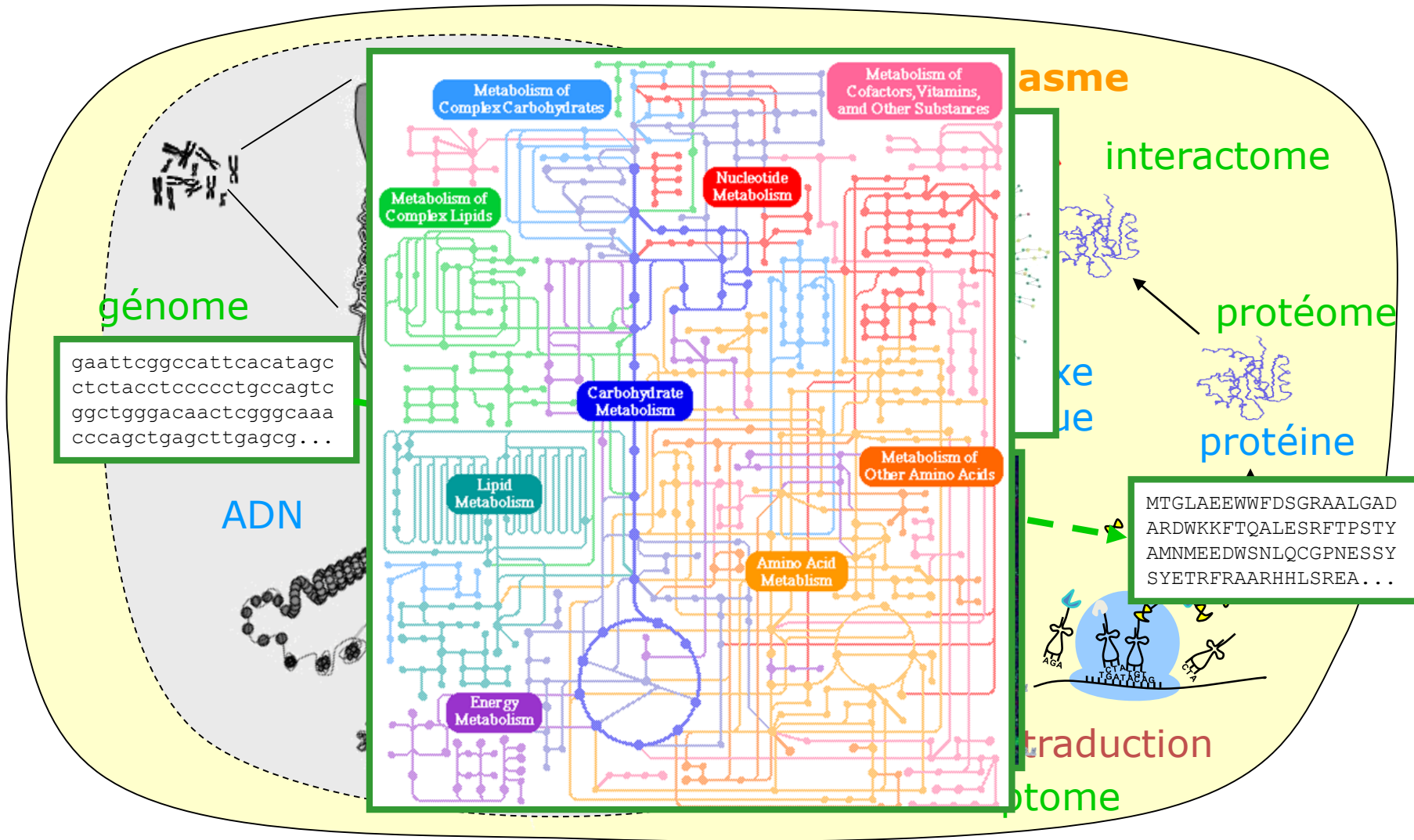


# Intégration de données hétérogènes

Master 2 MABS

Bioinformatique et Biologie des Systèmes

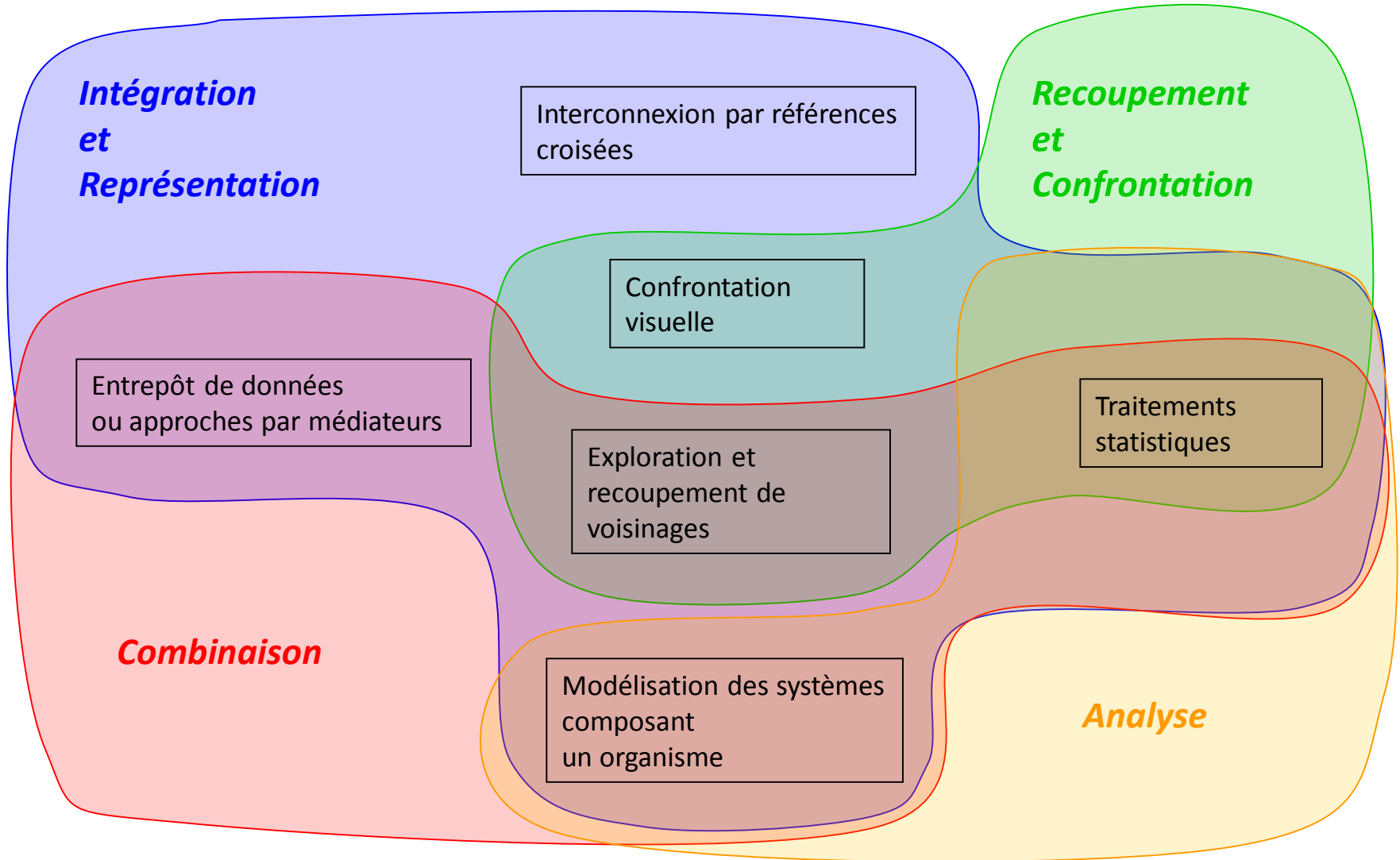
- Pourquoi ?
- Qu'est-ce que l'intégration ?
  - ◆ Interconnexion
  - ◆ Fusion
  - ◆ Médiation
  - ◆ Modélisation
  - ◆ Confrontation
  - ◆ Recoupement



Cellule eucaryote

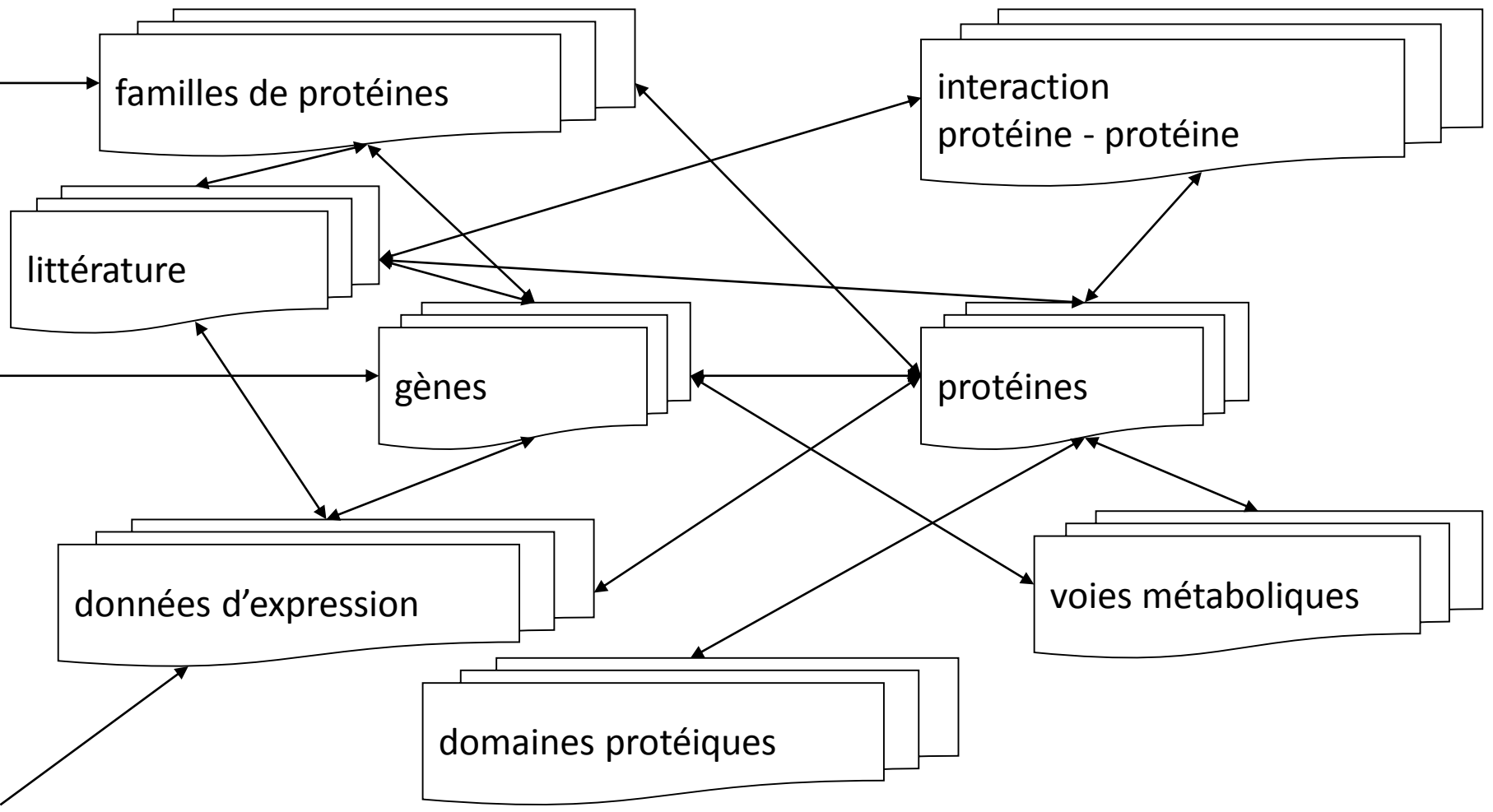
## Apperçu des données disponibles

- En quantité
- Dispersées
  - gènes, protéines, expression, interaction, ...
  - NCBI, EBI, KEGG, SIB, ...
- Hétérogènes : type, structure et sémantique
  - ♦ mots : séquence génome, gène, protéine
  - ♦ attributs
    - nominaux : mots-clés, ontologies, vocabulaires contrôlés
    - numériques :
      - ♦ niveaux d'expression,
      - ♦ usage des codons
  - ♦ graphes : interaction protéique, réactions enzymatiques, transduction du signal, structures classificatoires
  - ♦ texte
    - vocabulaire contrôlé
    - littérature

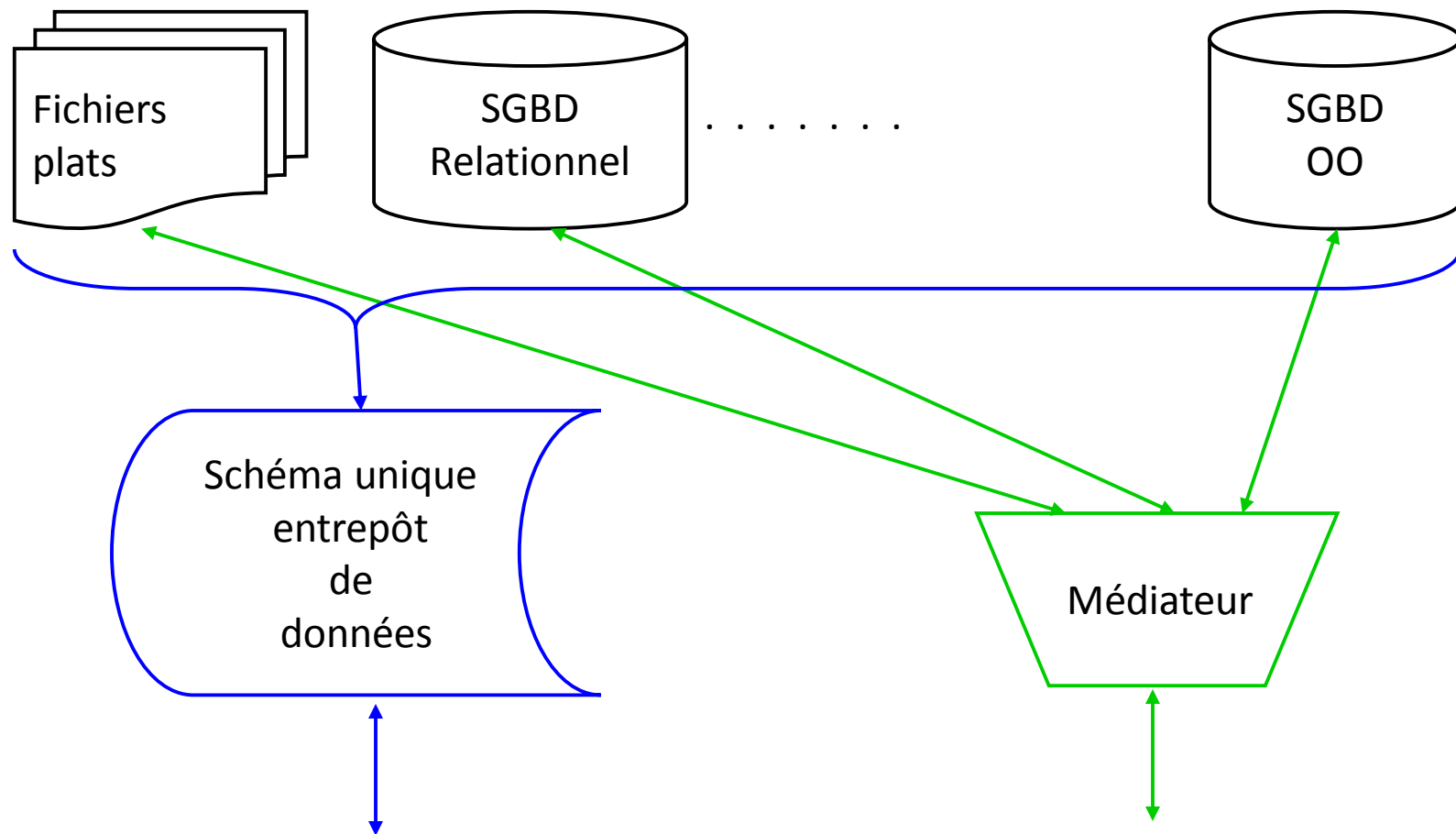


- Exploitation des références (croisées)
  - ◆ interconnexion
  - ◆ schéma unifié matérialisé : entrepôt
  - ◆ schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - ◆ exploration
  - ◆ recoupement
  - ◆ confrontation
  - ◆ fusion

# Intégration par interconnexion : principe



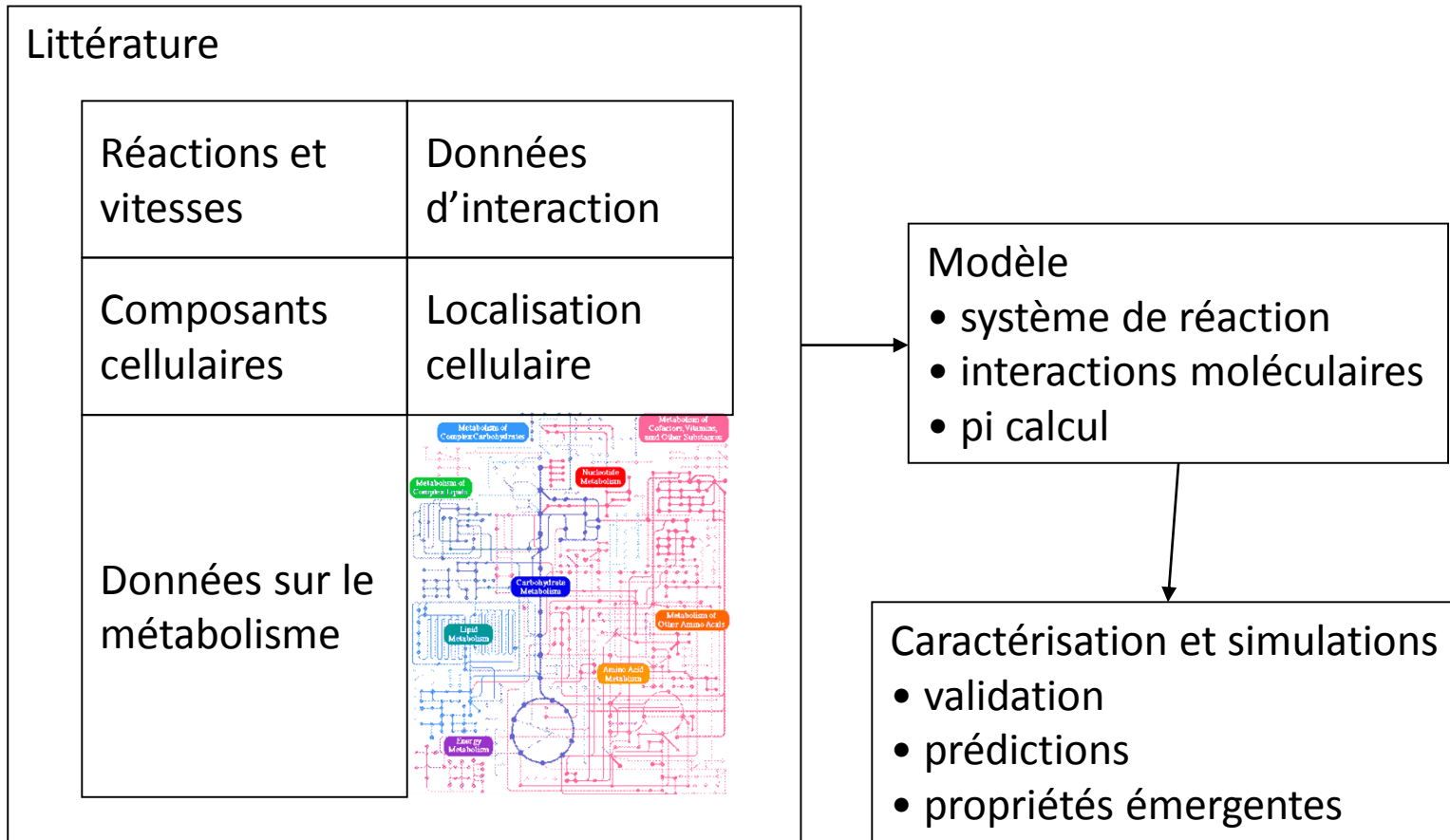
SRS [Etzold et al., 1996], Entrez [Schuler et al., 1996], ...



Integr8 [Kersey et al., 2005], BioMart [Kasprzyk et al., 2004],  
WInGS [Abergel et al., 2004], BioKleisli [Davidson et al., 1997], ...



- Exploitation des références (croisées)
  - ◆ interconnexion
  - ◆ schéma unifié matérialisé : entrepôt
  - ◆ schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - ◆ exploration
  - ◆ recoupement
  - ◆ confrontation
  - ◆ fusion



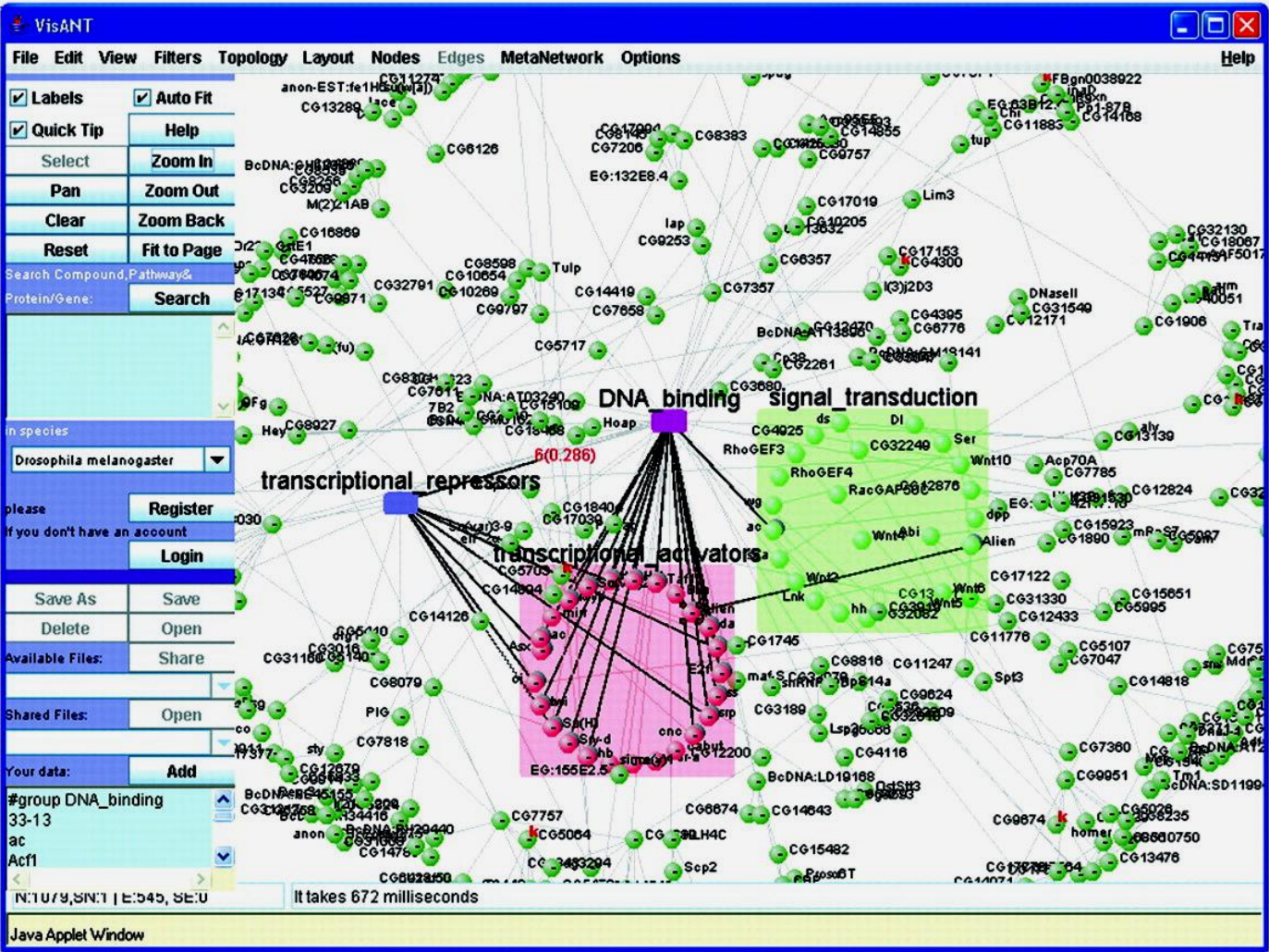
Virtual Cell [Loew et Schaff, 2001], E-CELL [Tomita et al., 1999],  
 Cellerator [Shapiro et al., 2003], ...

- Exploitation des références (croisées)
  - ◆ interconnexion
  - ◆ schéma unifié matérialisé : entrepôt
  - ◆ schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - ◆ exploration
  - ◆ recoupement
  - ◆ confrontation
  - ◆ fusion



- Exploitation des références (croisées)
  - ◆ interconnexion
  - ◆ schéma unifié matérialisé : entrepôt
  - ◆ schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - ◆ exploration
  - ◆ recoupement
  - ◆ confrontation
  - ◆ fusion

# Confrontation visuelle



Visant [Hu et al., 2005]

- Exploitation des références (croisées)
  - ◆ interconnexion
  - ◆ schéma unifié matérialisé : entrepôt
  - ◆ schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - ◆ exploration
  - ◆ recoupement
  - ◆ confrontation
  - ◆ fusion

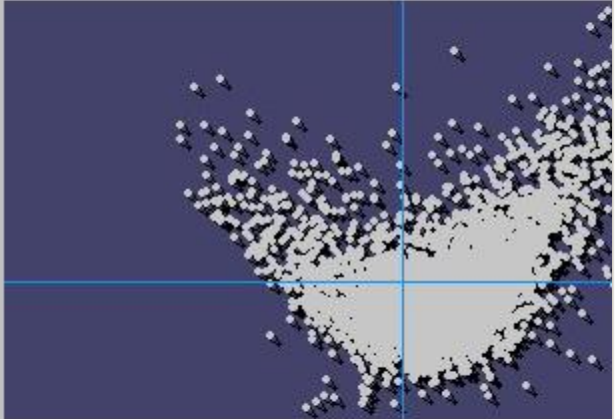


# Exploration visuelle de voisinages

### argB neighbours

Swiss Prot
Classification
Codons

Bibliography
pI
Save / Print



**Neighbor genes**

- argA
- ybjD
- ybbB
- yeiE

Gene Neighborhood
Codons Usage

### argH neighbours

Swiss Prot
Classification
Codons
Pathway

Bibliography
pI
Save / Print

#### Bibliography

Select a level below to display neighbor genes

**Escherichia coli and Salmonella typhimurium cellular and molecular biology.**  
 F. Neidhardt, R. Curtiss III, J. Ingraham, E. Lin, K. Brooks Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter and H. Umberg

**Variations on a Theme by Escherichia**

[Genome Structure](#)

**Biosynthesis of Arginine and Polyamines**

[Arginine Biosynthetic Enzymes](#)

- N-Acetylglutamokinase
- N-Acetylglutamylphosphate Reductase
- Argininosuccinase

[Arginine Reulon](#)

**Neighbor genes**

Gene Neighborhood

Codons Usage

Java Applet Window

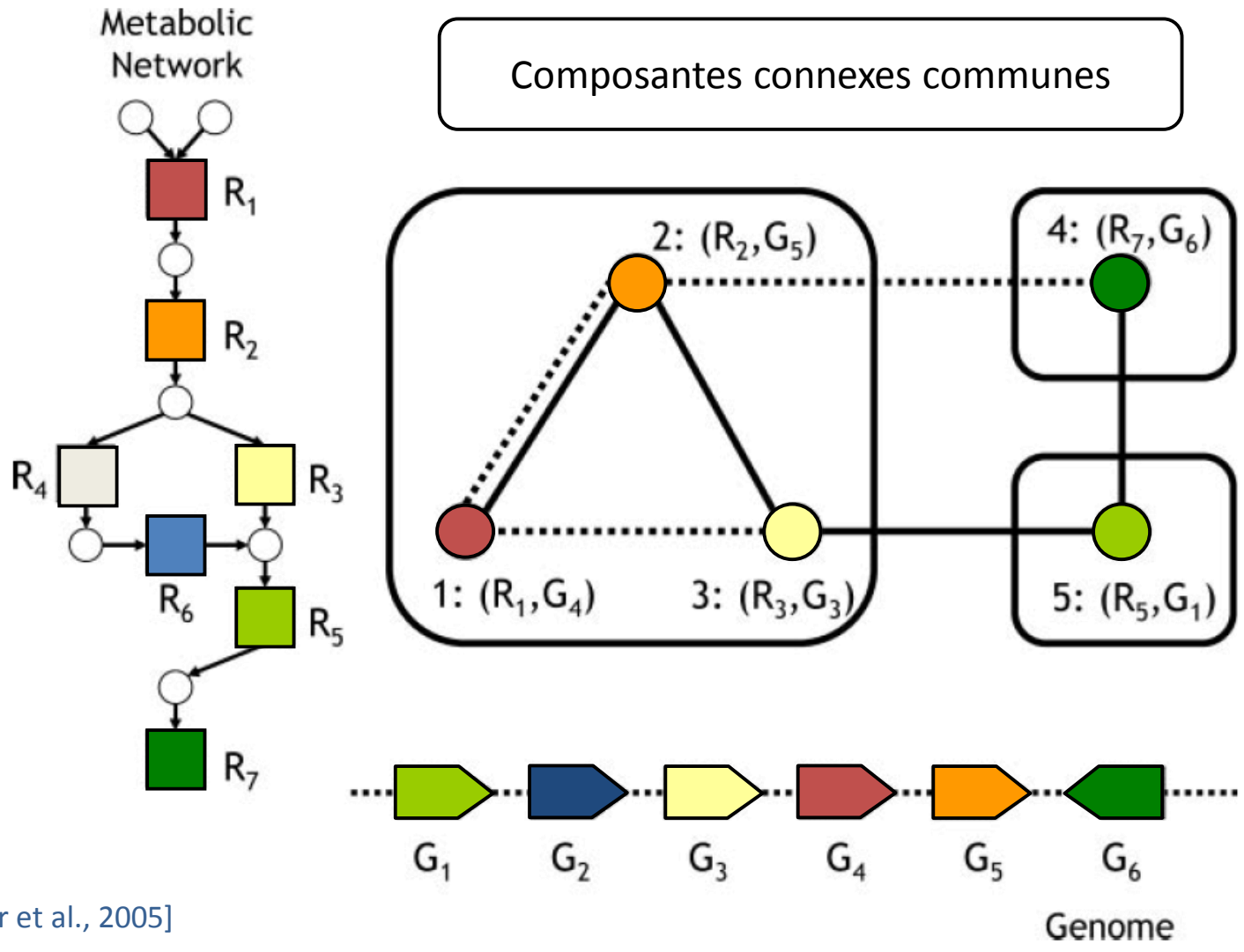
Delete

Java Applet Window

Java Applet Window

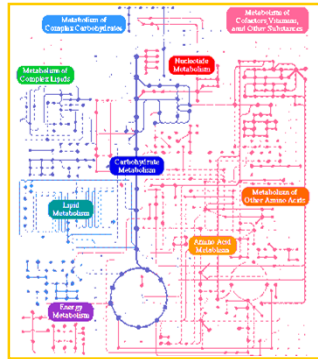


# Recoupement de voisinages : approche graphique



[Boyer et al., 2005]

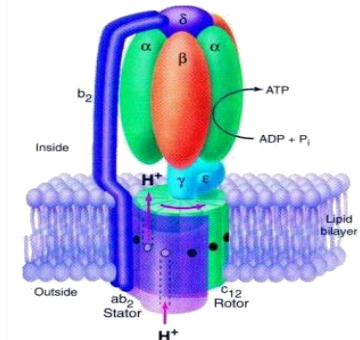
# Recoupement de voisinages : approche ensembliste



voies  
métaboliques

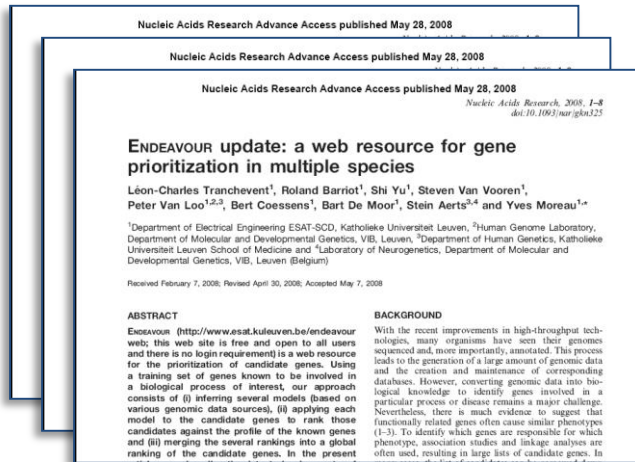


localisation chromosomique

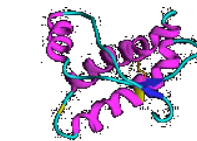


complexes  
protéiques

ensembles de gènes



co-citation

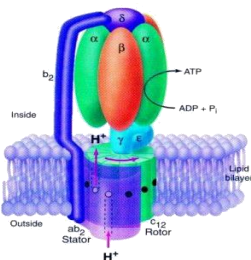
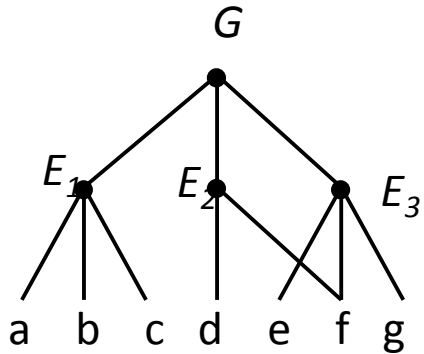


domaines  
protéiques

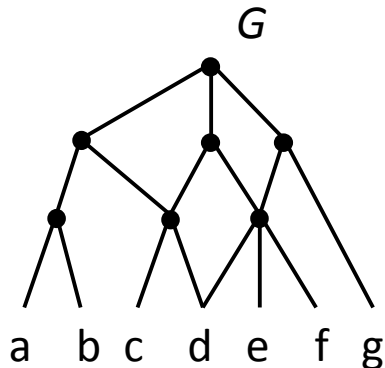
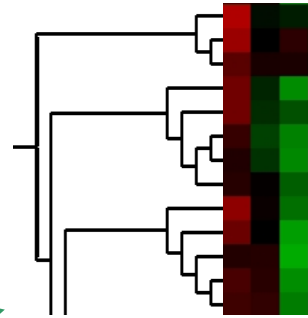


Gene  
Ontology

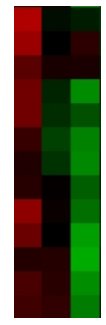
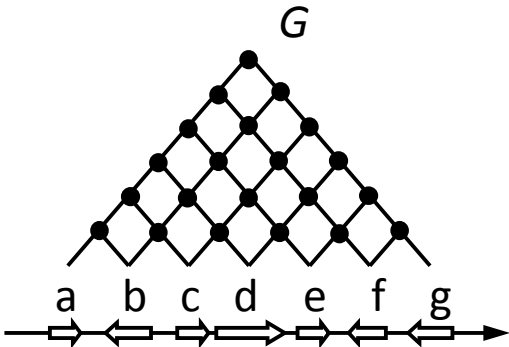
Ensembles



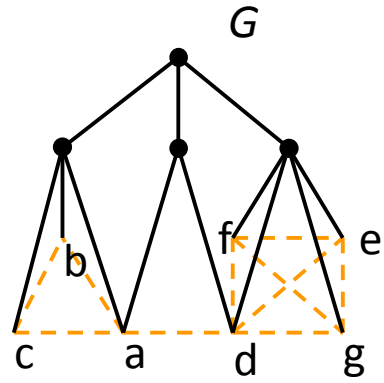
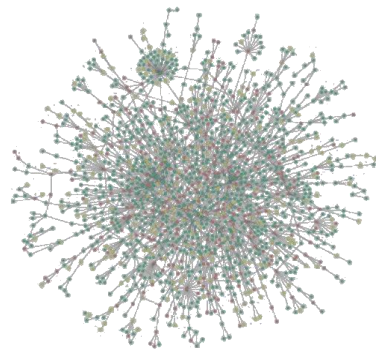
Hiérarchies



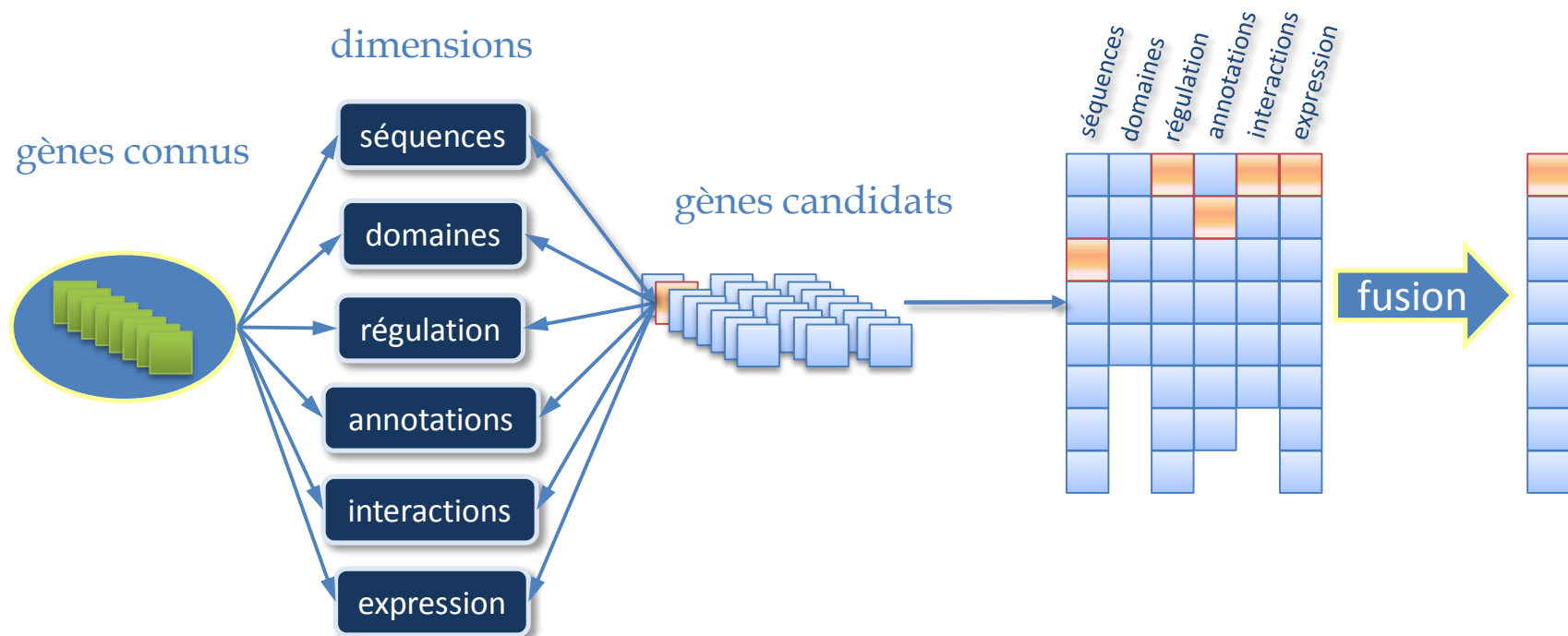
Vecteurs



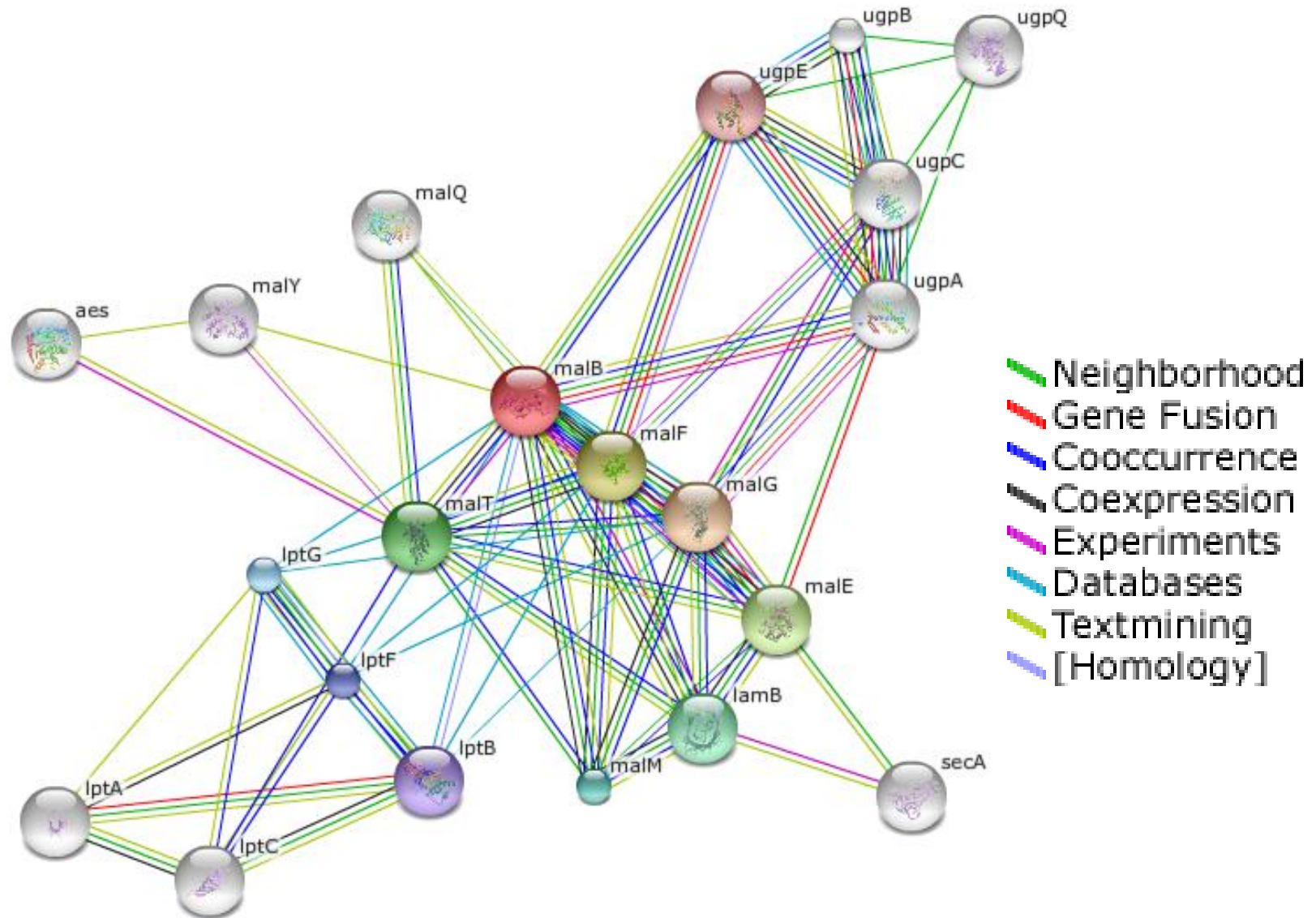
Graphes



- Exploitation des références (croisées)
  - ◆ interconnexion
  - ◆ schéma unifié matérialisé : entrepôt
  - ◆ schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - ◆ exploration
  - ◆ recoupement
  - ◆ confrontation
  - ◆ fusion



# Fusion : approche graphique



- Sémantique
  - ◆ Gene Ontology BP/KEGG pathways/BioCyc
  - ◆ structure d'un gène/peptide
- Modèle et format
  - ◆ SGBDr, SGBDoo, LDAP, fichier
    - schémas, attributs et unités
  - ◆ XML, FASTA, EMBL/GenBank, SWISSPROT
- Architecture et accès
  - ◆ SGBD, SOAP, REST, pipeline (galaxy, ergatis)

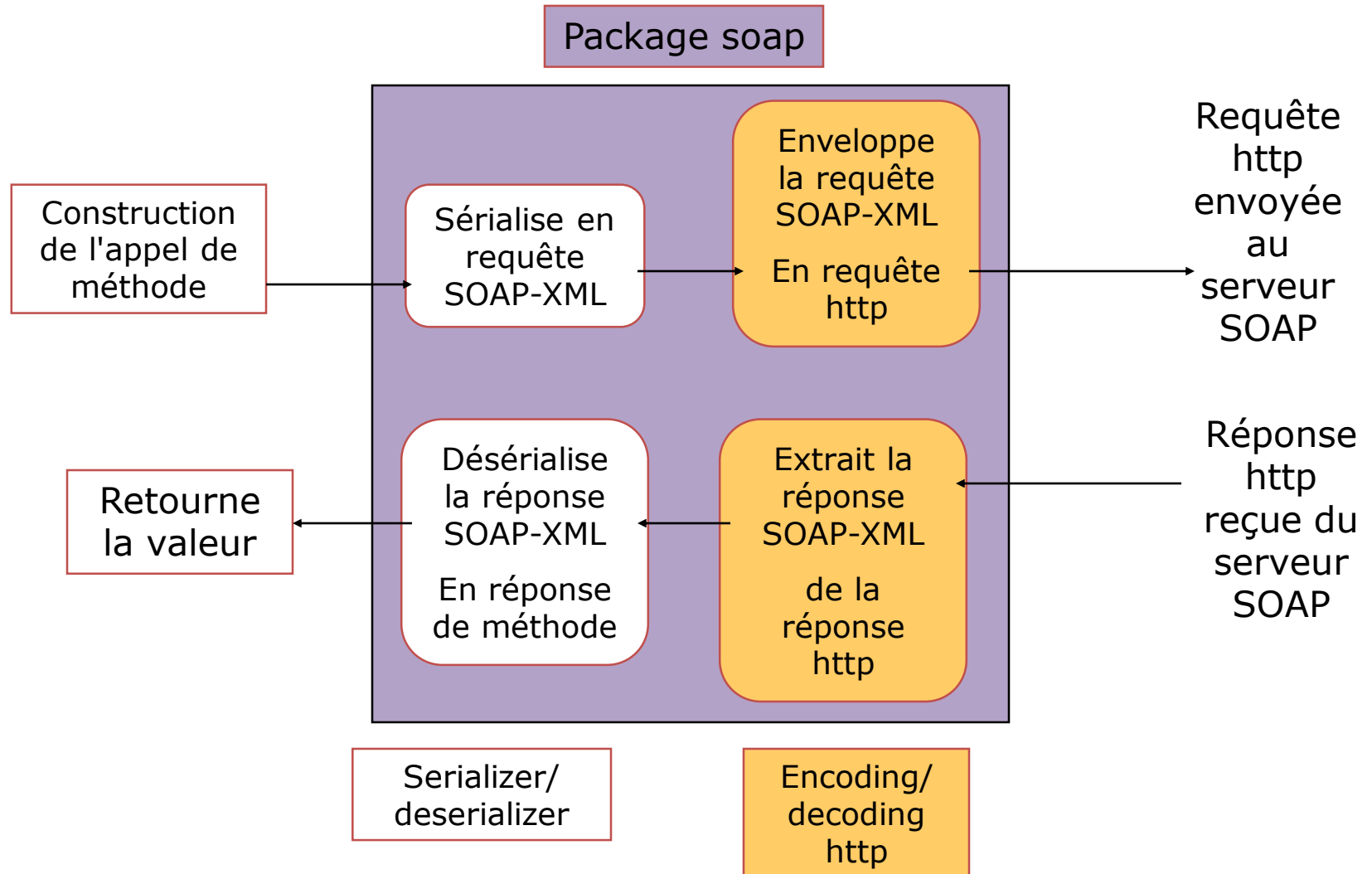
- Hétérogénéité des systèmes et des représentations
- Besoin de standards
- Une solution :
  - ◆ Format d'échange : XML
  - ◆ Protocole d'accès : services Web



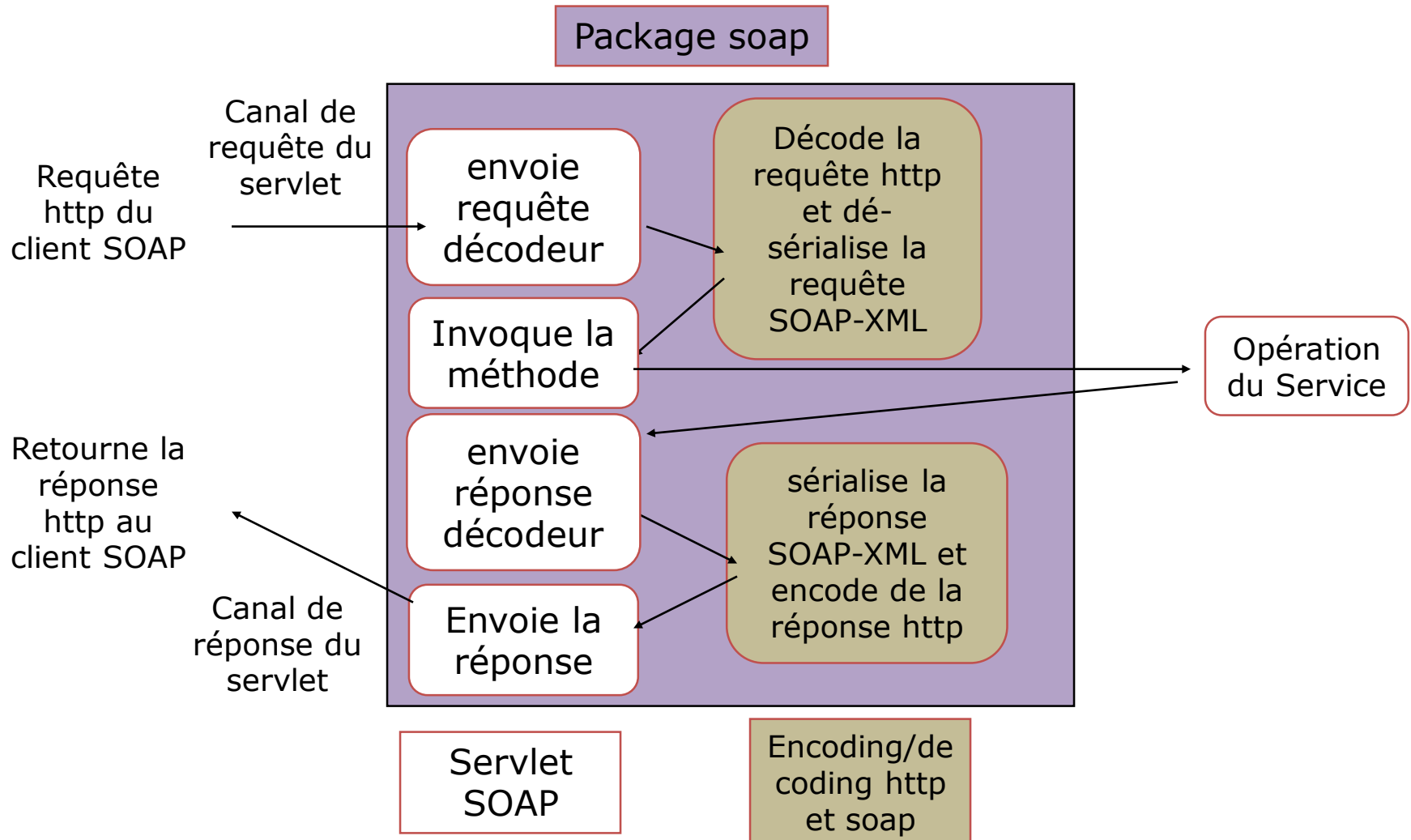
- Principe
  - ◆ appel de procédure à distance
  - ◆ permet de faire communiquer différents systèmes (plateforme, OS, langage de programmation, ...) à travers XML
  - ◆ un fournisseur propose certaines fonctionnalités
  - ◆ description du mode d'accès à ces fonctions : WSDL (Web Service Description Language)
  - ◆ protocole de communication : SOAP (Simple Object Access Protocol)
    - couche transport (HTTP, SMTP, POP3, IMAP, ...)
    - représentation XML : encodage/décodage des données

- Côté fournisseur :
  - ◆ serveur : Web, SMTP ou autre
  - ◆ ex: script CGI (apache-perl-SOAP)
- Côté client :
  - ◆ de nombreuses bibliothèques disponibles
    - perl, ruby, python, Java, ...

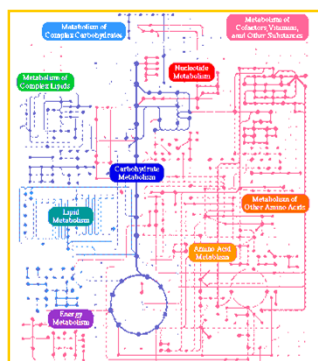
- Processus Client avec binding http



- Processus Serveur avec binding http



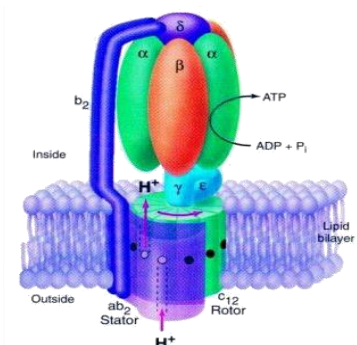
# Recoupement de voisinages : approche ensembliste



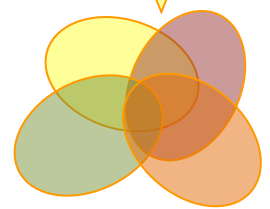
voies métaboliques



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/gln225

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

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

<sup>1</sup>Department of Electrical Engineering ESAT-SCD, Katholieke Universiteit Leuven, <sup>2</sup>Human Genome Laboratory, Department of Molecular and Developmental Genetics, VIB, Leuven, <sup>3</sup>Department of Human Genetics, Katholieke Universiteit Leuven School of Medicine and <sup>4</sup>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, 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 (1-5). 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

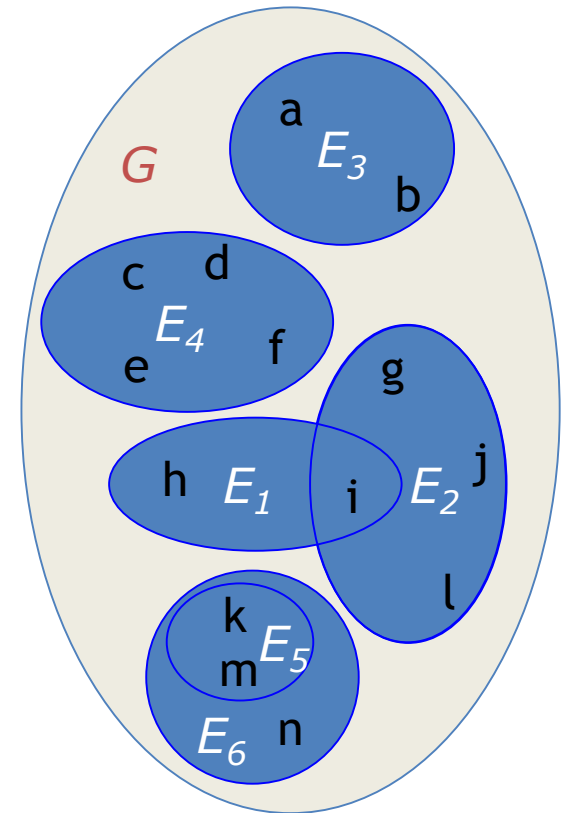


domaines protéiques



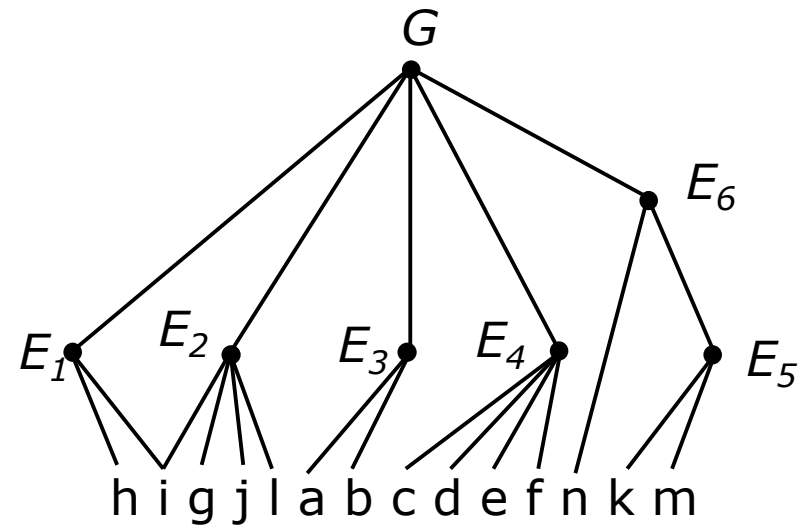
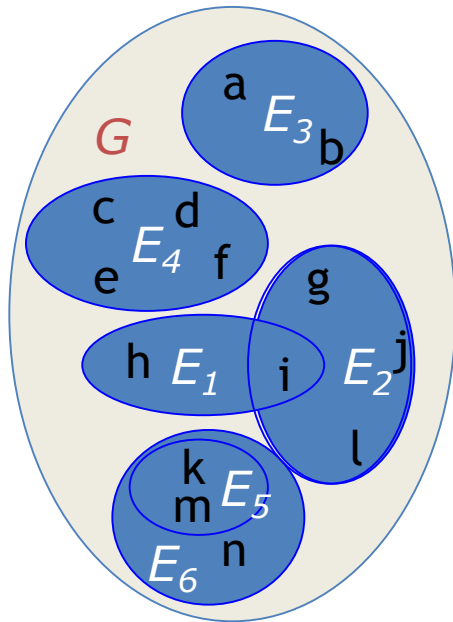
Gene Ontology

- (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.



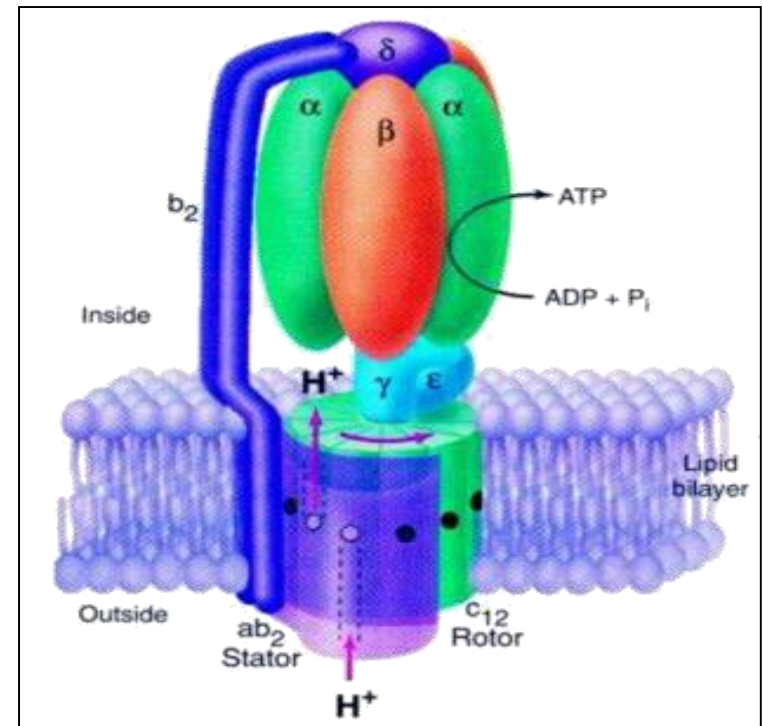
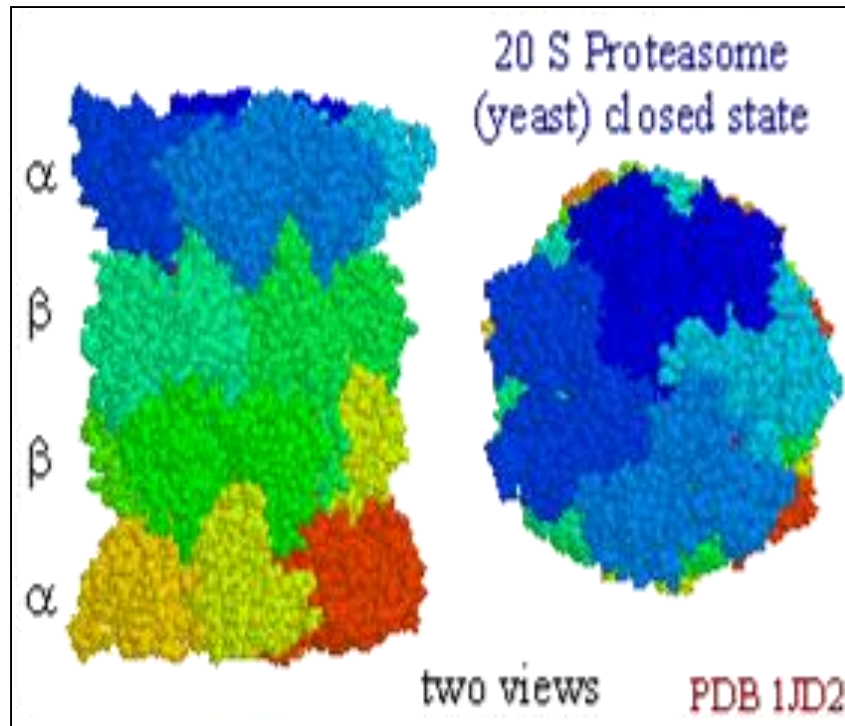
$$V = \{E_1, E_2, E_3, E_4, E_5, E_6\} \subseteq P(G)$$

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

diagramme de Hasse de  $V$ 

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

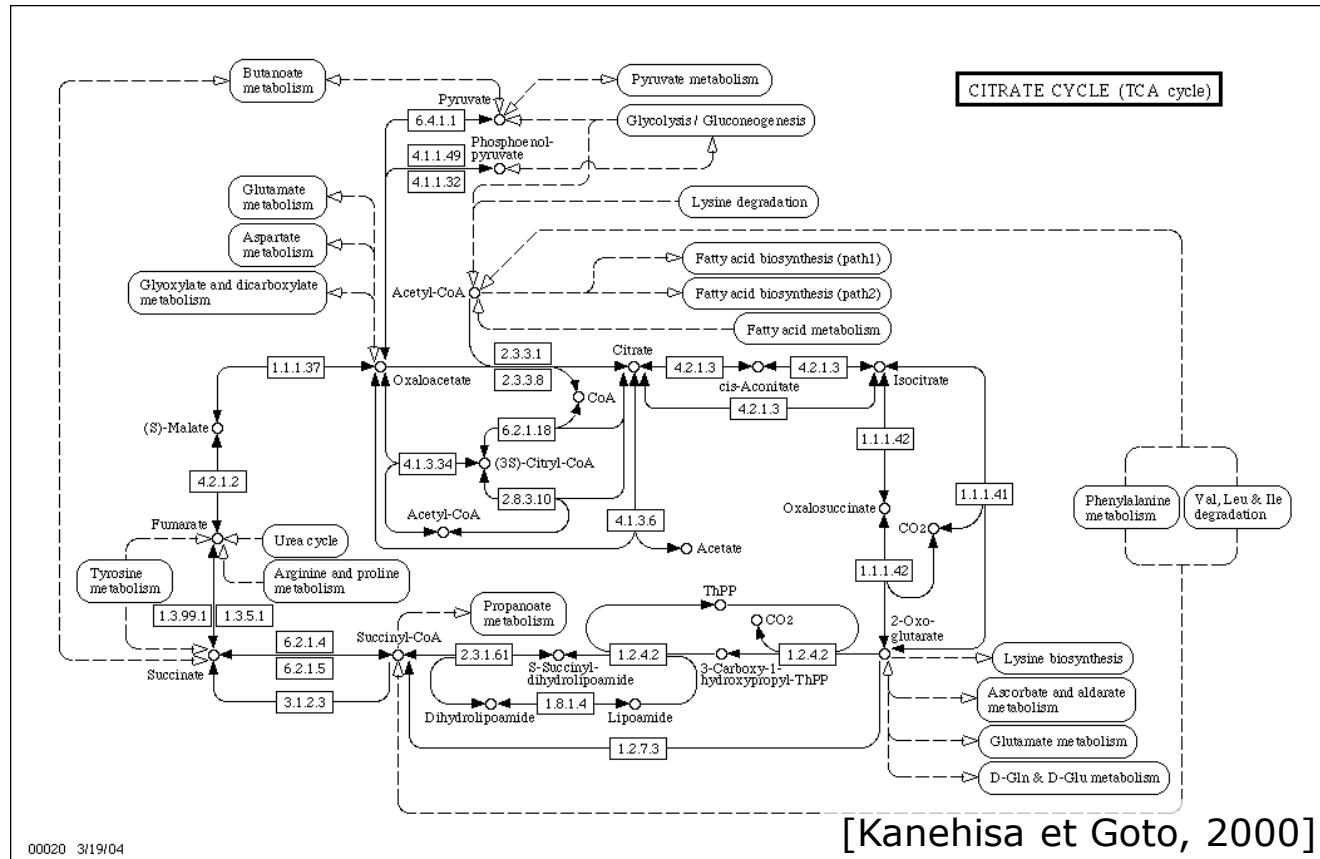
# Exemple de fonction de regroupement : complexes protéiques



un complexe → un ensemble de protéines



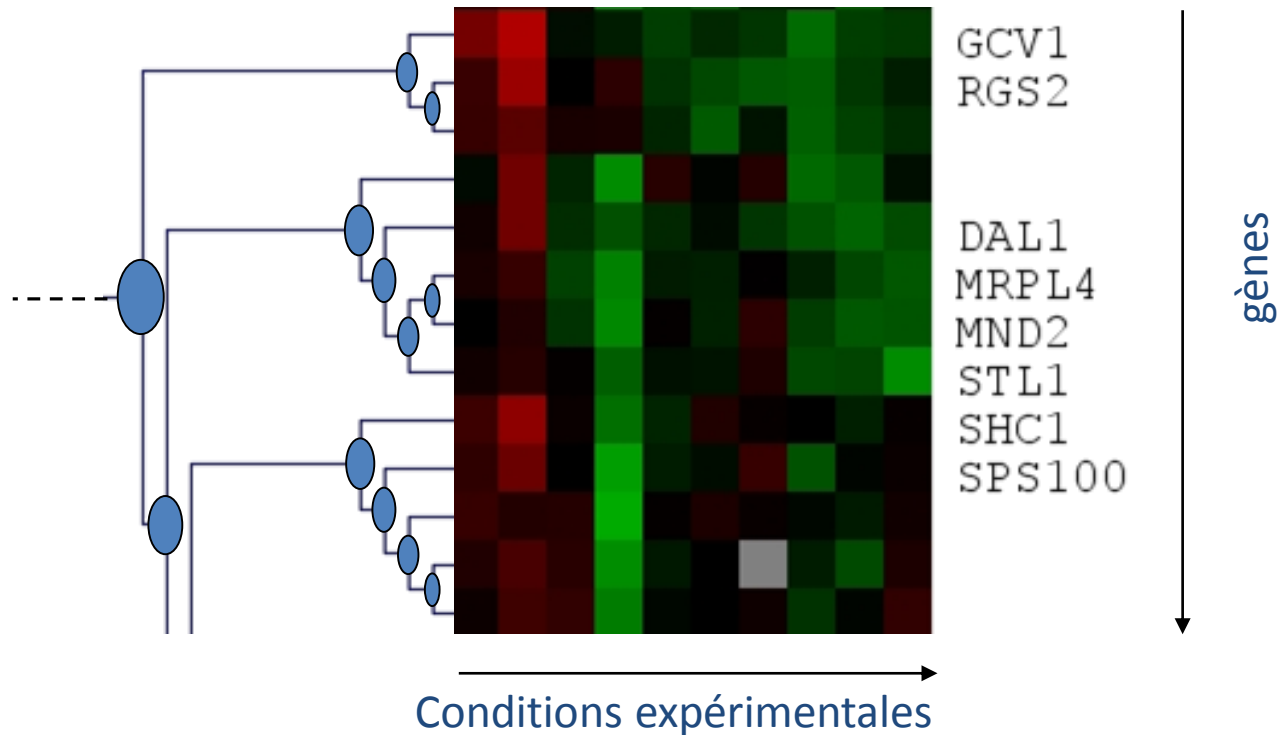
# Exemple de critère de regroupement : voies métaboliques



une voie métabolique → un ensemble de protéines

# Exemple de critère de regroupement : données d'expression

clustering hiérarchique  
des profils

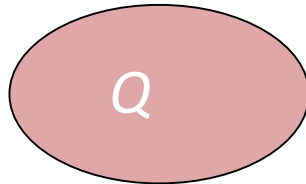


un cluster → un ensemble de gènes

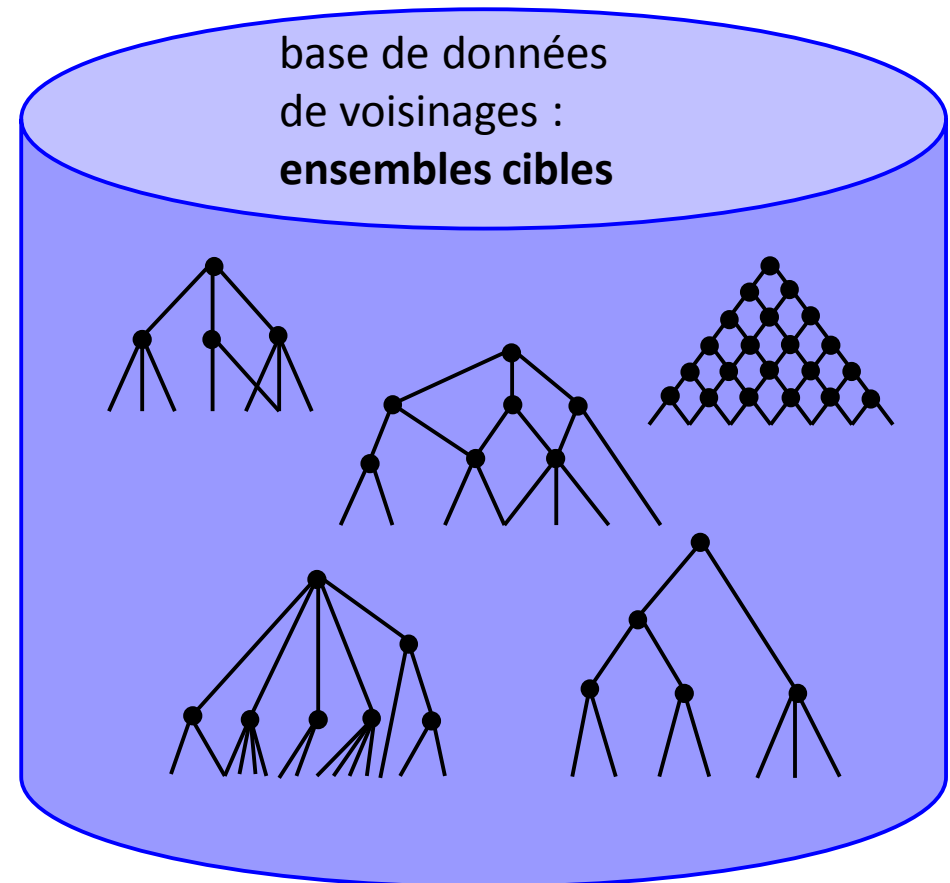
- Recherche d'ensembles similaires

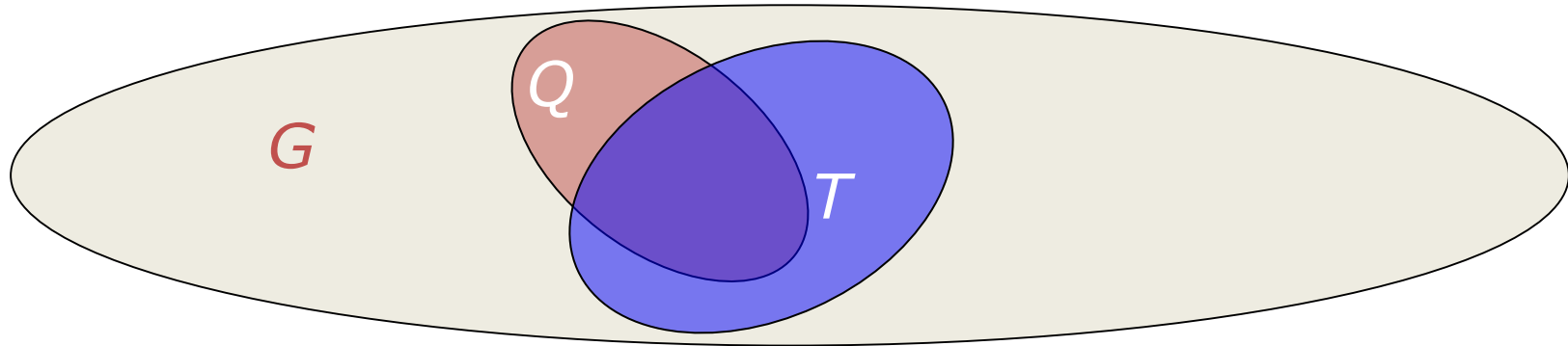
ensemble requête

$$Q \subseteq G$$



Quels sont les  
ensembles cibles  
qui lui sont similaires ?





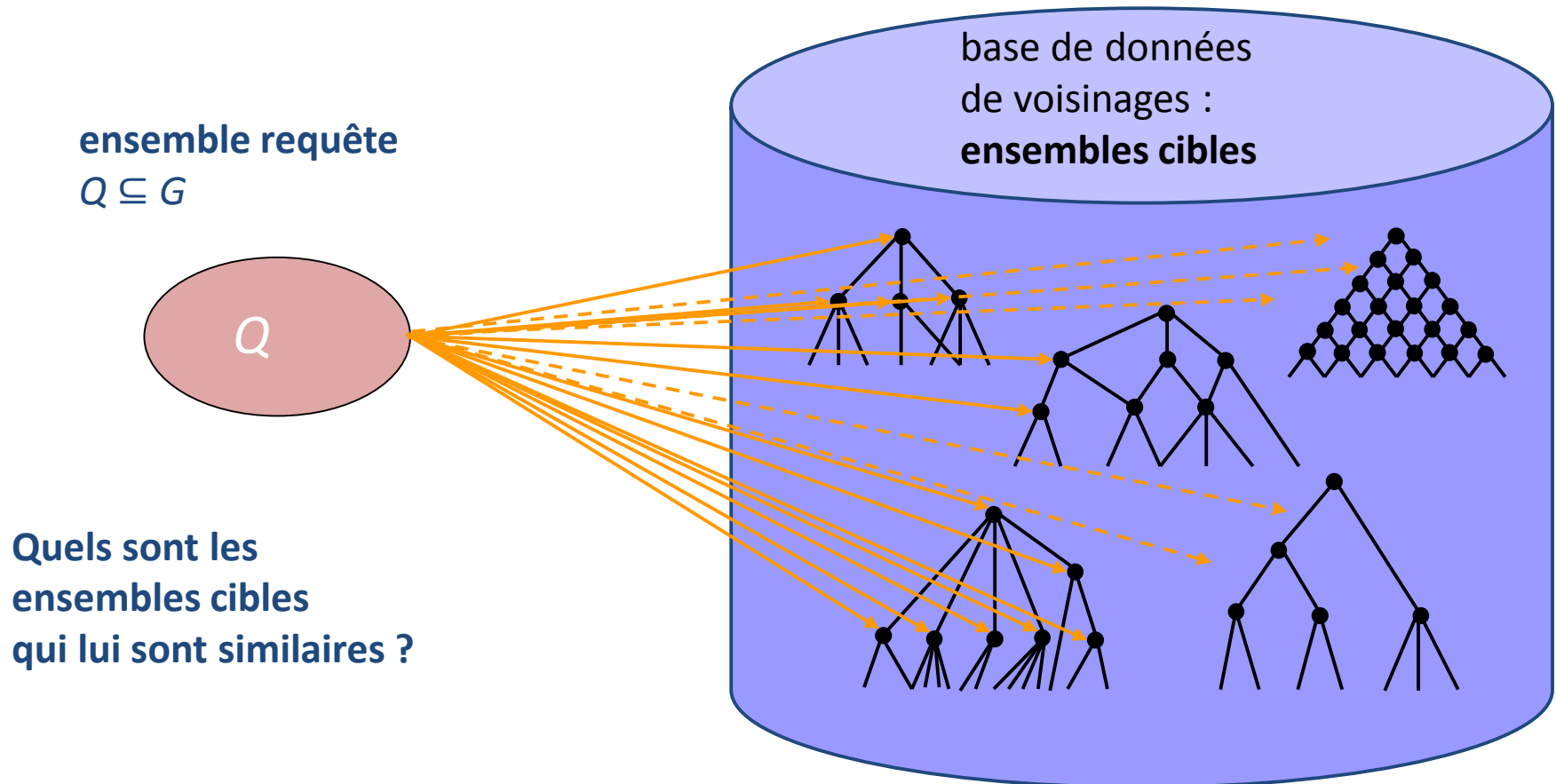
- Loi hypergéométrique : probabilité d'avoir au moins le nombre d'éléments communs observé entre 2 échantillons issus d'une même population

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

*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
    - ♦  $\chi^2$
    - ♦ ratio, pourcentage

- Recherche d'ensembles similaires



- 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

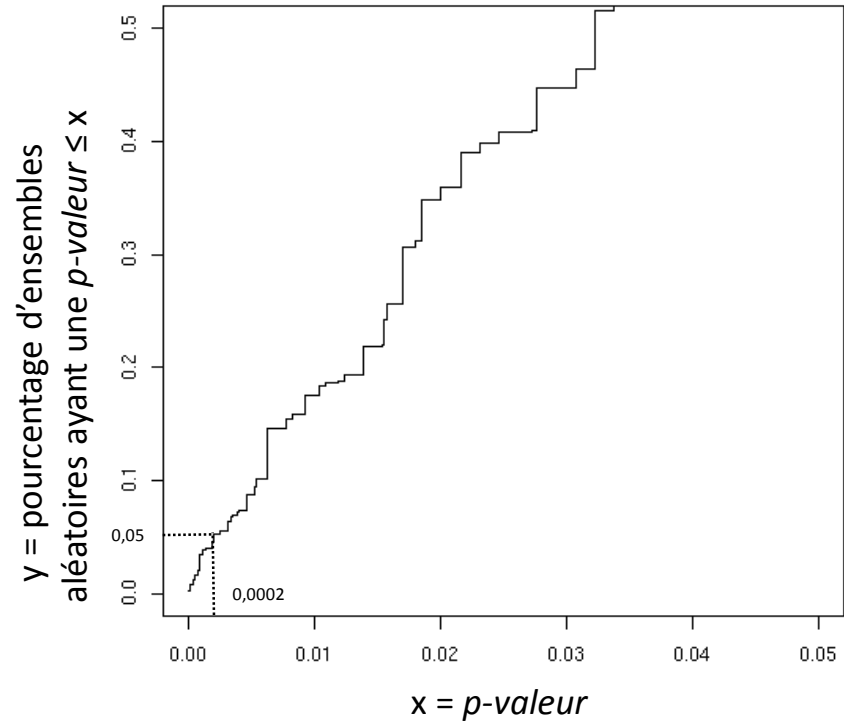
.

.

RandomSet\_n, minPi = Mn

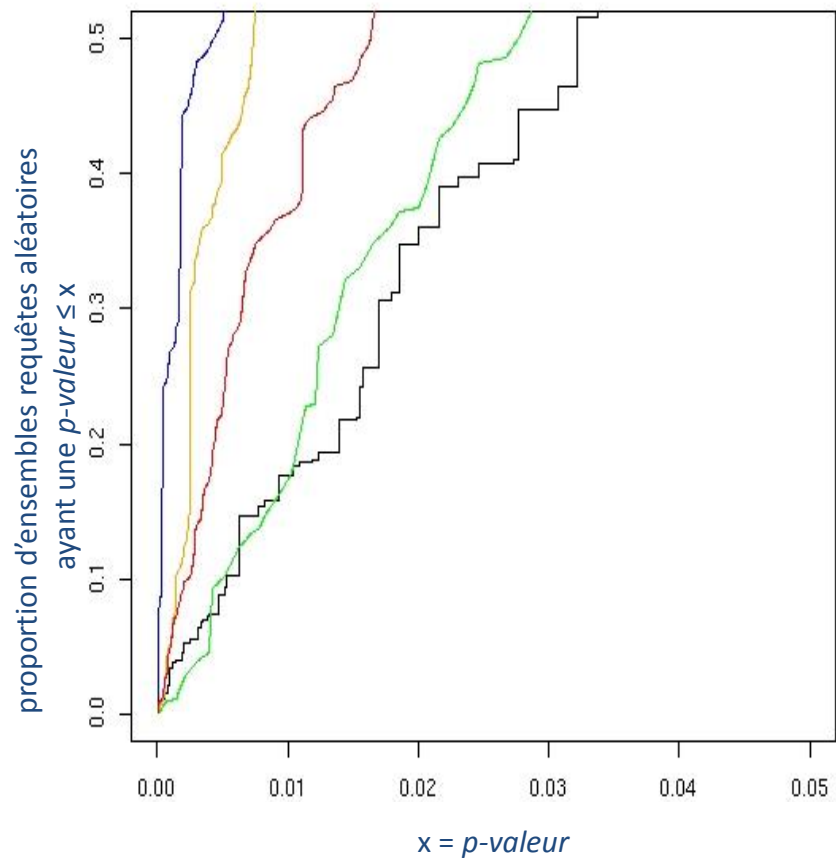
Étant donnée une p-valeur  $p$

Combien ont un meilleur score ?

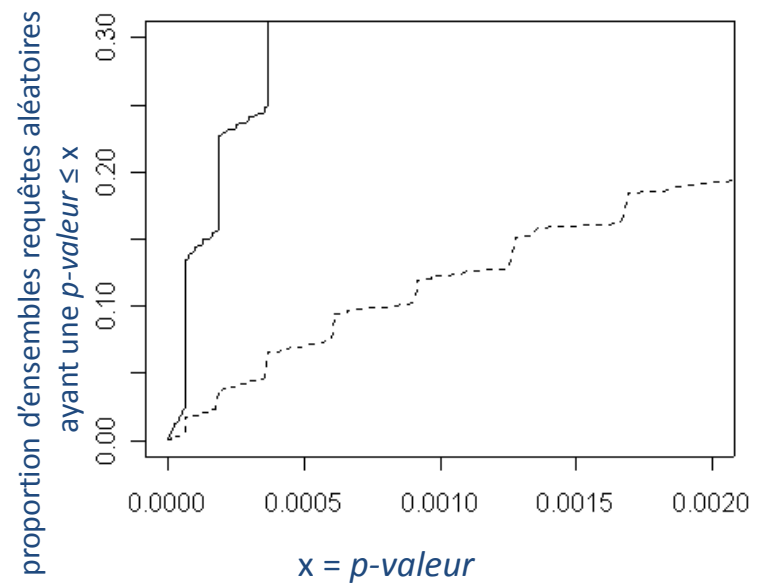


levure *Saccharomyces cerevisiae*  
n=500, q=9, g=5786, KEGG Pathways

# Significativité des p-valeurs obtenues



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

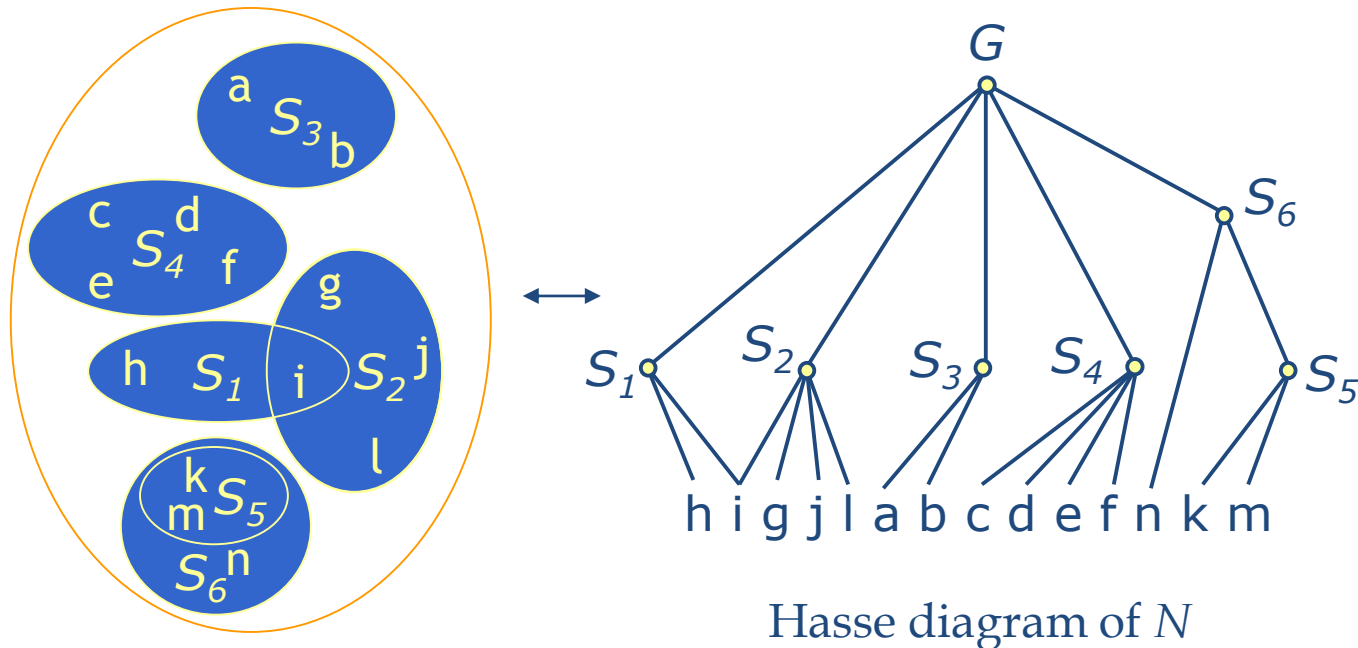


*Saccharomyces cerevisiae*

$n=500$ ,  $q=50$ ,  $g=5786$ ,

— GO molecular function,

- - - Ferea et al., 1999



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

a target set  $T$  is **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

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



## Pertinence definition

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

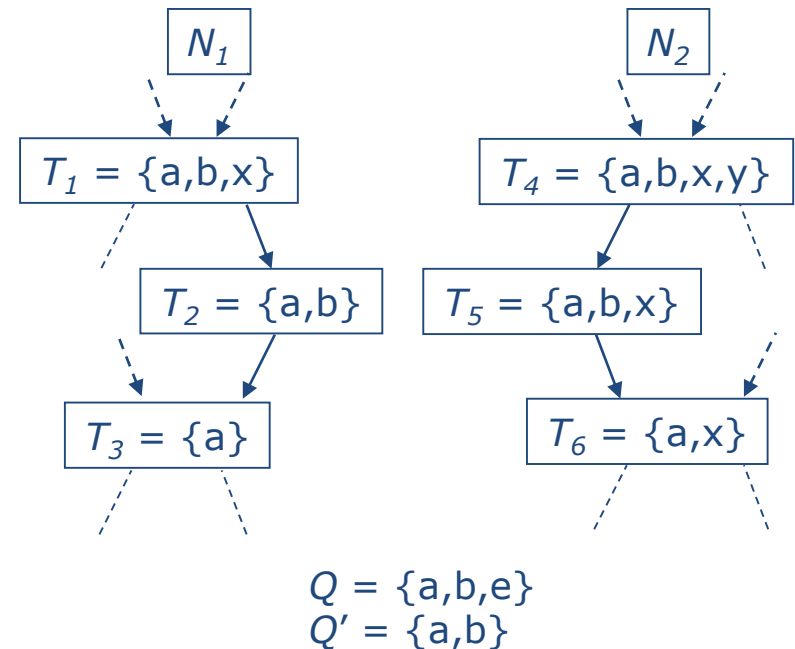
$$Q \cap T \neq \emptyset$$

and

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

and

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

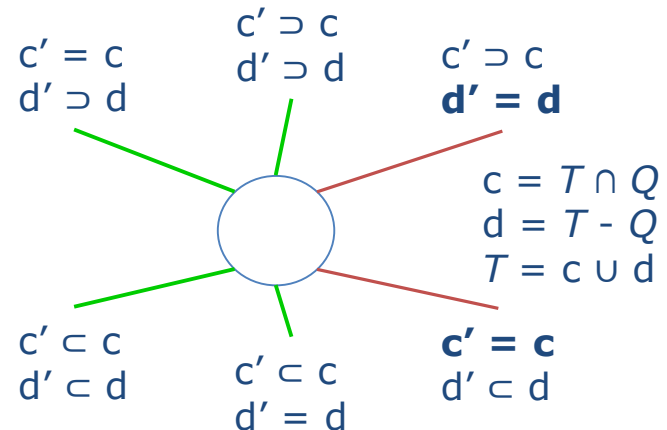


## Local decision

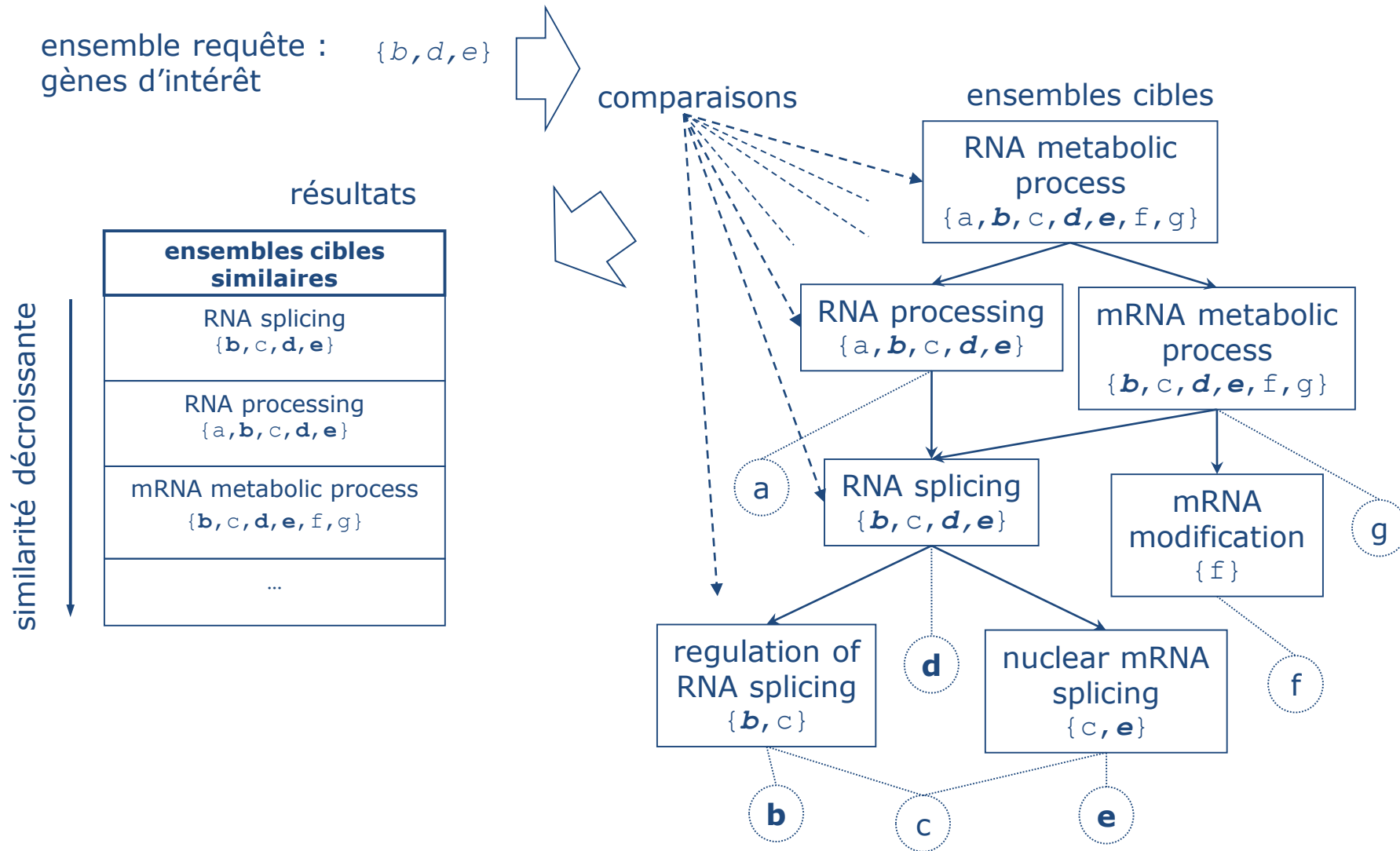
$$|c| > 0$$

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

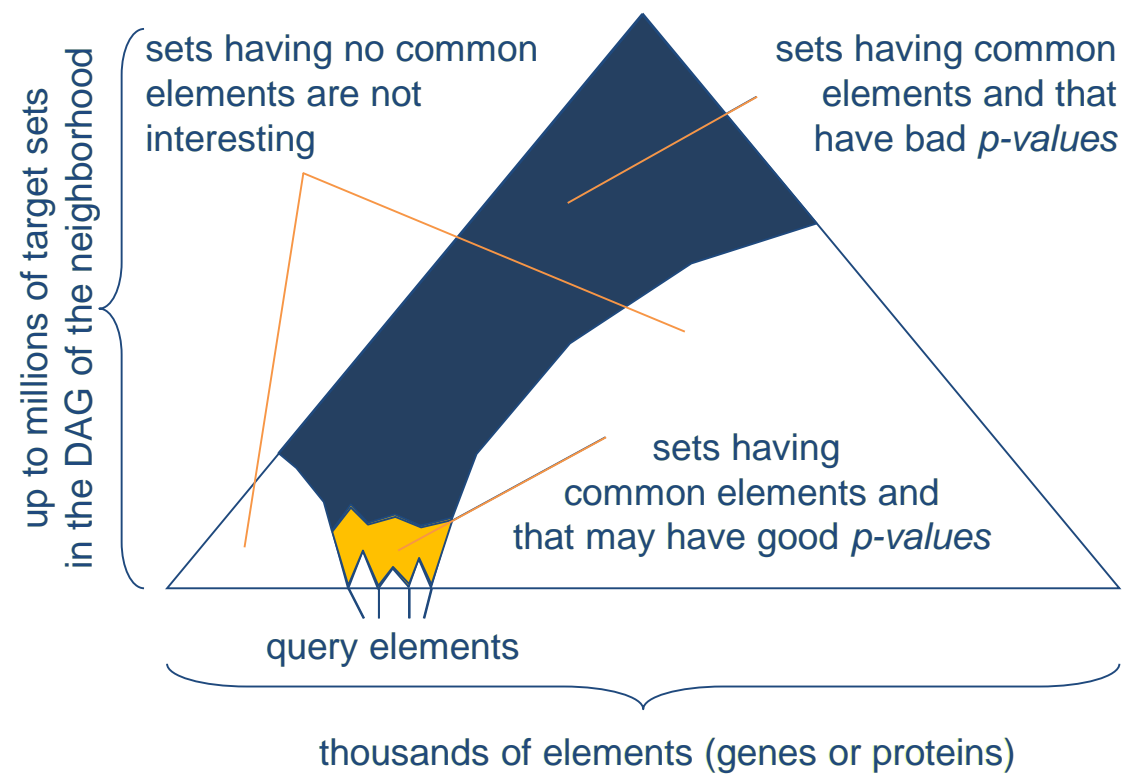
$$|c| > \max(\{c_{\text{children}}\})$$



- Pertinence des comparaisons & redondance des résultats

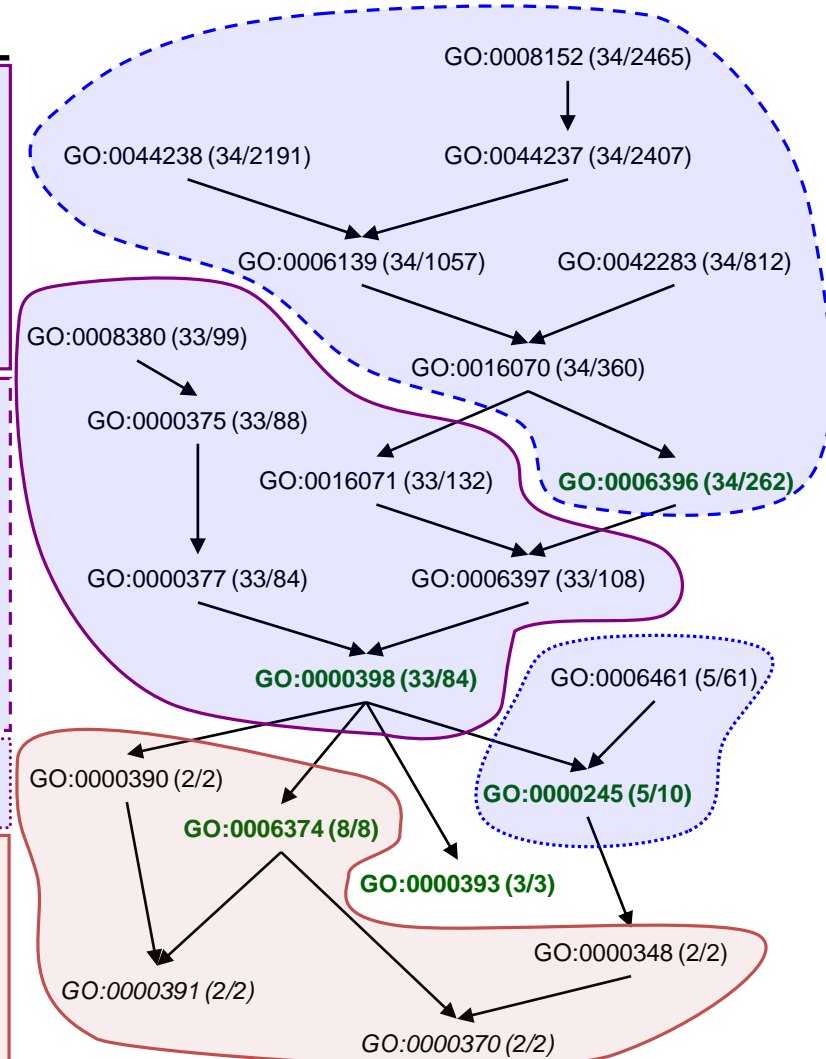


# A small portion of the DAG is searched

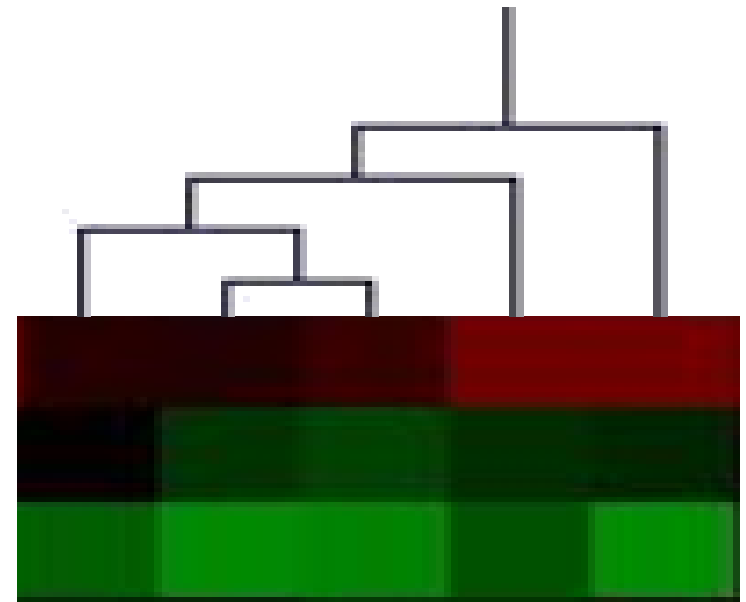


# Complex 440.30.10 mRNA splicing

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	U2-type spliceosome disassembly	2	2
GO:0000390	spliceosome disassembly	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



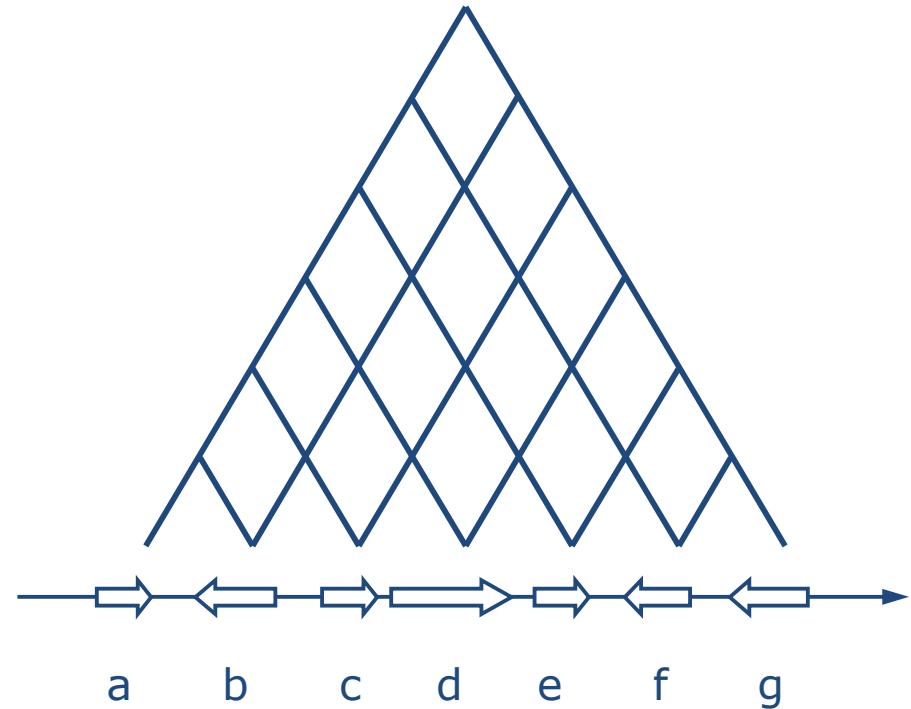
- each node has only 1 parent
- Algorithm
  - ◆ parses the input with a stack of stacks at the time it is loaded
  - ◆  $O(|G|)$  time



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

**tree**

- 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

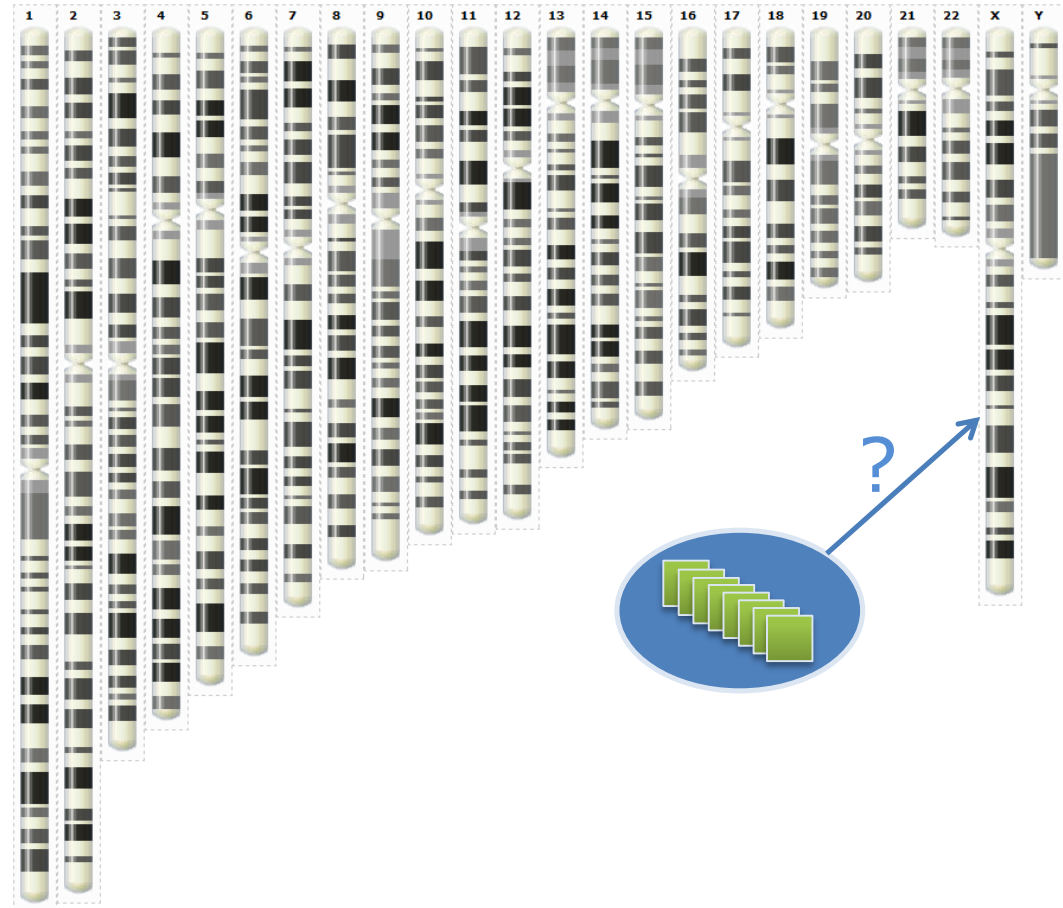


**implicit**

## Set of genes of interest

### Examples

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



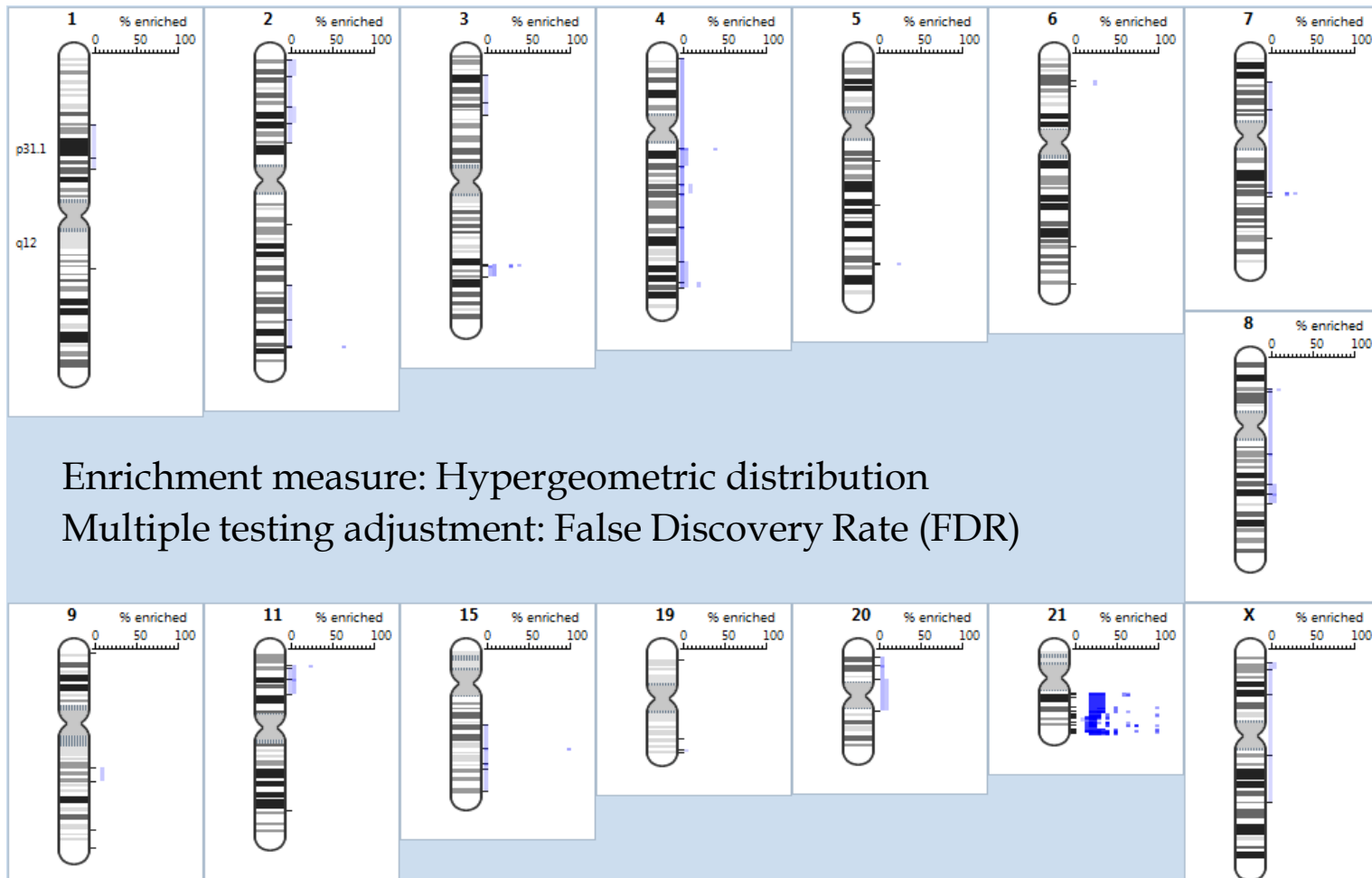
→ Question: Do those genes surprisingly cluster in the genome?

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

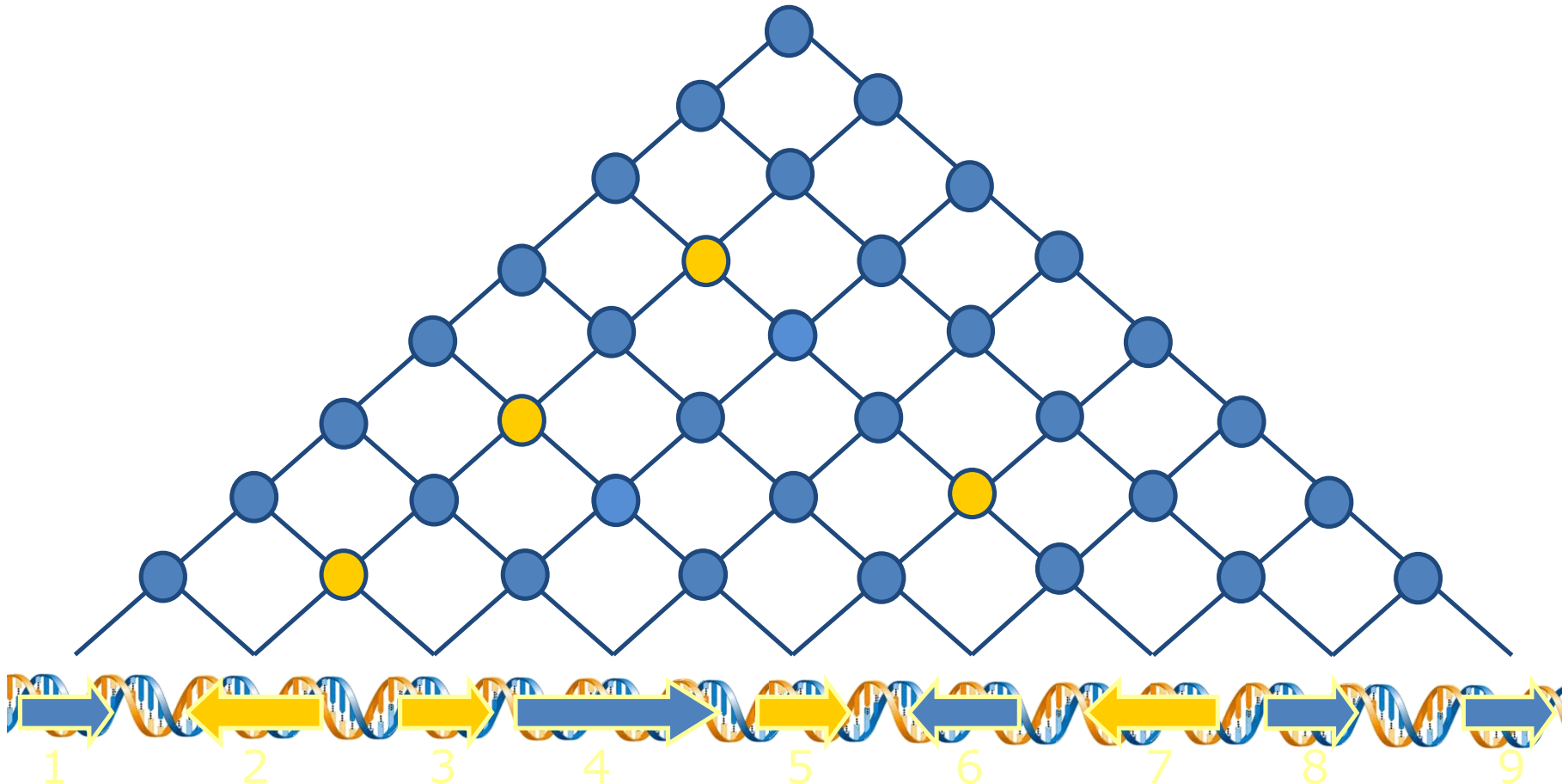
Multiple testing adjustment: False Discovery Rate (FDR)



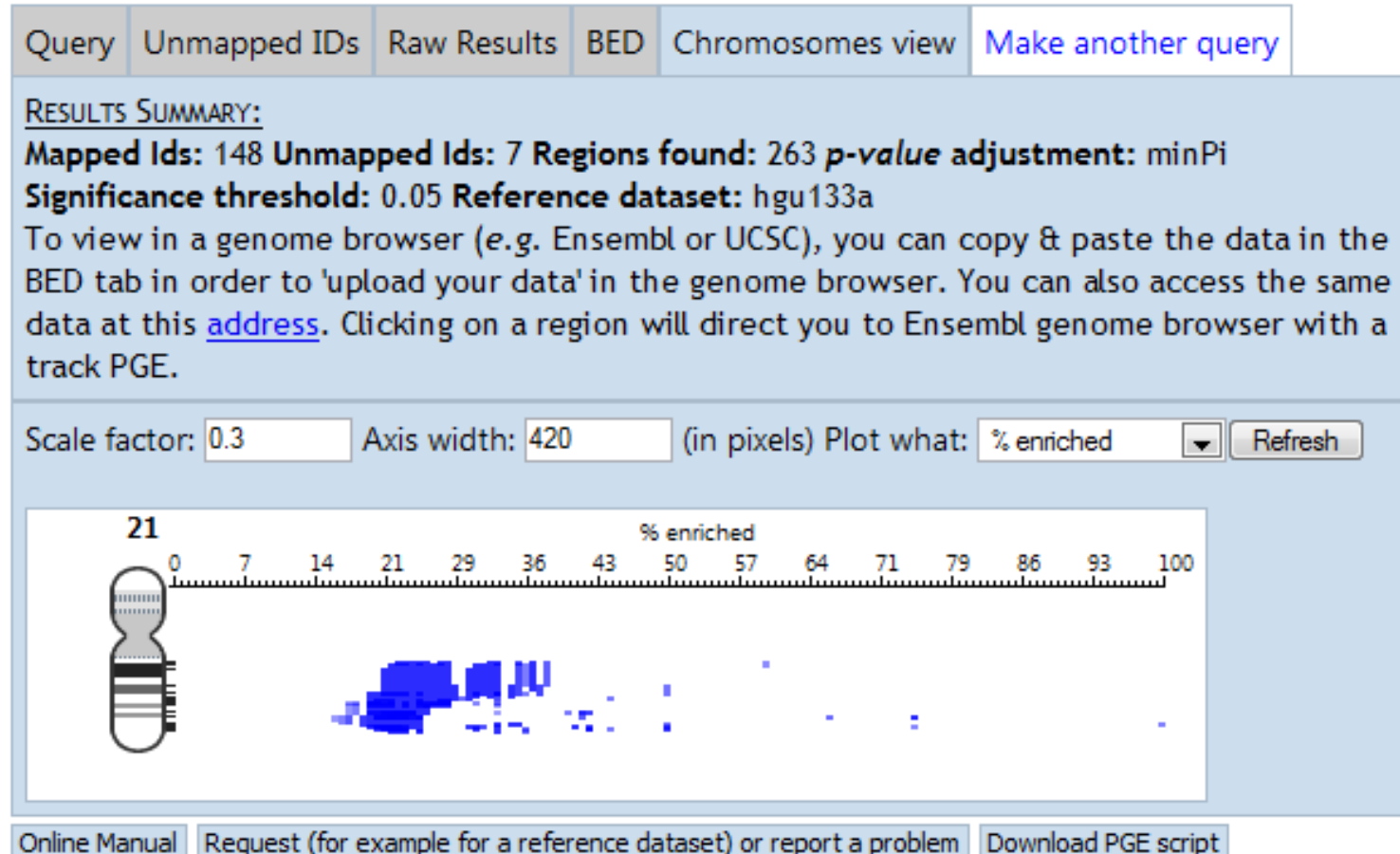
# Pertinent Regions

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

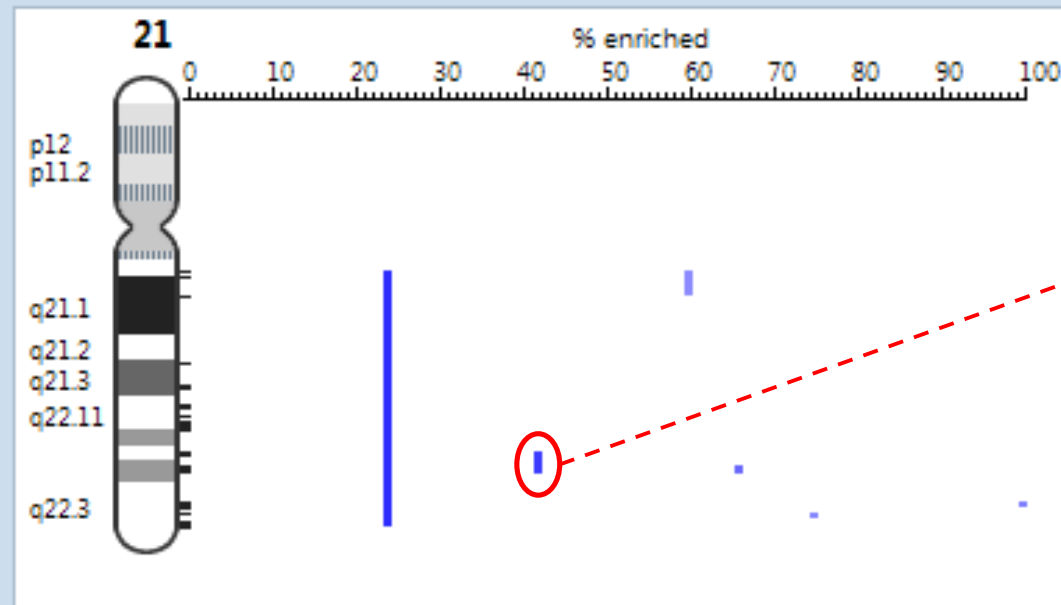
Query Unmapped IDs Raw Results BED Chromosomes view [Make another query](#)

RESULTS SUMMARY:

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

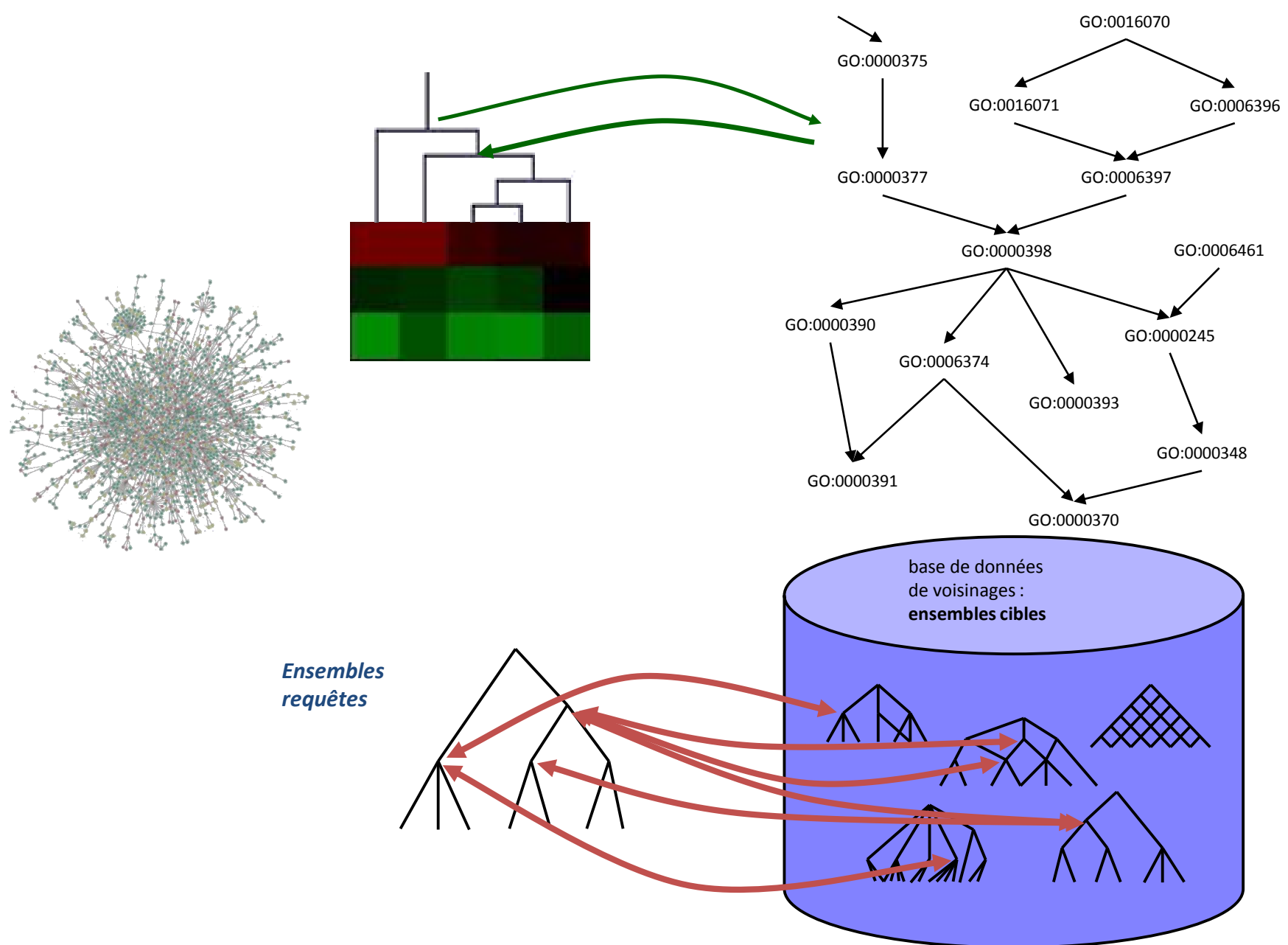
Scale factor:  Axis width:  (in pixels) Plot what:

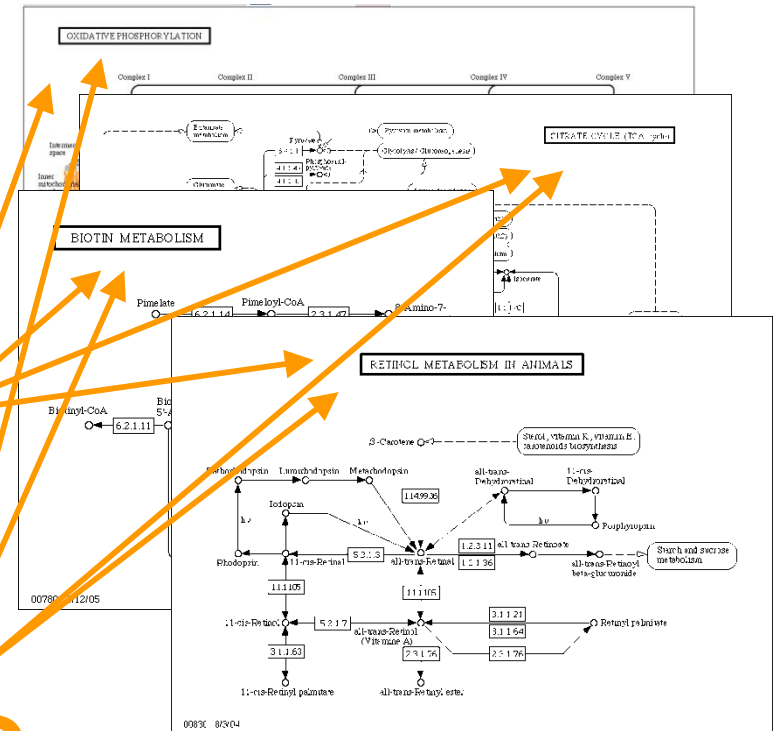
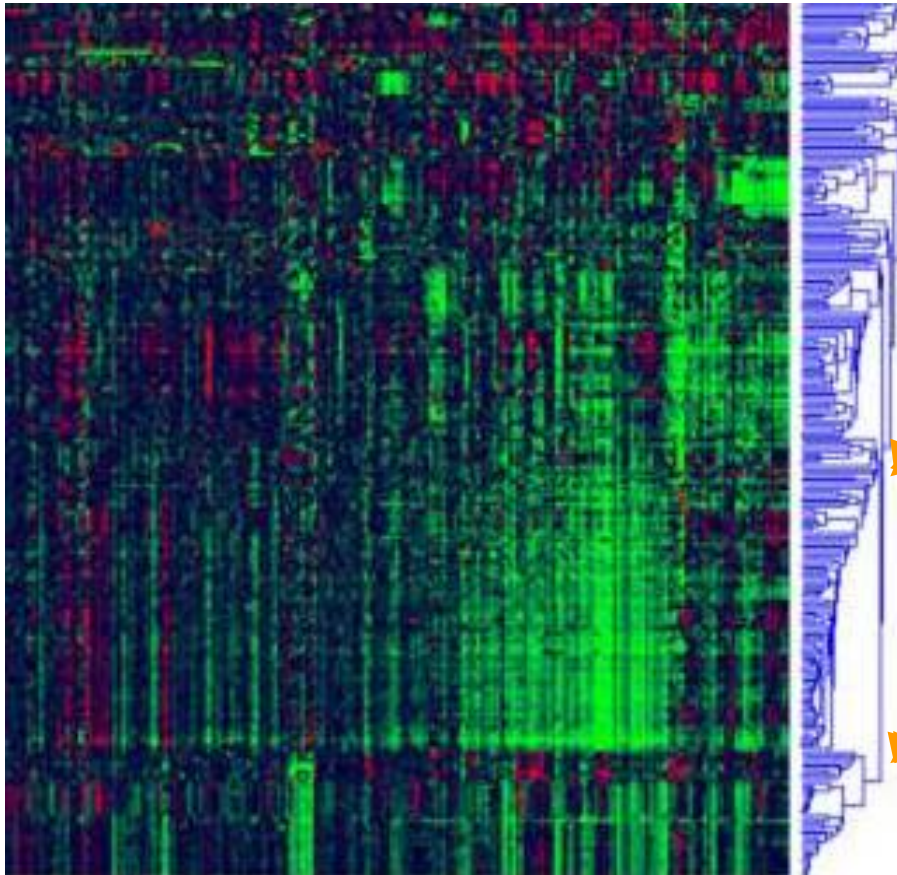
Refresh



*p*-value: 7.87E-10  
*p*-value<sub>adj</sub>: <0.002  
 Score: 91.039  
 Score<sub>adj</sub>: 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

# Challenges actuels

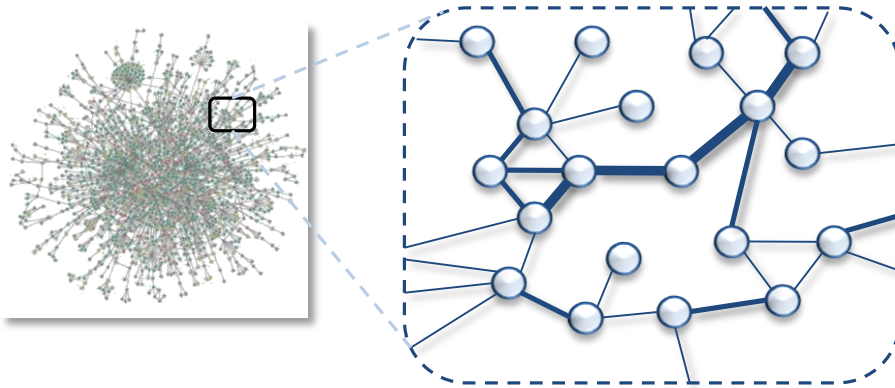




[Ferea et al., 1999]

[Kanehisa et Goto, 2000]

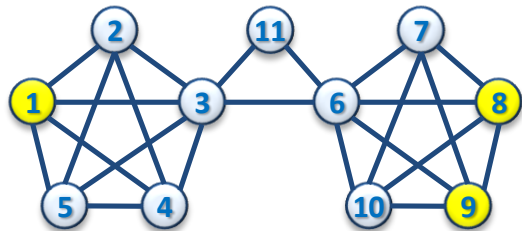
## Extraction de sous-graphe pertinent & visualisation



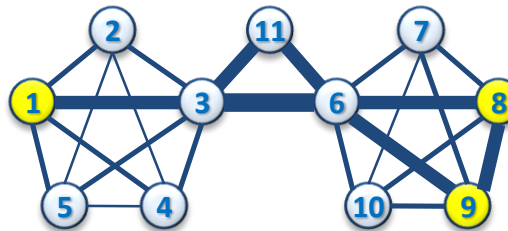
Idée :

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

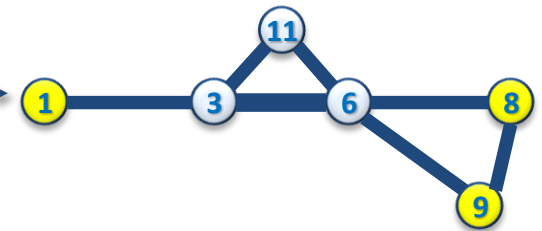
Gènes ayant la même annotation  
ex : interaction with host



Marche aléatoire :  
pondération des arcs

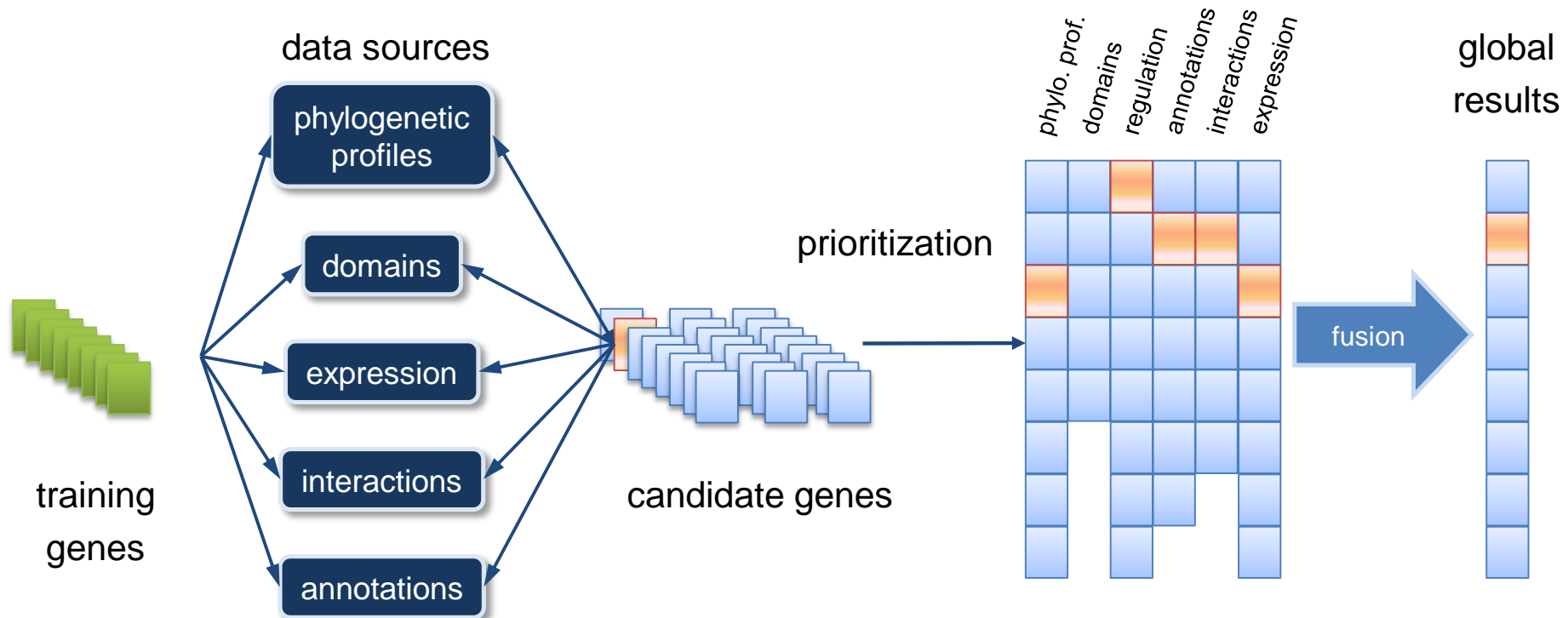


Surreprésentation :  
sous-graphe pertinent



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

- Observation: more and more post-genomic data available
- Paradox: more difficult to select the best candidates
- Goal: objective and comprehensive evaluation of candidates

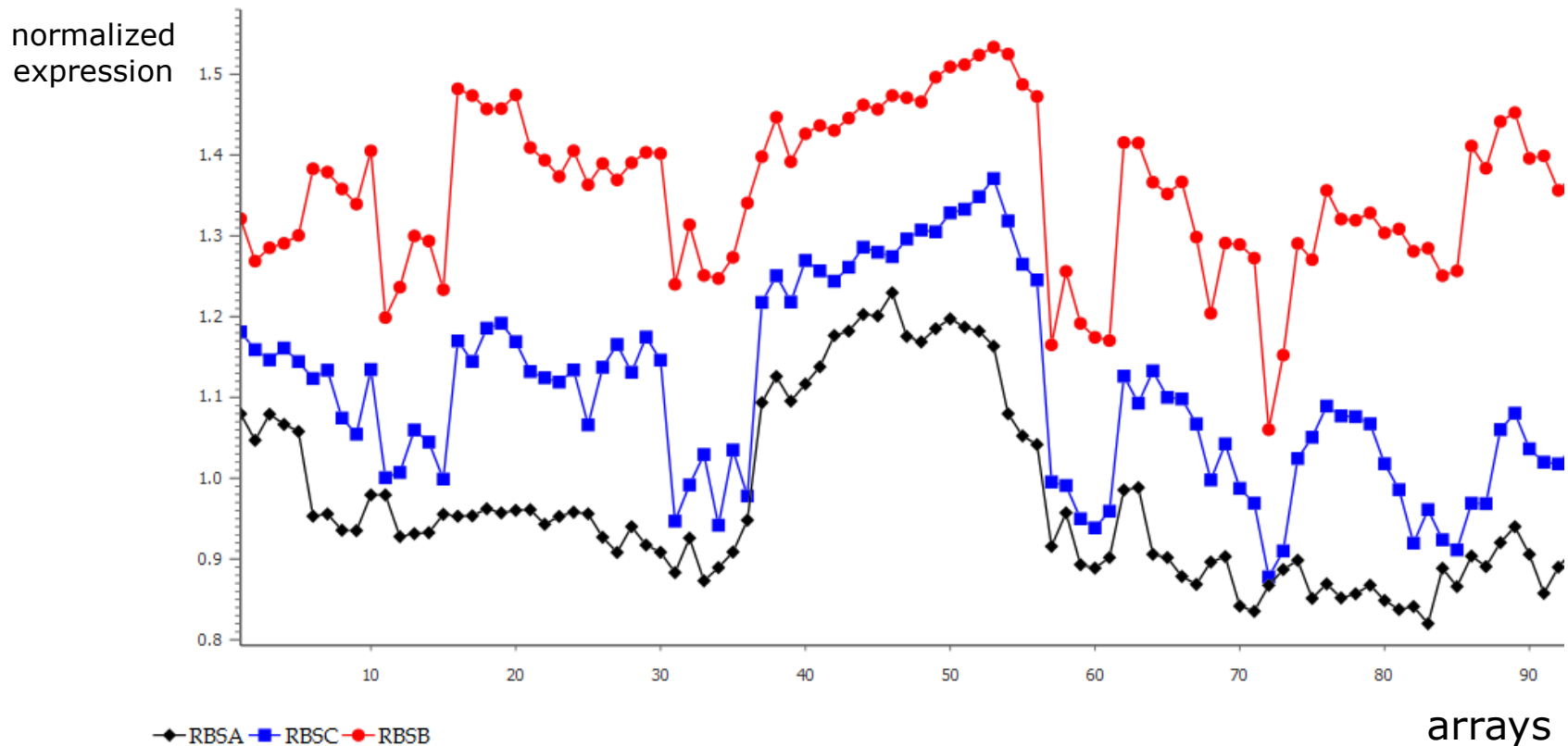


[Aerts *et al.* 2006]



# Gene expression

- a gene: set of expression values in various experimental conditions
- a pair of genes: dissimilarity index based on Pearson's correlation coefficient
- score : average dissimilarity

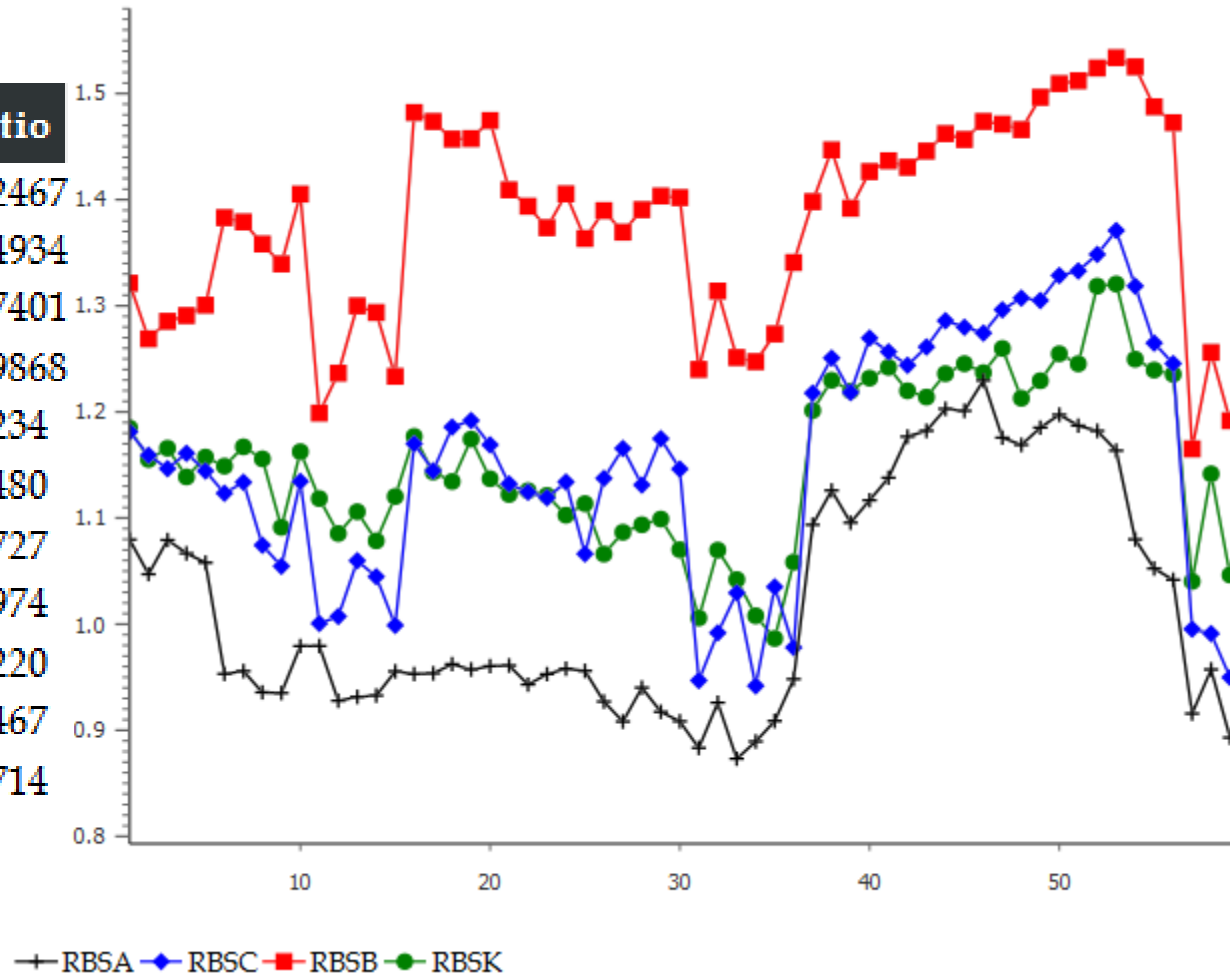




## Gene expression example

- training: rbsA, rbsB, rbsC in *E. coli* K-12

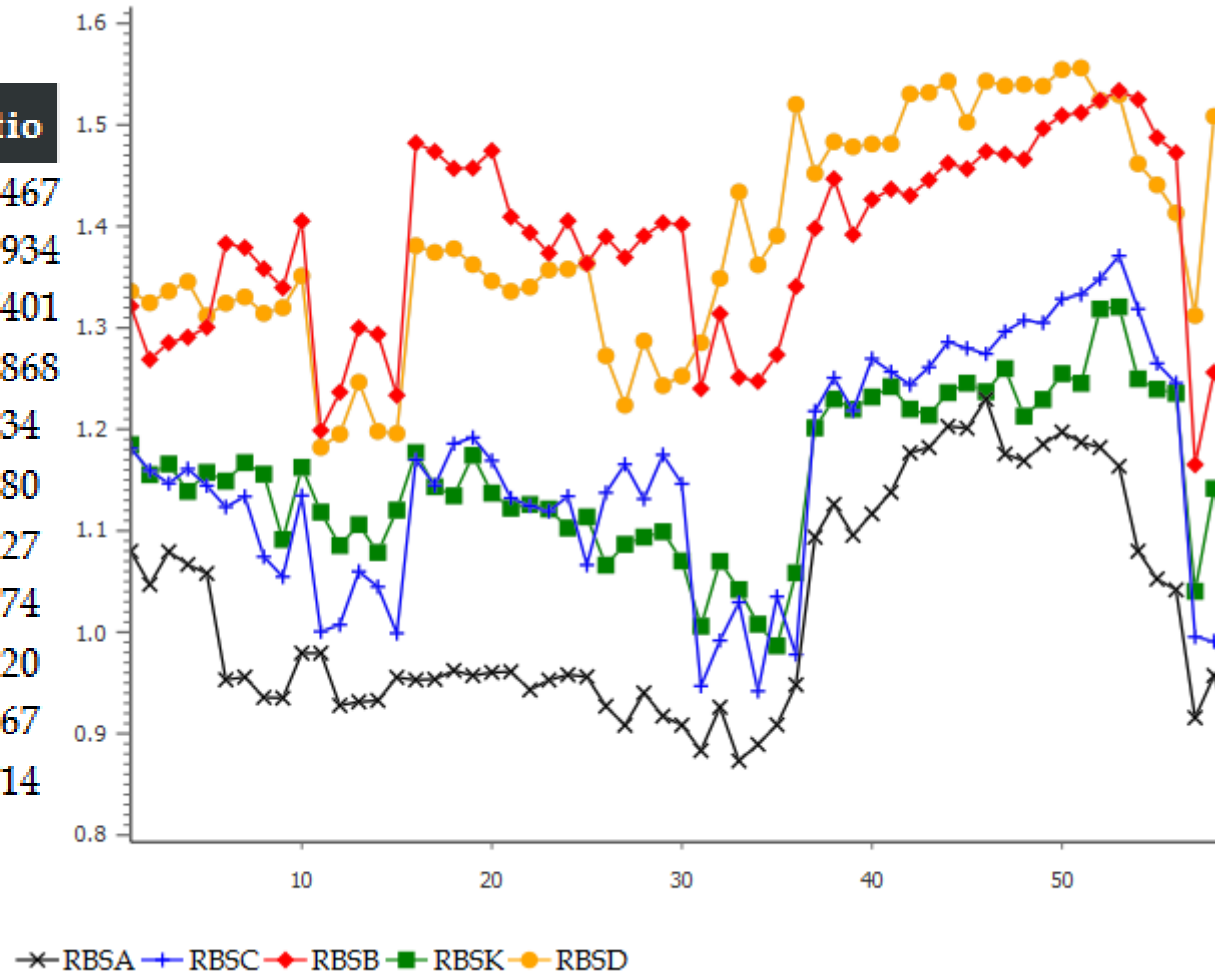
candidate	score	rank	rank ratio
<input type="checkbox"/> RBSK	0.1870	1	0.0002467
<input type="checkbox"/> RBSD	0.2695	2	0.0004934
<input type="checkbox"/> FDOI	0.3288	3	0.0007401
<input type="checkbox"/> MALE	0.3514	4	0.0009868
<input type="checkbox"/> MALK	0.3537	5	0.001234
<input type="checkbox"/> FDOG	0.3551	6	0.001480
<input type="checkbox"/> FDOH	0.3670	7	0.001727
<input type="checkbox"/> TREB	0.3679	8	0.001974
<input type="checkbox"/> NUPG	0.3841	9	0.002220
<input type="checkbox"/> LAMB	0.3850	10	0.002467
<input type="checkbox"/> MALF	0.3933	11	0.002714



## Gene expression example

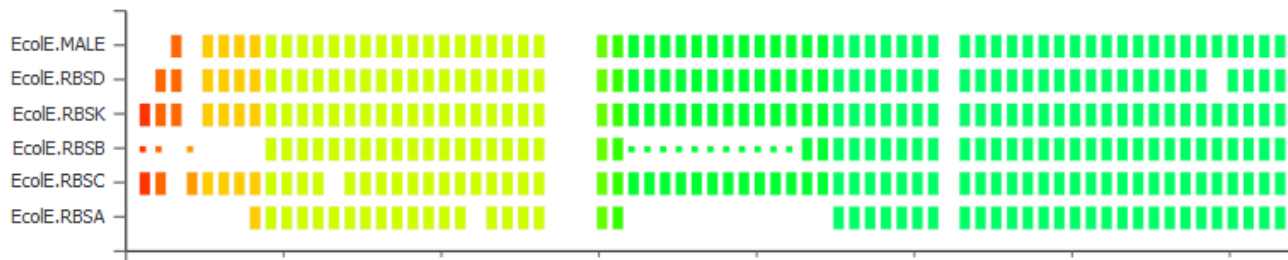
- training: rbsA, rbsB, rbsC

candidate	score	rank	rank ratio
<input type="checkbox"/> RBSK	0.1870	1	0.0002467
<input type="checkbox"/> RBSD	0.2695	2	0.0004934
<input type="checkbox"/> FDOI	0.3288	3	0.0007401
<input type="checkbox"/> MALE	0.3514	4	0.0009868
<input type="checkbox"/> MALK	0.3537	5	0.001234
<input type="checkbox"/> FDOG	0.3551	6	0.001480
<input type="checkbox"/> FDOH	0.3670	7	0.001727
<input type="checkbox"/> TREB	0.3679	8	0.001974
<input type="checkbox"/> NUPG	0.3841	9	0.002220
<input type="checkbox"/> LAMB	0.3850	10	0.002467
<input type="checkbox"/> MALF	0.3933	11	0.002714



- a gene: presence/absence of isorthologs in other genomes
- pair of genes: dissimilarity index based on the Jaccard index
- score: average dissimilarity

training: rbsA, rbsB, rbsC



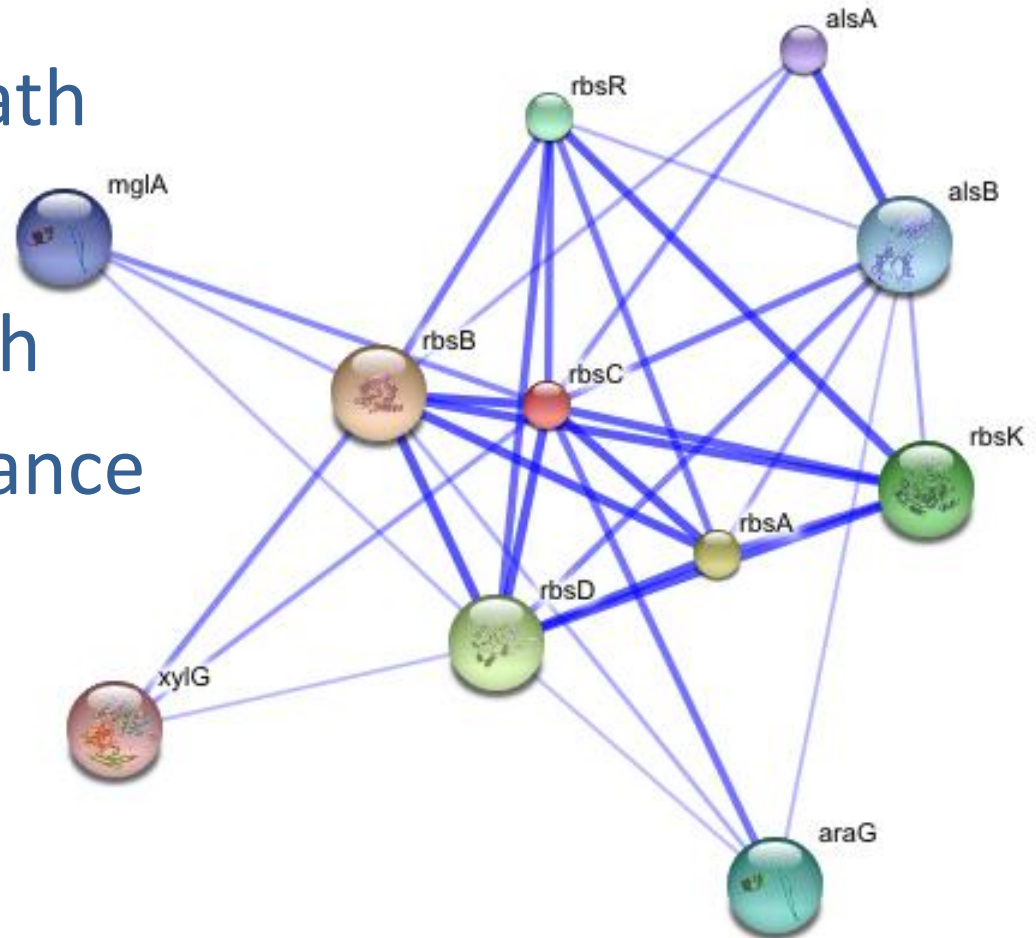
candidate	score	rank	rank ratio
<input type="checkbox"/> RBSD	0.6304	1	0.0002369
<input type="checkbox"/> MGSA	0.7274	2	0.0004739
<input type="checkbox"/> CDAR	0.7280	3	0.0007108
<input type="checkbox"/> CYTR	0.7285	4	0.0009478
<input type="checkbox"/> GLPT	0.7416	5	0.001185
<input type="checkbox"/> PTSG	0.7474	6	0.001422
<input type="checkbox"/> MALG	0.7475	7	0.001659
<input type="checkbox"/> RBSK	0.7486	8	0.001896
<input type="checkbox"/> CPDB	0.7533	9	0.002132
<input type="checkbox"/> POTB	0.7536	10	0.002369
<input type="checkbox"/> FLYI	0.7560	11	0.002606

■ Shigella dysenteriae Iso  
■ Shigella dysenteriae Ort  
■ Shigella boydii Iso  
■ Shigella boydii Ort  
■ Sodalis glossinidius Iso  
■ Sodalis glossinidius Ort  
■ Shigella Iso  
■ Salmonella Iso  
■ Photorhabdus Iso  
■ Pectobacterium Iso  
■ Yersinia Iso  
■ Yersinia Ort

- all pairs shortest path
- a pair of gene:  
shortest path length
- score: average distance

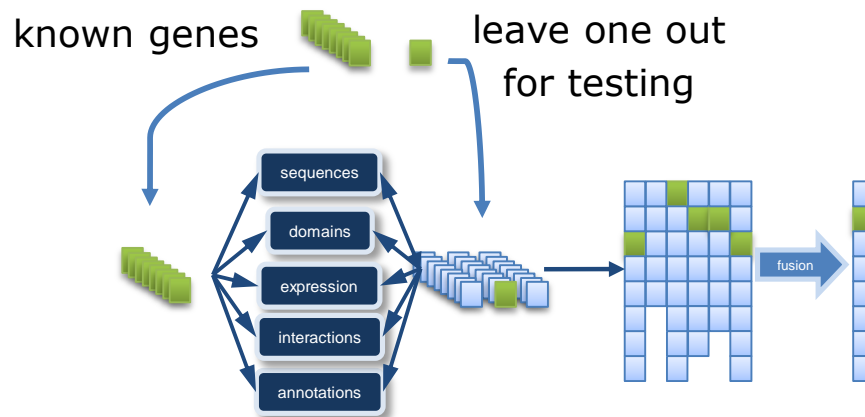
training: rbsA, rbsB, rbsC

candidate	score	rank	rank ratio
<input type="checkbox"/> RBSK	1.000	2	0.0005136
<input type="checkbox"/> RBSD	1.000	2	0.0005136
<input type="checkbox"/> RBSR	1.000	2	0.0005136
<input type="checkbox"/> ALSB	1.333	5	0.001284
<input type="checkbox"/> ALSC	1.333	5	0.001284
<input type="checkbox"/> YPHD	1.333	5	0.001284
<input type="checkbox"/> MGLC	1.667	10.5	0.002696
<input type="checkbox"/> XYLG	1.667	10.5	0.002696
<input type="checkbox"/> ALSA	1.667	10.5	0.002696
<input type="checkbox"/> YTFT	1.667	10.5	0.002696



from STRING  
<http://string-db.org>

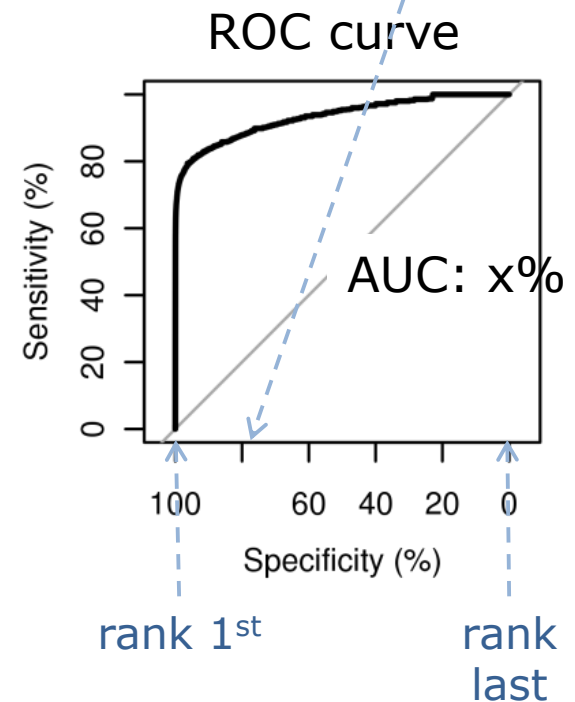
- Leave-one-out cross validation (LOOCV)



How well does it rank?  
*e.g.* rank ratio =  $2/8 = 0.25$

- for each manually curated ABC system

- ◆ perform LOOCV on each gene: rank ratio
- ◆ plot Receiver Operating Characteristic (ROC) curve and consider Area Under the Curve (AUC)



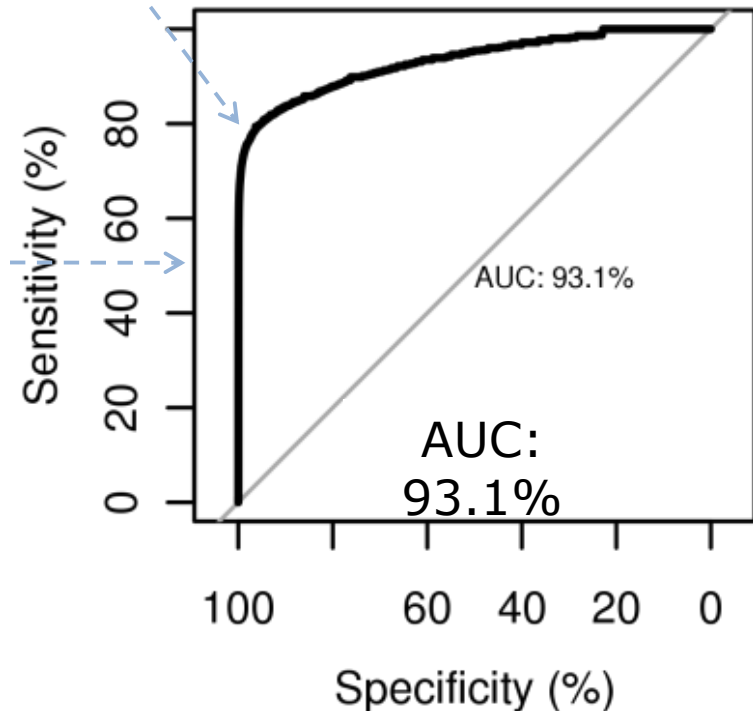
## Gold standard

- ABCdb, manually curated ABC systems:
  - ◆ 135 genomes
  - ◆ 14,450 genes
  - ◆ 4,586 ABC systems

80% of the left out genes rank in the top 5%

top 5%

**fusion , tests: 14450**



# Prioritization illustration

## Organism [Hide](#)

Organism	<b>Escherichia coli</b> ( strain K12 )
External Links	[ <a href="#">UNIPROT</a> ] [ <a href="#">NCBI</a> ]
Taxonomic Lineage	> <a href="#">Bacteria</a> > <a href="#">Proteobacteria</a> > <a href="#">Gammaproteobacteria</a> > <a href="#">Enterobacteriales</a> > <a href="#">Enterobacteriaceae</a> > <a href="#">Escherichia</a> > <a href="#">Escherichia coli</a> > <a href="#">EcolE</a>
Strain Name	K12
ABCdb identifier	<a href="#">EcolE</a>
Chromosomes	<a href="#">EcolE01</a>

## Assembly [Hide](#)

Assembly	NBD	MSD	SBP	Class
<a href="#">EcolE01.RBSB</a>	★ <a href="#">EcolE01.RBSA</a>	★ <a href="#">EcolE01.RBSC</a>	★ <a href="#">EcolE01.RBSB</a>	A_1a

## Proteins [Hide](#)

Protein	Domain	Subfamily	TCdb
★ <a href="#">EcolE01.RBSB</a>	SBP	S_1aa	3.A.1.2.1 Ribose porter (RbsC has 10 TMSs with N- and C-termini in the cytoplasm (Stewart and Hermodson, 2003))
★ <a href="#">EcolE01.RBSC</a>	MSD	M_1aa	3.A.1.2.1 Ribose porter (RbsC has 10 TMSs with N- and C-termini in the cytoplasm (Stewart and Hermodson, 2003))
★ <a href="#">EcolE01.RBSA</a>	NBD-NBD	N_1aN&N_1aC	3.A.1.2.1 Ribose porter (RbsC has 10 TMSs with N- and C-termini in the cytoplasm (Stewart and Hermodson, 2003))

from ABCdb

<http://www-abcdb.biotoul.fr>

# Prioritization illustration

Prioritization **Hide**

**Hide**

Run prioritization.

Show  entries

Search:

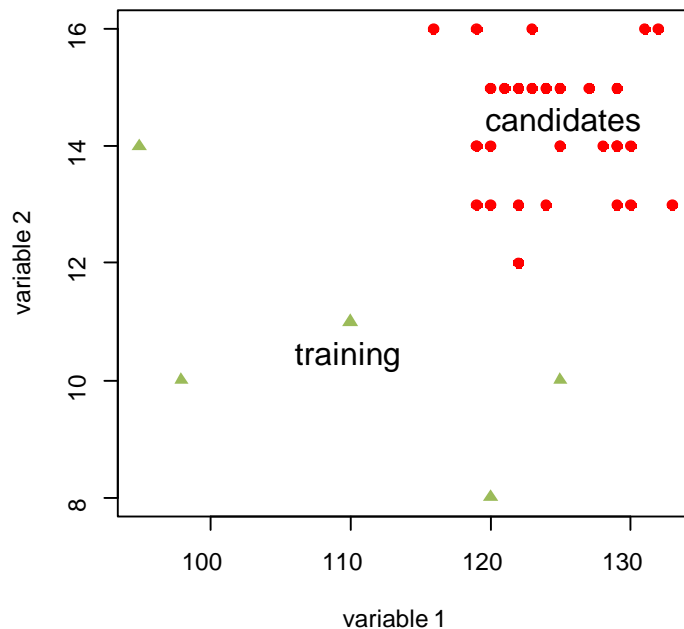
rank	Global results	pathways (fusion)	string (fusion)	transcriptome (fusion)	phylogenetic_profiles EcolE	go (fusion)	interactome EcolE
1	<b>RBSD (1)</b> S: 0, RR: 0	D-ribose pyranase					
2	<b>RBSK (2)</b> S: 0, RR: 0	Ribokinase					
3	<b>MALE (3)</b> S: 0, RR: 0.001	SBP of maltose/maltodextrin/maltoogisaccharide ABC transporter					
4	<b>DEOC (4)</b> S: 0, RR: 0.001	Deoxyribose-phosphate aldolase					
5	<b>RBSR (5)</b> S: 0.001, RR: 0.001	Ribose operon repressor					
6	<b>UDP (6)</b> S: 0.001, RR: 0.001	Uridine phosphorylase					
7	<b>MGLA (7)</b> S: 0.001, RR: 0.002	NBD of galactose/glucose (methyl galactoside) ABC transporter (same subfamily)					
8	<b>MUKF (8)</b> S: 0.002, RR: 0.002	Chromosome partition protein mukF					
9	<b>GAPA (9)</b> S: 0.002, RR: 0.002	<i>CITT (2056)</i> S: 1, RR: 1	<b>XYLF (9)</b> S: 0, RR: 0.002	<b>UCPA (9)</b> S: 0.003, RR: 0.002	<b>CPDB (9)</b> S: 0.753, RR: 0.002	<b>RPLN (16)</b> S: 0.004, RR: 0.004	<b>UDP (34)</b> S: 1.5, RR: 0.009



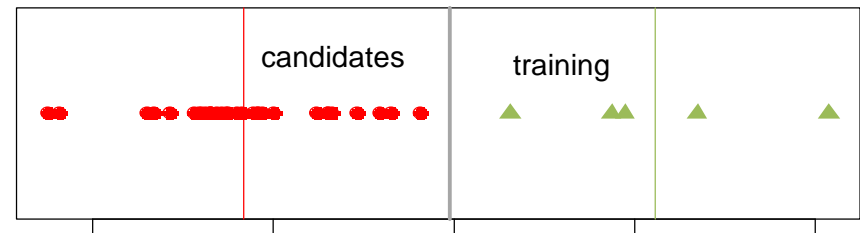
# Weighted fusion through linear discriminant analysis

## • Principle

- ♦ prioritize the candidate genes and including the training genes
- ♦ consider each data source as a measure for classification with classes: training/candidate
- ♦ perform discriminant analysis to weigh and separate training genes from background (candidates)

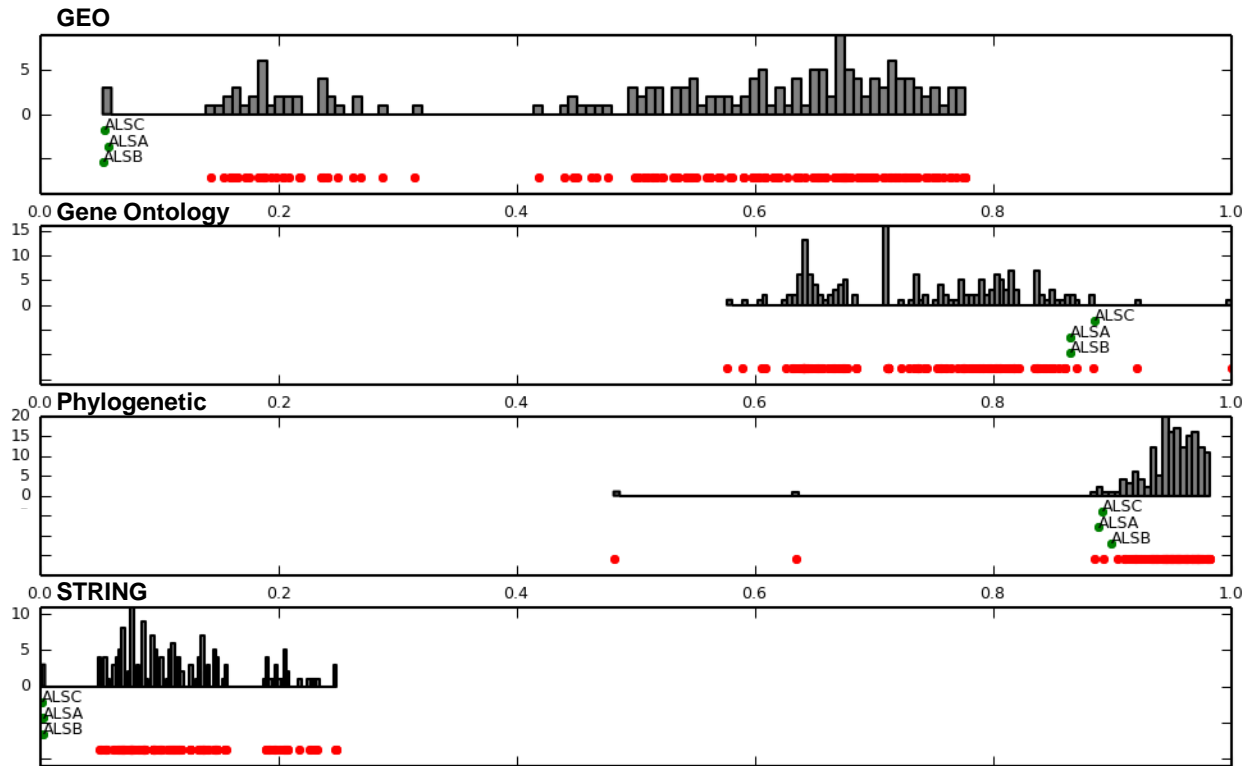


dimension	weight
variable 1	-0.1307346
variable 2	-0.7031850



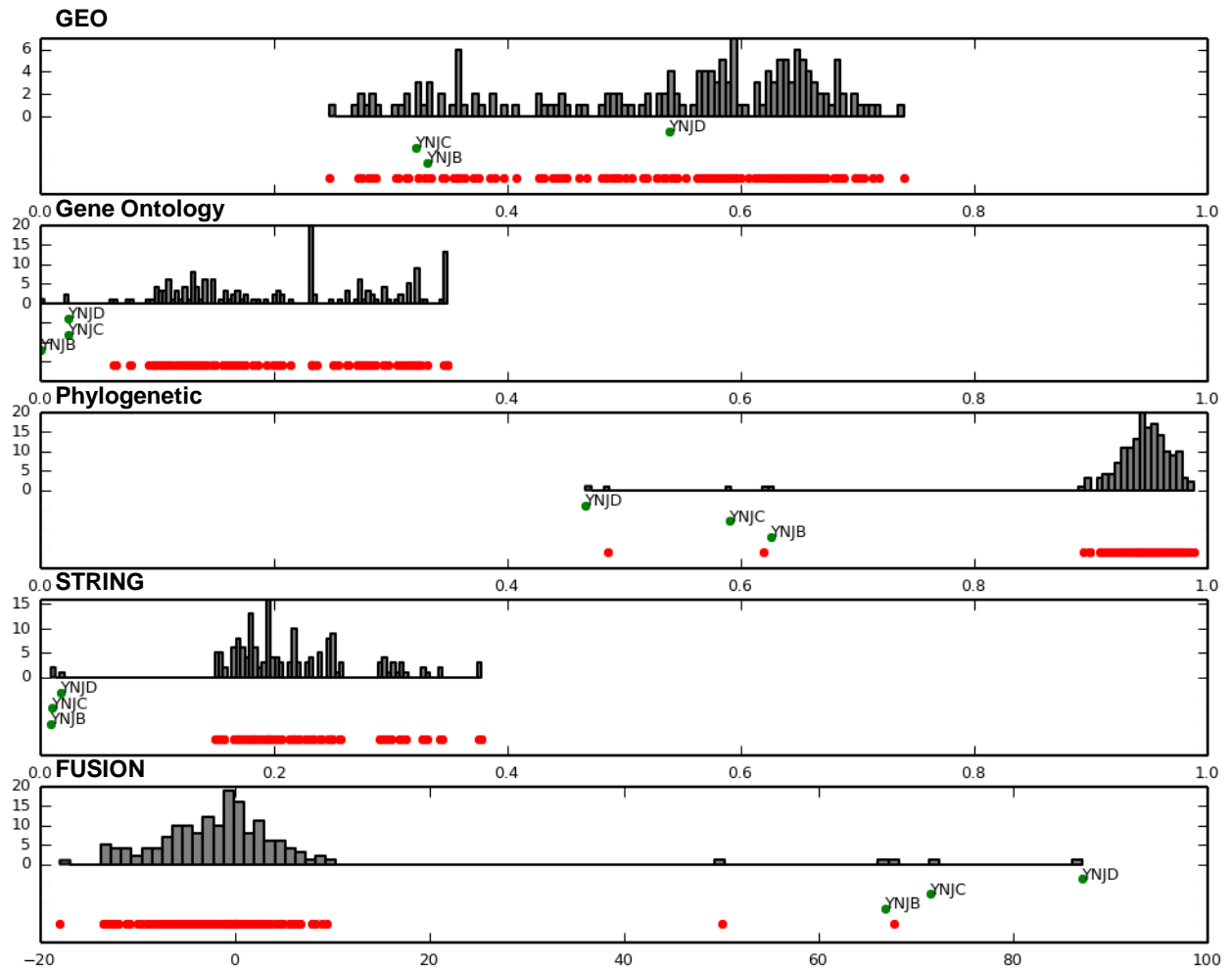
# Application to *E. coli* *alsA* system: *alsA*, *alsB*, *alsC*

Data source	Weight
Expression (GEO)	3.5
Annotations (Gene Ontology)	-4.7
Phylogenetic	4.0
Interactions (STRING)	12.3

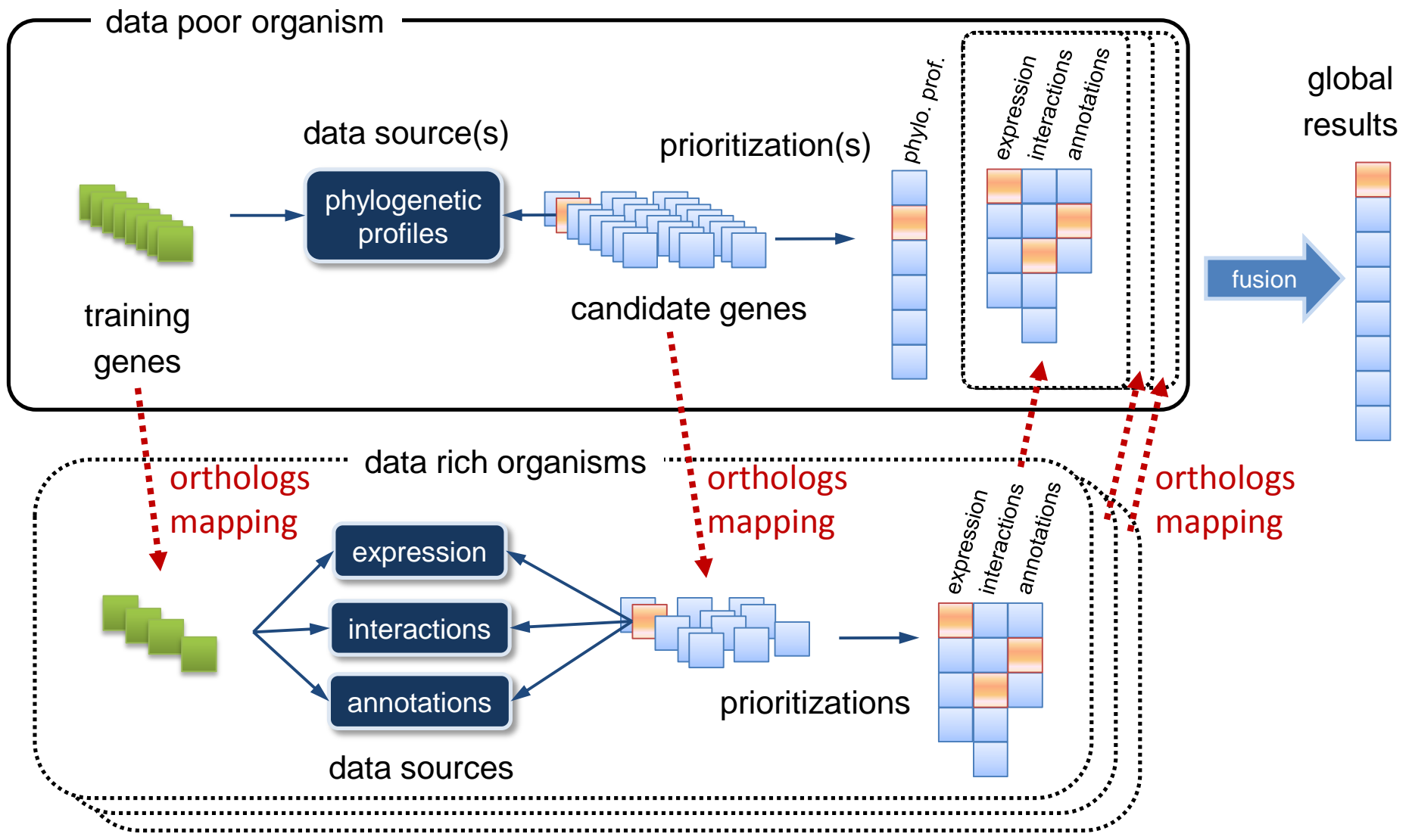


# Application to *E. coli* ynjD system: ynjB, ynjC, ynjD

Data source	Weight
Expression (geo)	1.3
Annotations (go)	3.4
Phylogenetic (microsynteny)	17.4
Interactions (string)	6.6



# Extension of the method to data poor organisms

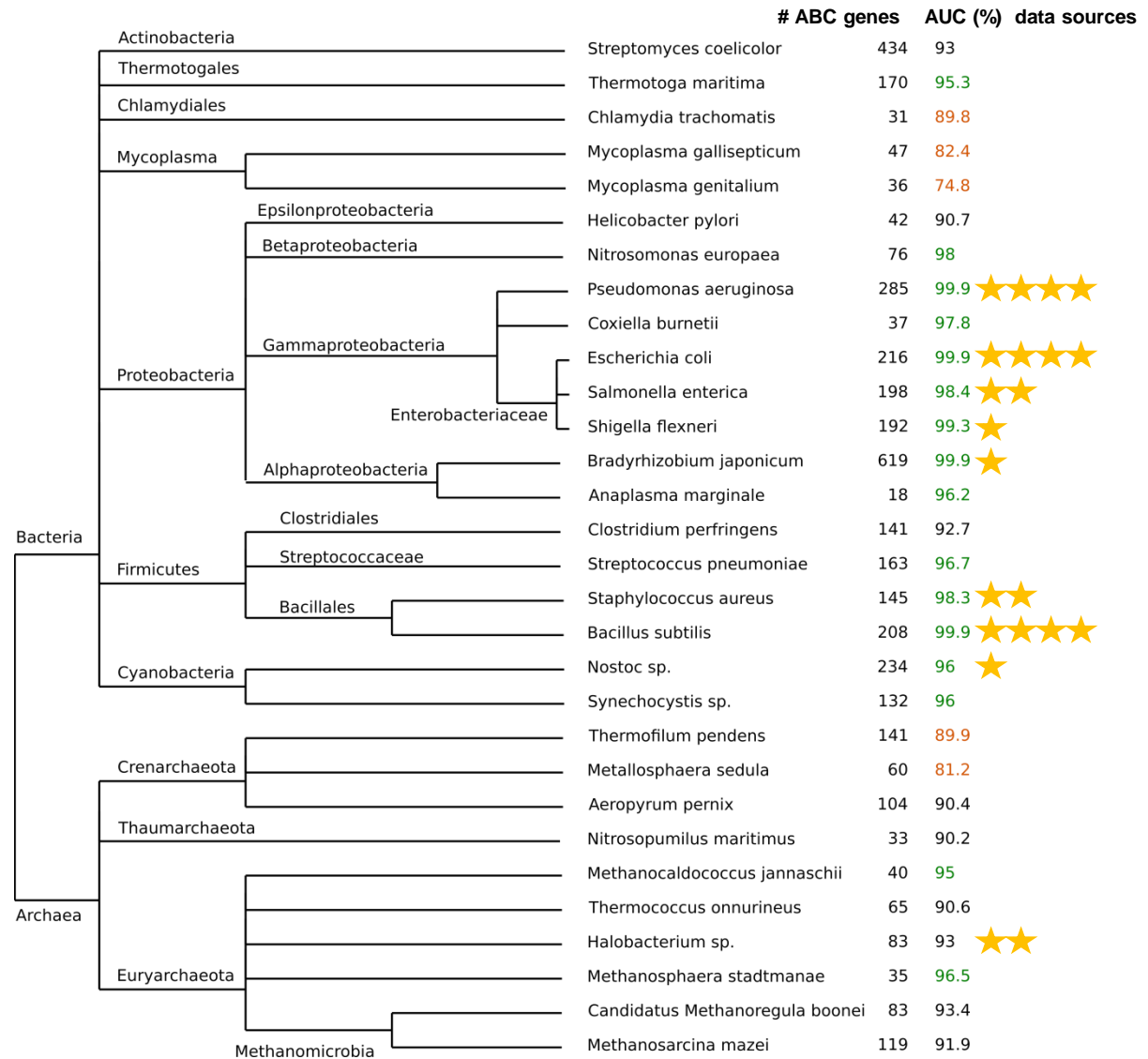
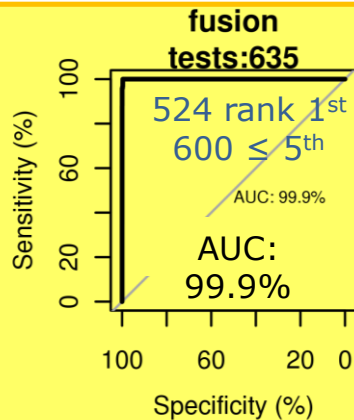
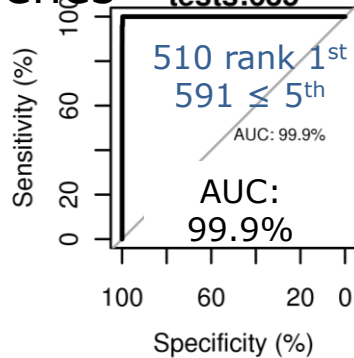


# Performances: using other organisms data through orthology

Organisms:

*B. subtilis*, *E. coli*, *P. aeruginosa*

192 ABC systems, 635 genes



# Performances: using other organisms data through orthology

			# ABC genes	AUC (%)	data sources		
Bacteria	Actinobacteria						
		Thermotogales	Streptomyces coelicolor	434	93		
			Thermotoga maritima	170	95.3		
		Chlamydiales	Chlamydia trachomatis	31	89.8		
		Mycoplasma	Mycoplasma gallisepticum	47	82.4		
			Mycoplasma genitalium	36	74.8		
		Epsilonproteobacteria	Helicobacter pylori	42	90.7		
		Betaproteobacteria	Nitrosomonas europaea	76	98		
		Gammaproteobacteria	Pseudomonas aeruginosa	285	99.9	★★★★	
			Coxiella burnetii	37	97.8		
			Enterobacteriaceae	Escherichia coli	216	99.9	★★★★
				Salmonella enterica	198	98.4	★★
			Shigella flexneri	192	99.3	★	
		Alphaproteobacteria	Bradyrhizobium japonicum	619	99.9	★	
			Anaplasma marginale	18	96.2		
		Clostridiales	Clostridium perfringens	141	92.7		
		Firmicutes	Streptococcaceae	Streptococcus pneumoniae	163	96.7	
	Bacillales		Staphylococcus aureus	145	98.3	★★	
			Bacillus subtilis	208	99.9	★★★★	
	Cyanobacteria	Nostoc sp.	234	96	★		
			Synechocystis sp.	132	96		
		Thermophilum pendens	141	89.9			