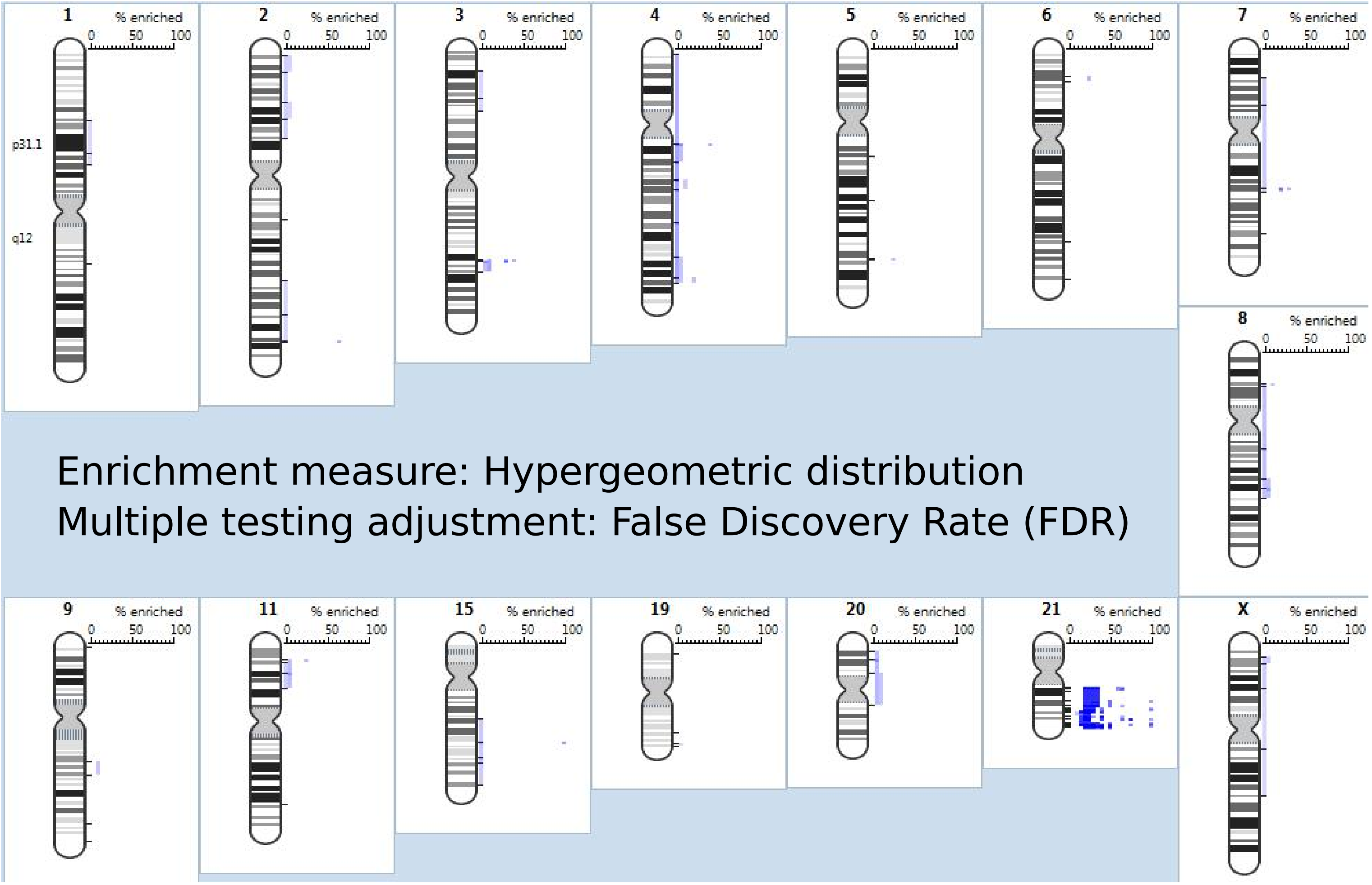
→ Question: Do those genes surprisingly cluster in the genome?

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.



Enrichment measure: Hypergeometric distribution
Multiple testing adjustment: False Discovery Rate (FDR)

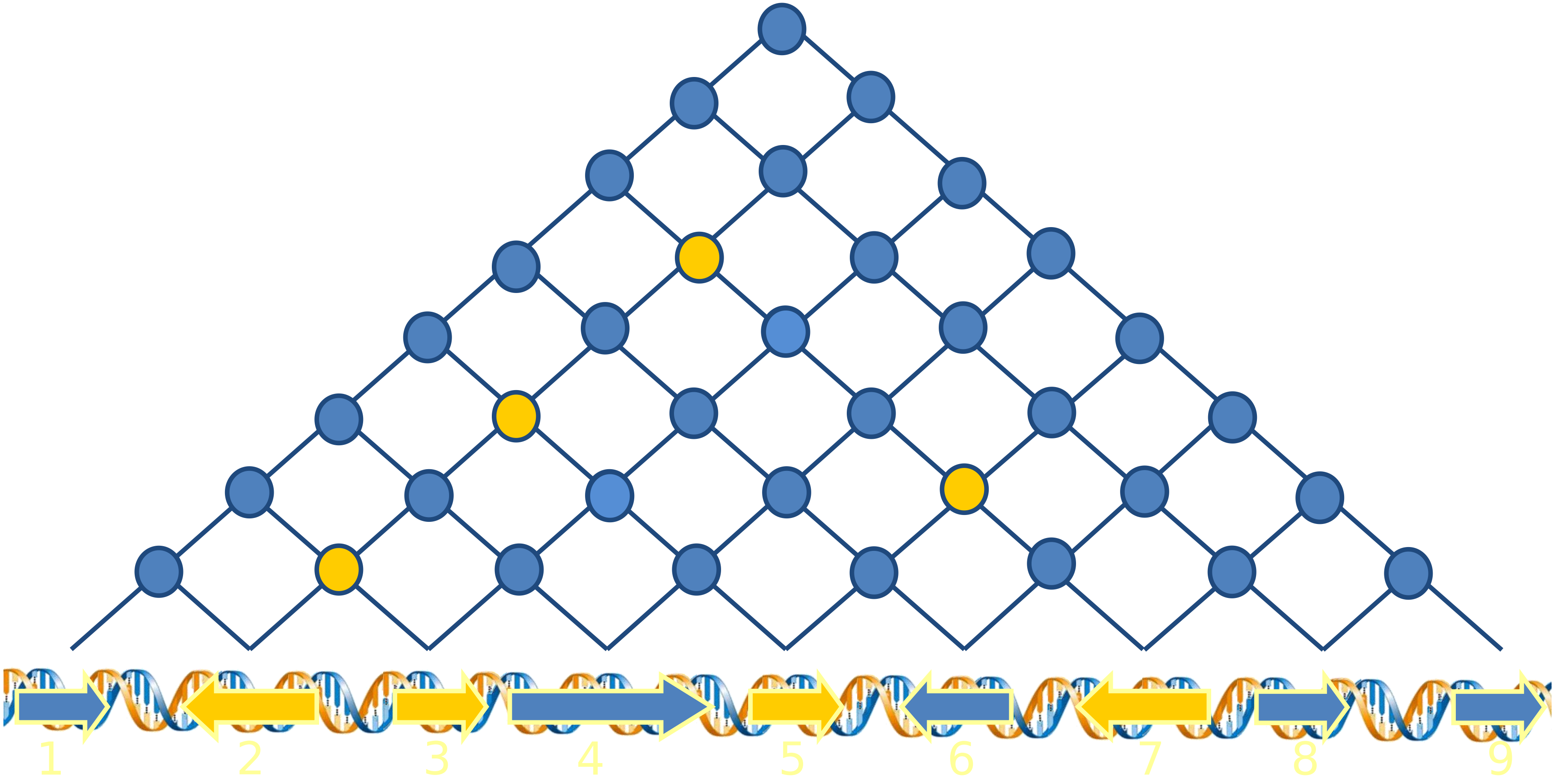
Issues:

- Number of regions to test
- False positives
- Redundancy

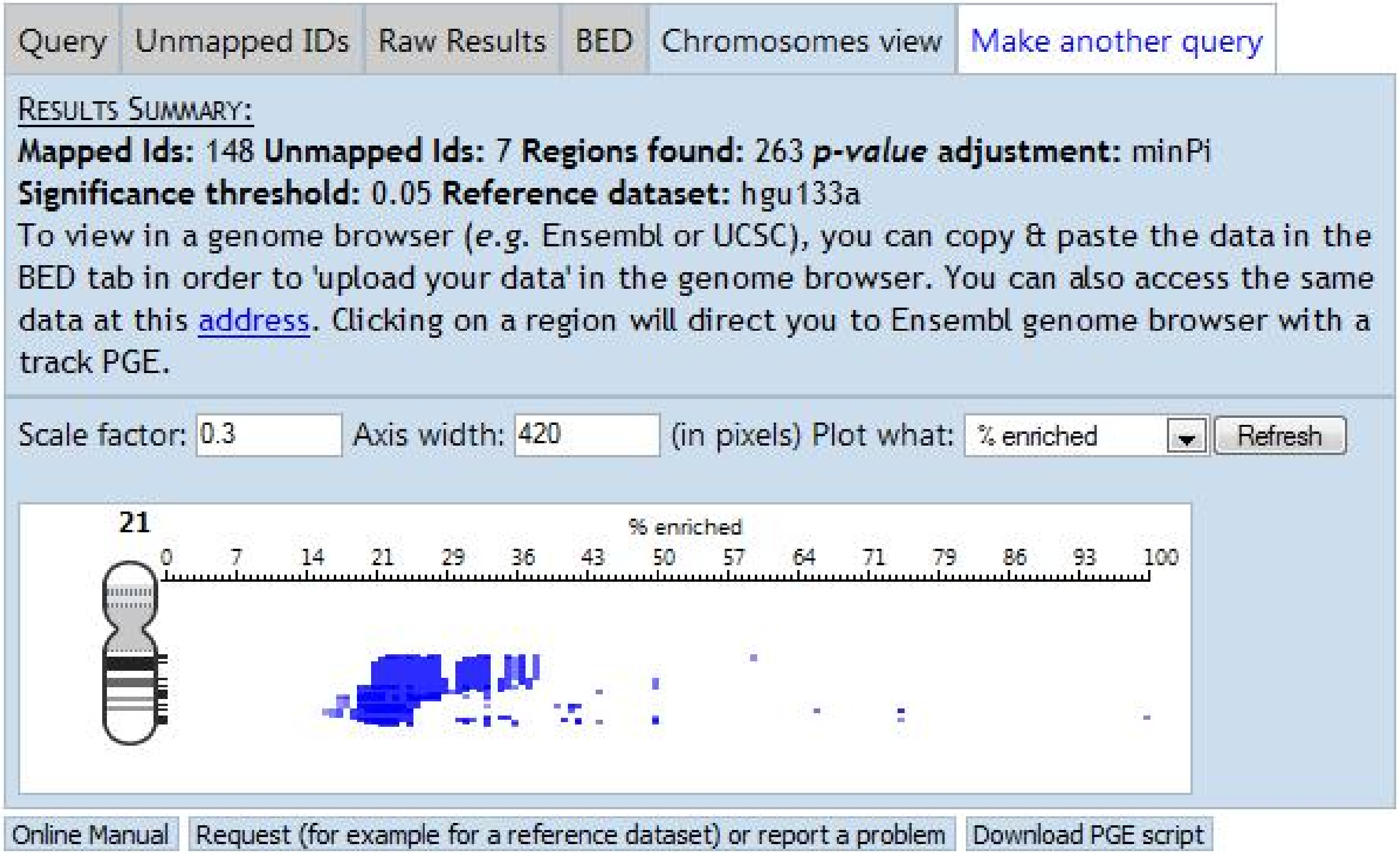
Pertinent Regions

A region is pertinent if it is:

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



Down Syndrome (minP_i)



RESULTS SUMMARY:

Mapped Ids: 148 **Unmapped Ids:** 7 **Regions found:** 263 ***p-value* adjustment:** minPi
Significance threshold: 0.05 **Reference dataset:** hgu133a

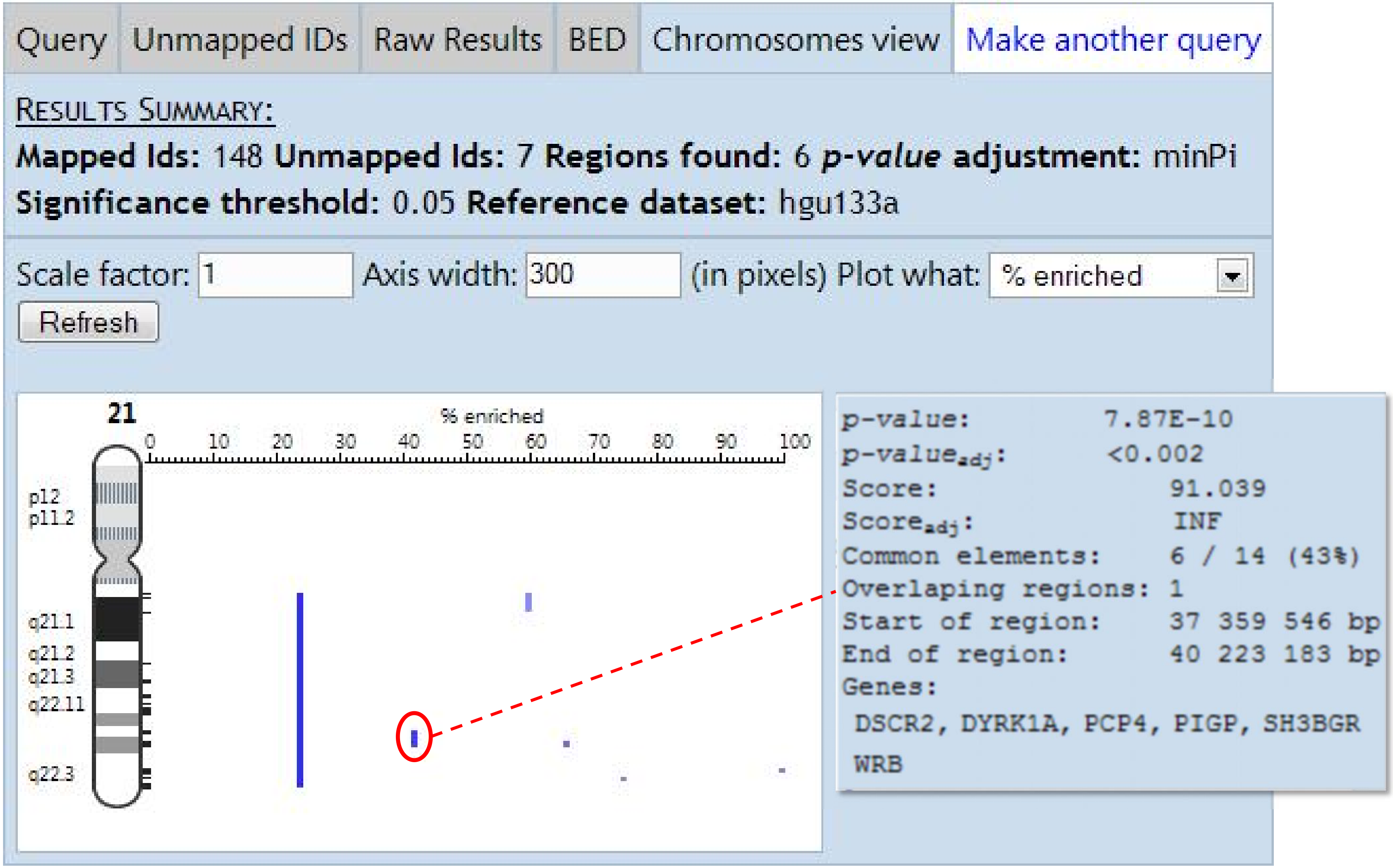
To view in a genome browser (*e.g.* Ensembl or UCSC), you can copy & paste the data in the BED tab in order to 'upload your data' in the genome browser. You can also access the same data at this address. Clicking on a region will direct you to Ensembl genome browser with a track PGE.

Scale factor: 0.3 Axis width: 420 (in pixels) Plot what: % enriched

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



RESULTS SUMMARY:

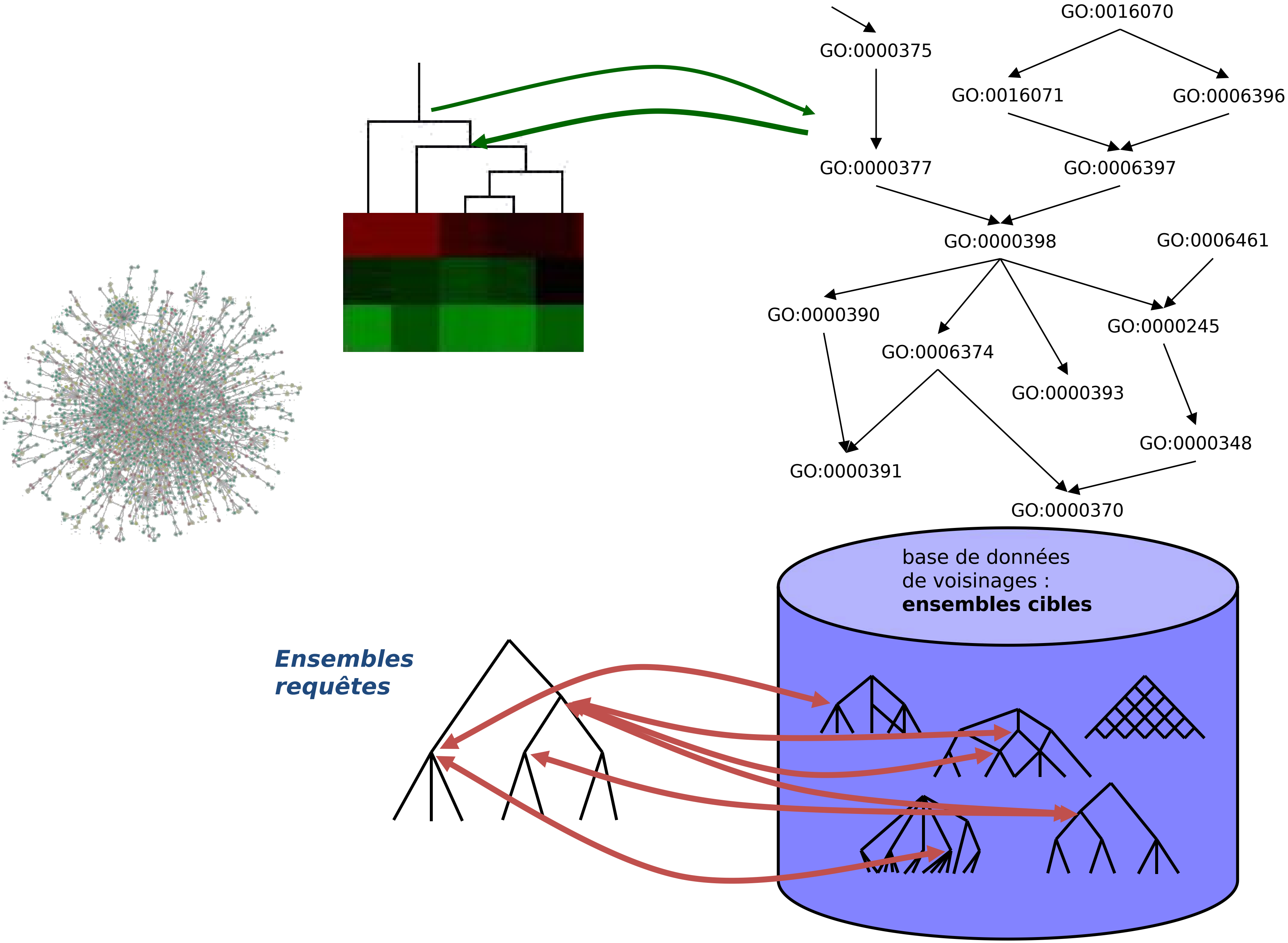
Mapped Ids: 148 **Unmapped Ids:** 7 **Regions found:** 6 ***p-value* adjustment:** minPi

Significance threshold: 0.05 **Reference dataset:** hgu133a

Scale factor: 1 Axis width: 300 (in pixels) Plot what: % enriched

```
p-value:             7.87E-10
p-value_adj:         <0.002
Score:                 91.039
Score_adj:             INF
Common elements:       6 / 14 (43%)
Overlaping 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



GO:0016070

GO:0000375

GO:0016071

GO:0006396

GO:0000377

GO:0006397

GO:0000398

GO:0006461

GO:0000390

GO:0000245

GO:0006374

GO:0000393

GO:0000348

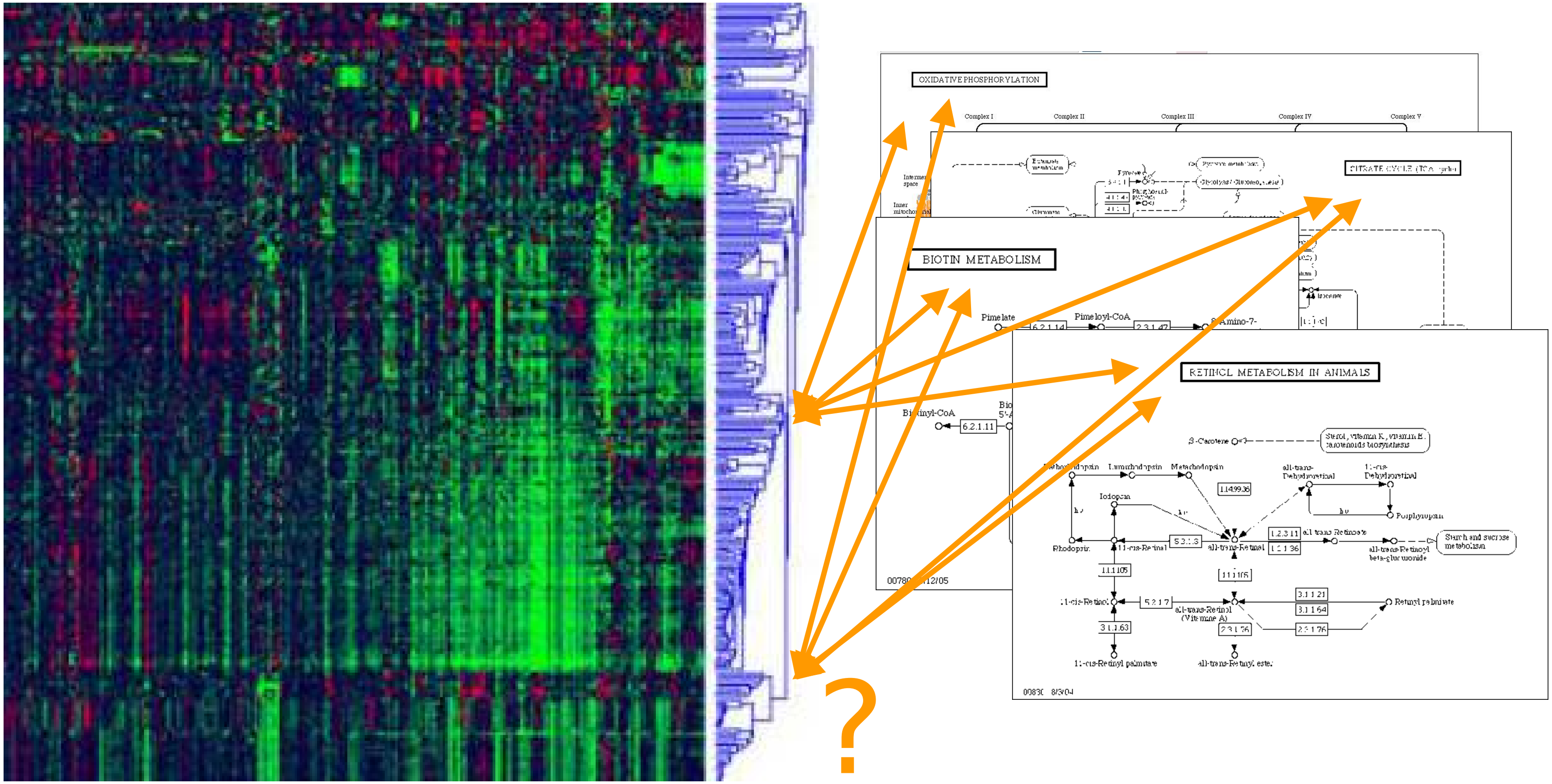
GO:0000391

GO:0000370

base de données
de voisinages :
ensembles cibles

***Ensembles
requêtes***

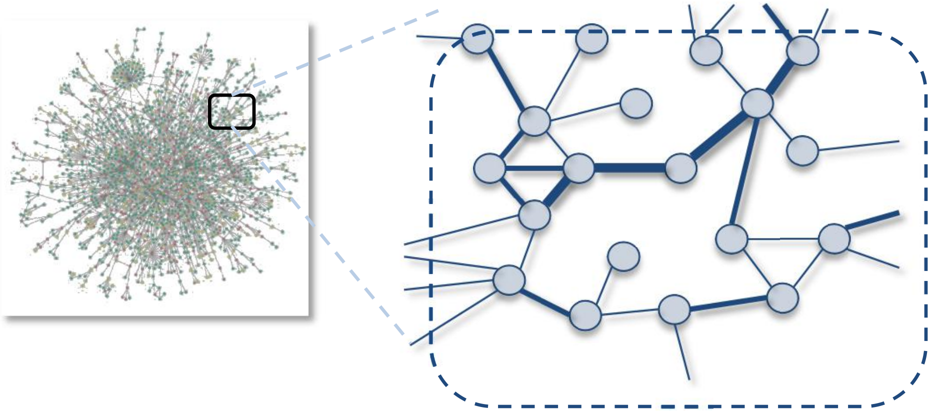
Analyse de données d’expression



[Ferea *et al.*, 1999]

[Kanehisa & Goto, 2000]

Extraction de sous-graphe pertinent & visualisation



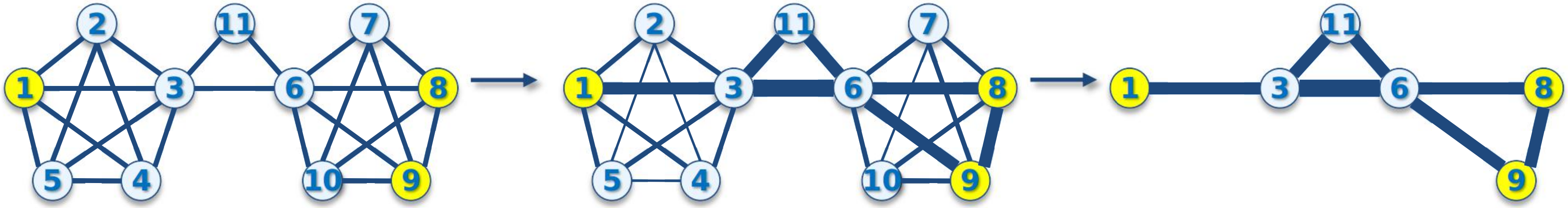
Idée :

- 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