Gestion de données non structurées - Applications Post-Génomiques

Généralités & Approches

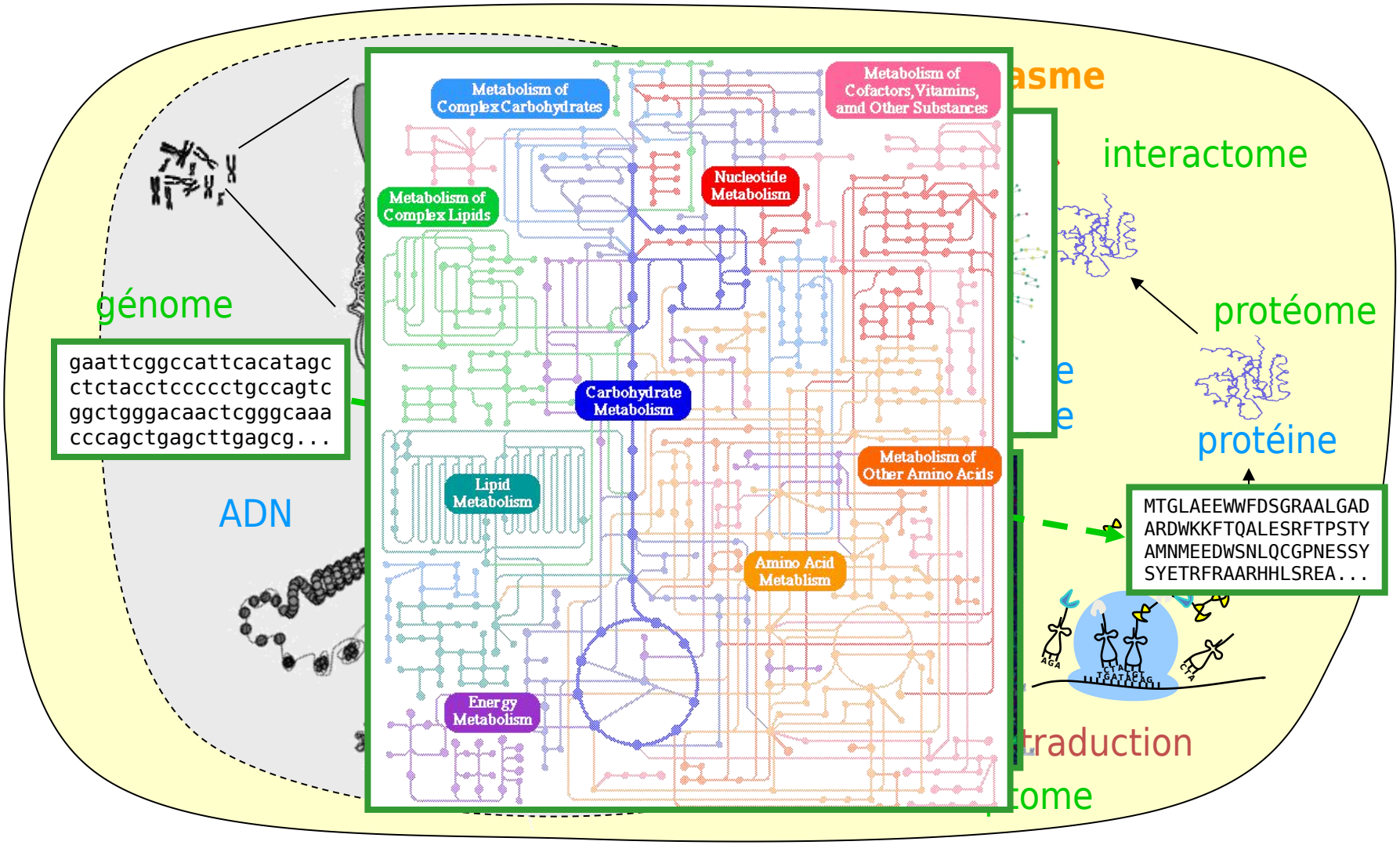
Master 2

Bioinformatique et Biologie des Systèmes

- Pourquoi ?

- Qu’est-ce que l’intégration ?
  - Interconnexion
  - Fusion
  - Médiation
  - Modélisation
  - Confrontation
  - Recoupement

Biologie, mesures, données et connaissances

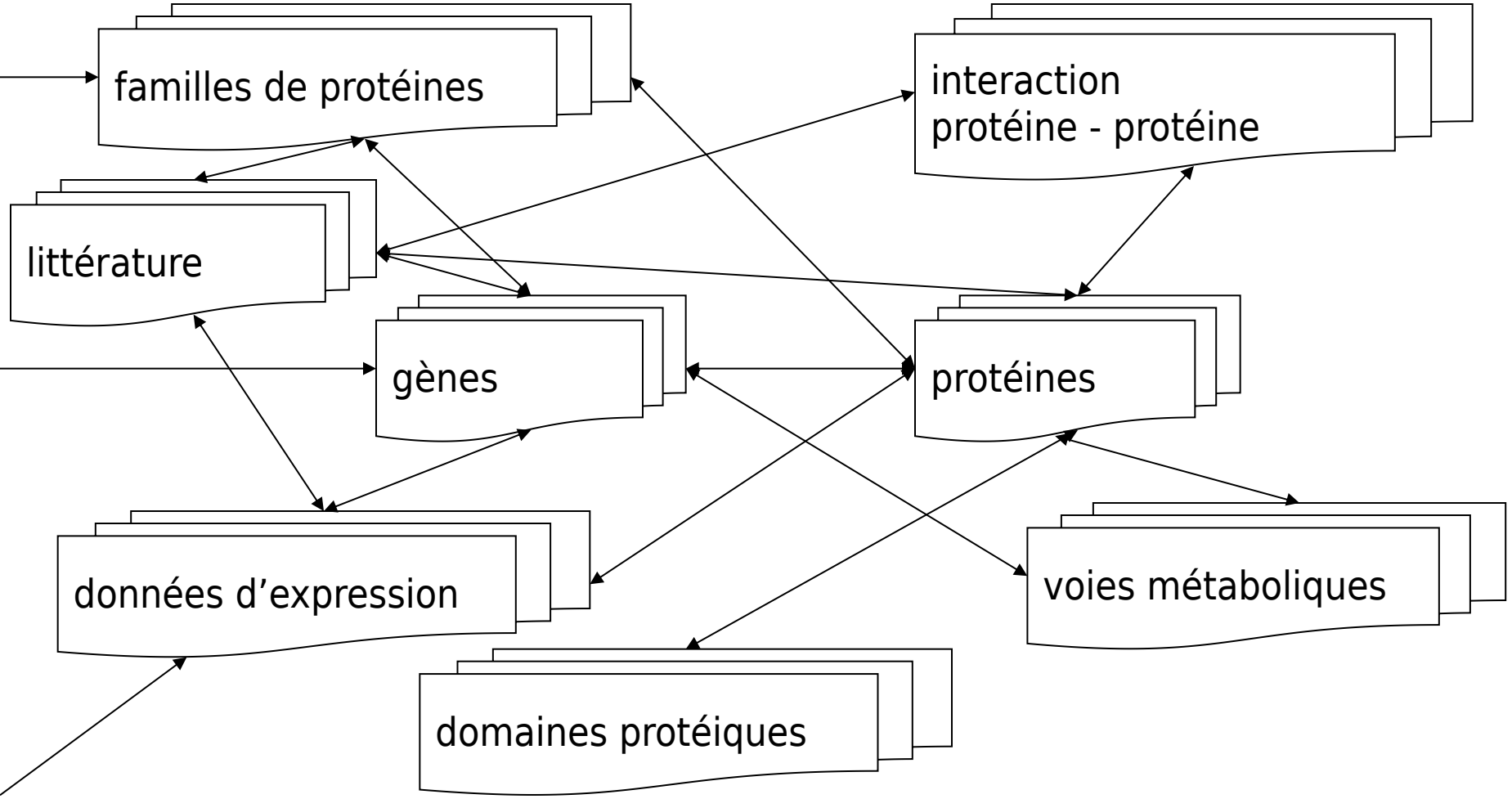


Cellule eucaryote

Intégration : vers une vue unifiée

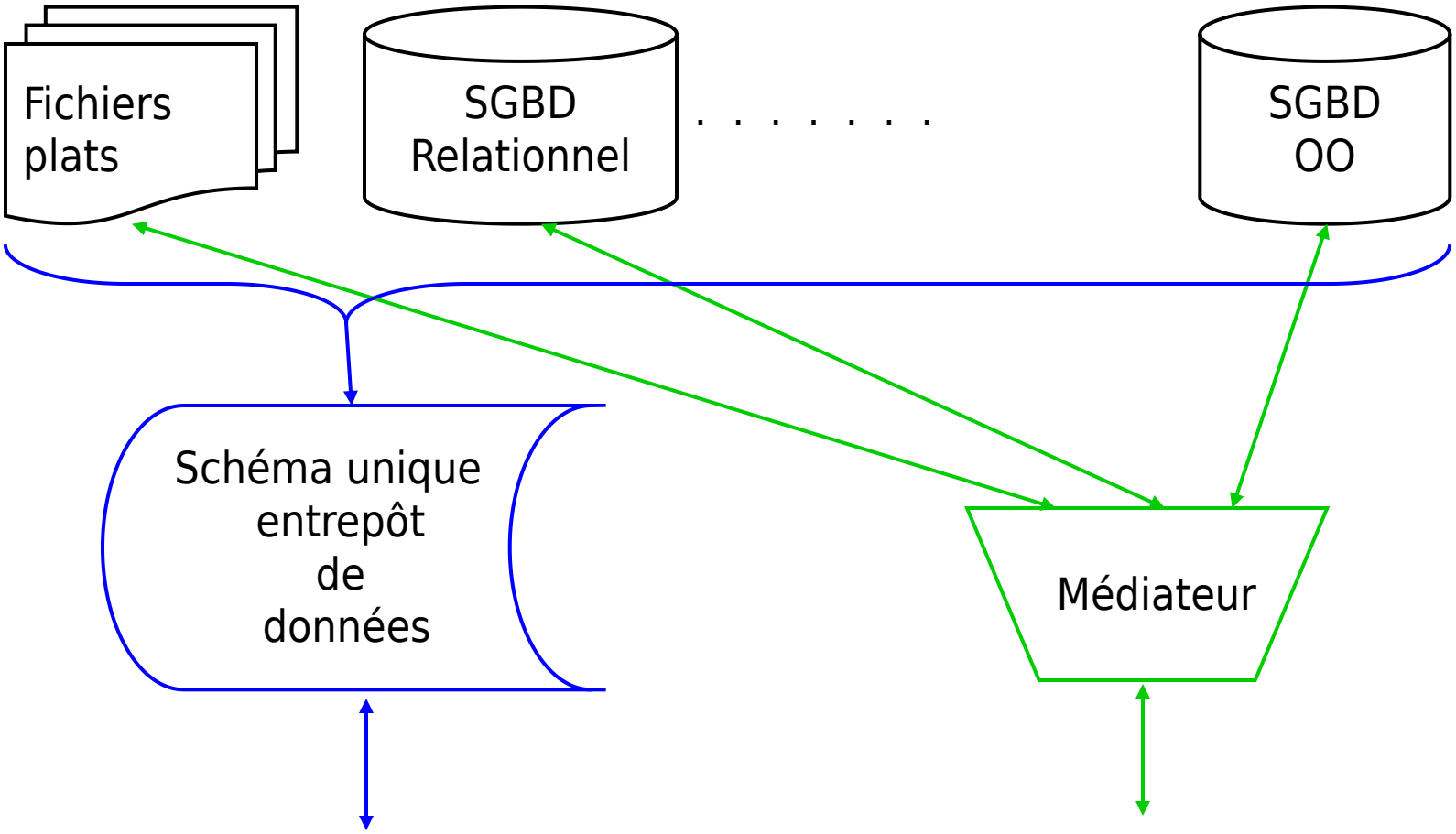
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

Intégration par interconnexion : principe



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

Intégration par fusion ou par médiateurs



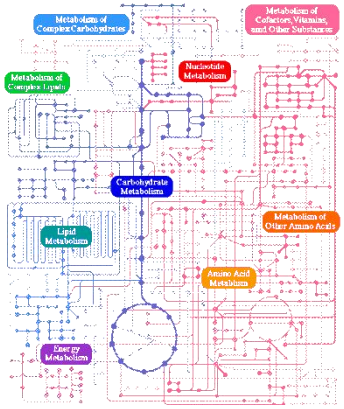
Integr8 [Kersey *et al.*, 2005], BioMart [Kasprzyk *et al.*, 2004],
WInGS [Abergel *et al.*, 2004], BioKleisli [Davidson *et al.*, 1997], ...

Intégration : vers une vue unifiée

- 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 la modélisation : vers la cellule virtuelle

Littérature

Réactions et vitesses	Données d'interaction
Composants cellulaires	Localisation cellulaire
Données sur le métabolisme	

Modèle
- système de réactions
- interactions moléculaires
- pi calcul

Caractérisation et simulations
- validation
- prédictions
- propriétés émergentes

Virtual Cell [Loew et Schaff, 2001], E-CELL [Tomita *et al.*, 1999],
Cellerator [Shapiro *et al.*, 2003],
MetExplore [Cottret *et al.*, 2010], ...

Intégration : vers une vue unifiée

- 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

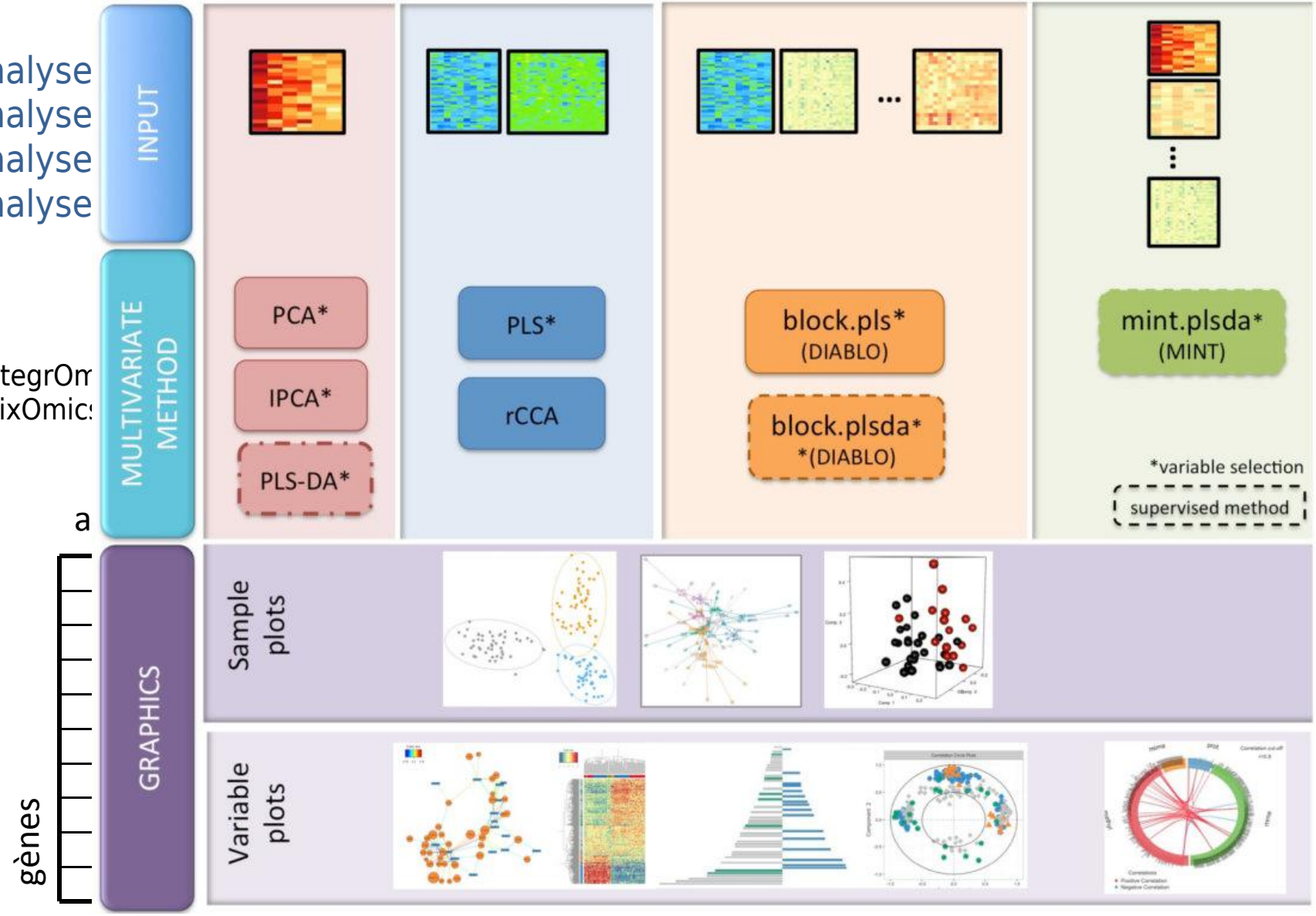
Statistiques

- Analyse
- Analyse
- Analyse
- Analyse
- …

integrOm
MixOmics

a

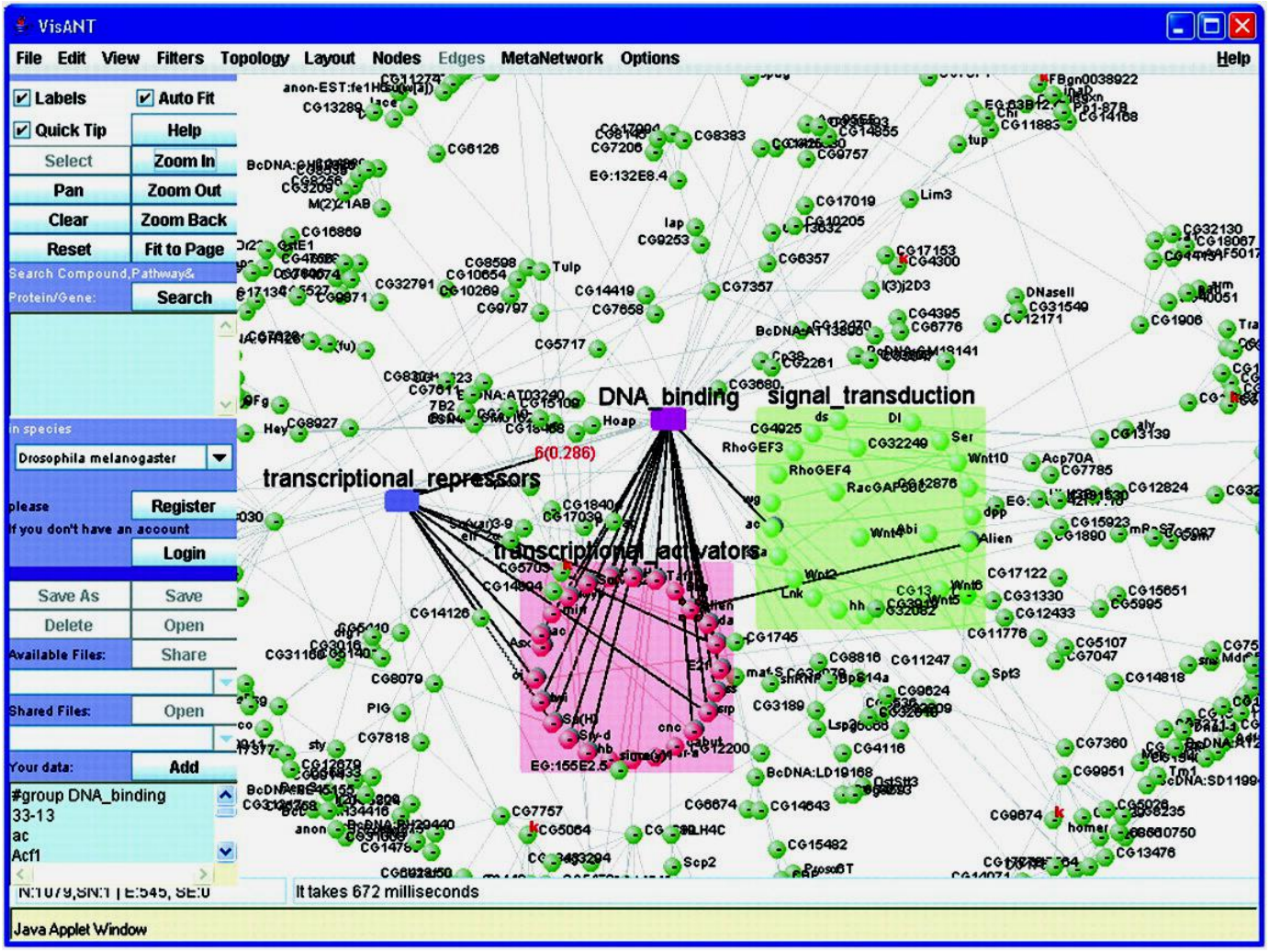
gènes



Intégration : vers une vue unifiée

- 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

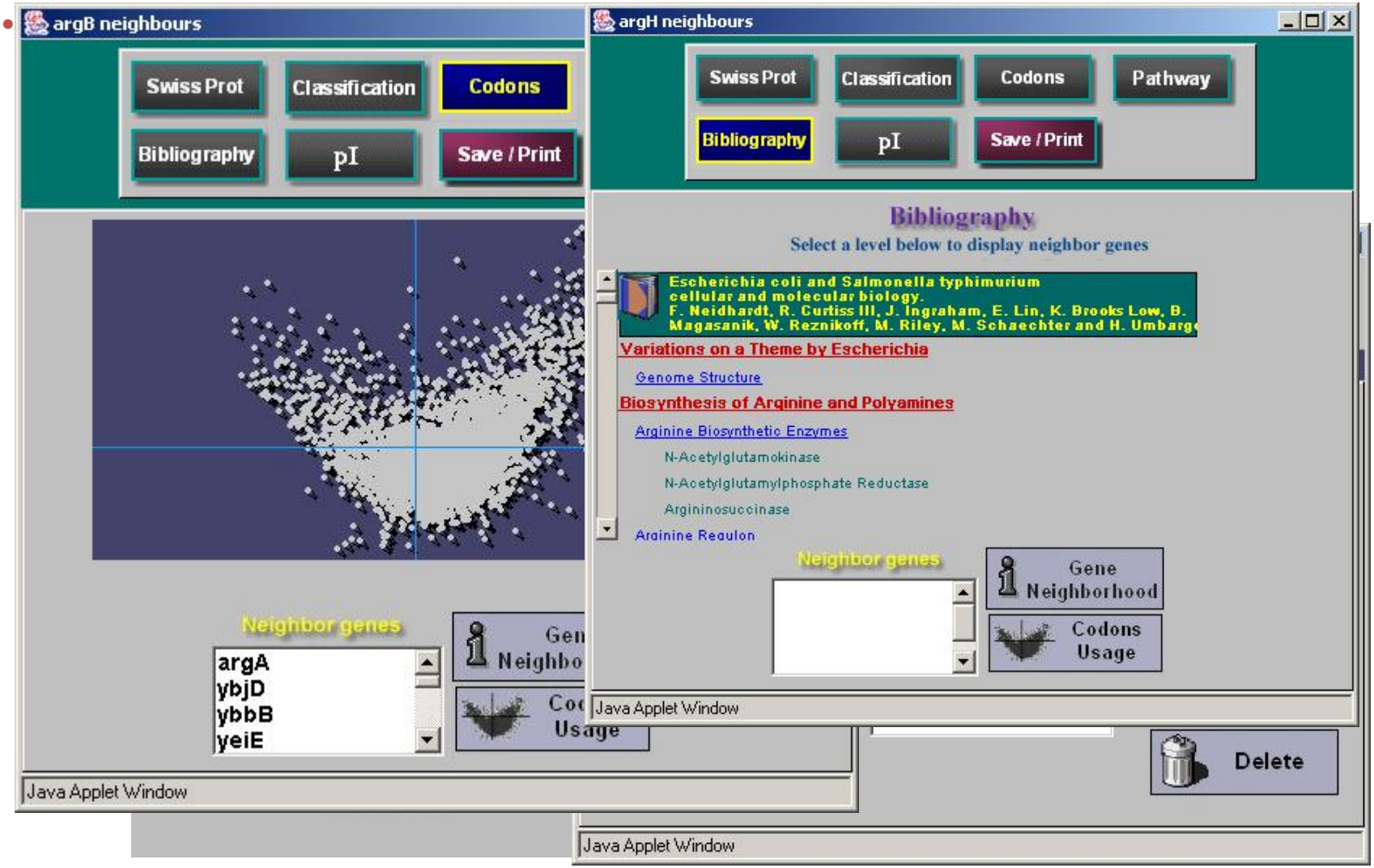


Visant [Hu *et al.*, 2005]

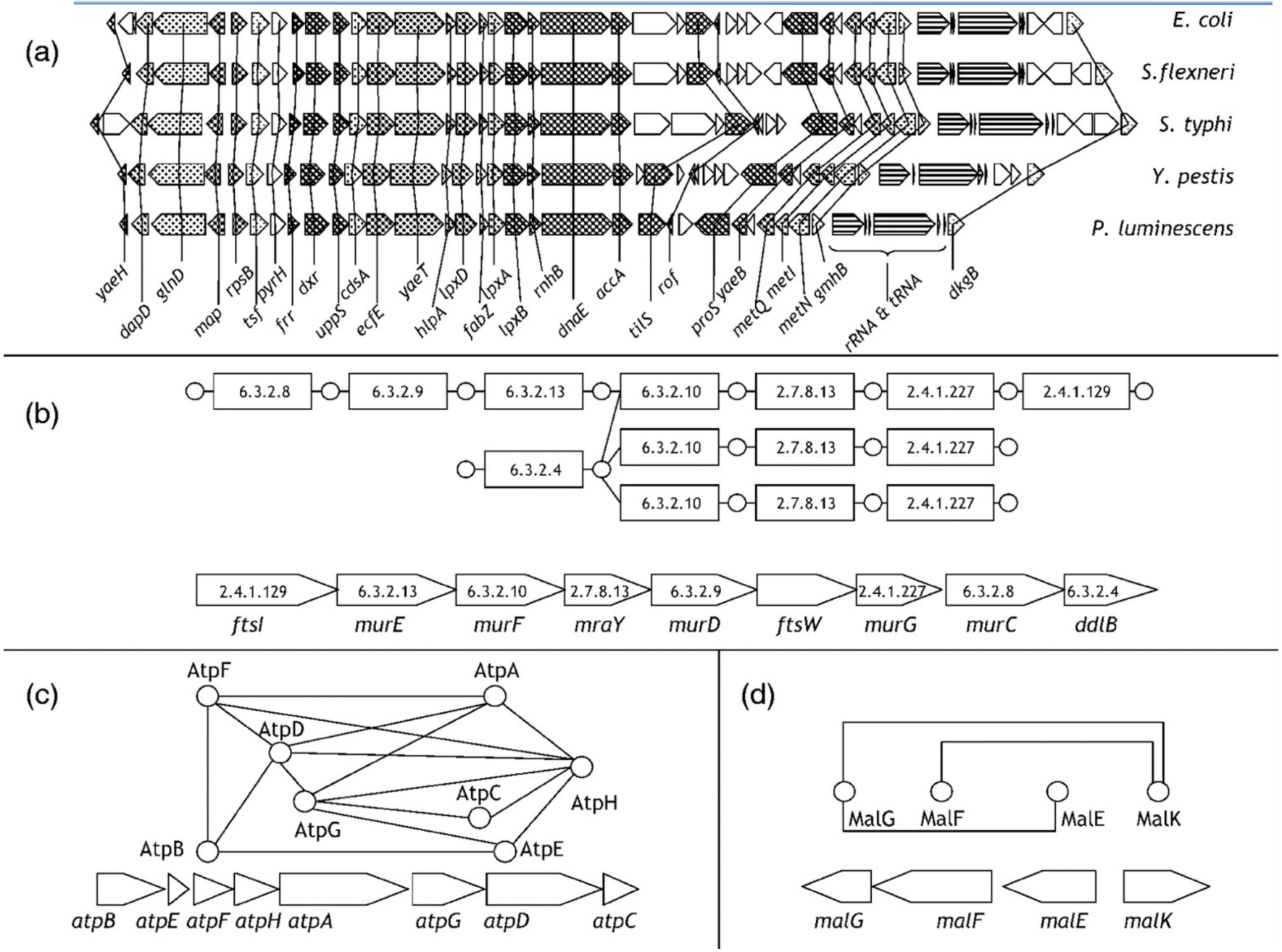
Intégration : vers une vue unifiée

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

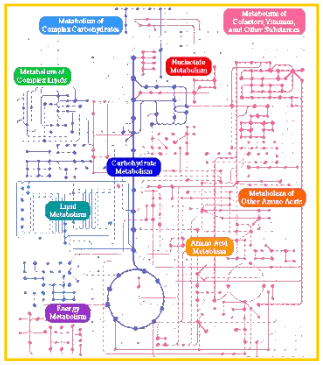
Exploration visuelle de voisinages



Recoupement de voisinages : approche basée sur les graphes

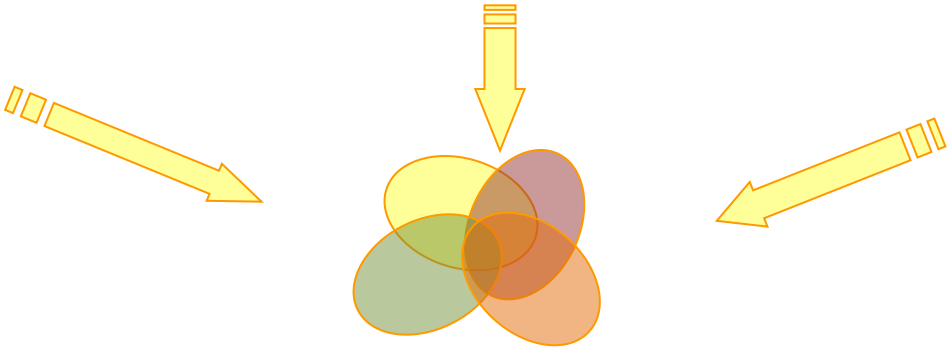


Recoupement de voisinages : approche ensembliste

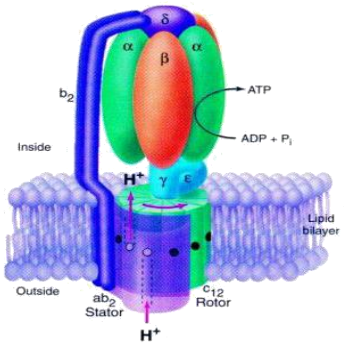


voies métaboliques

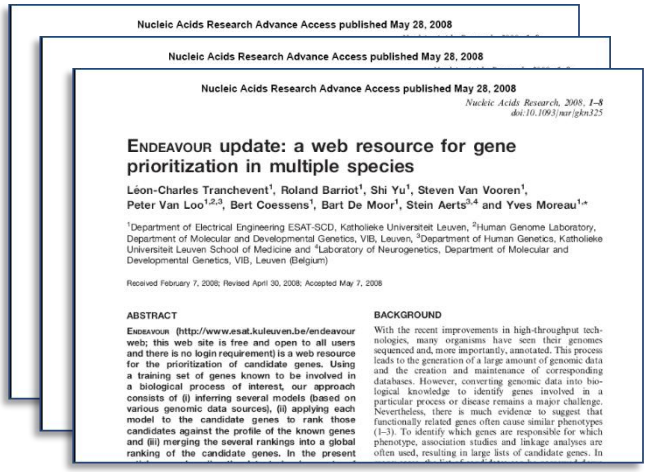
localisation chromosomique



ensembles de gènes



complexes protéiques



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

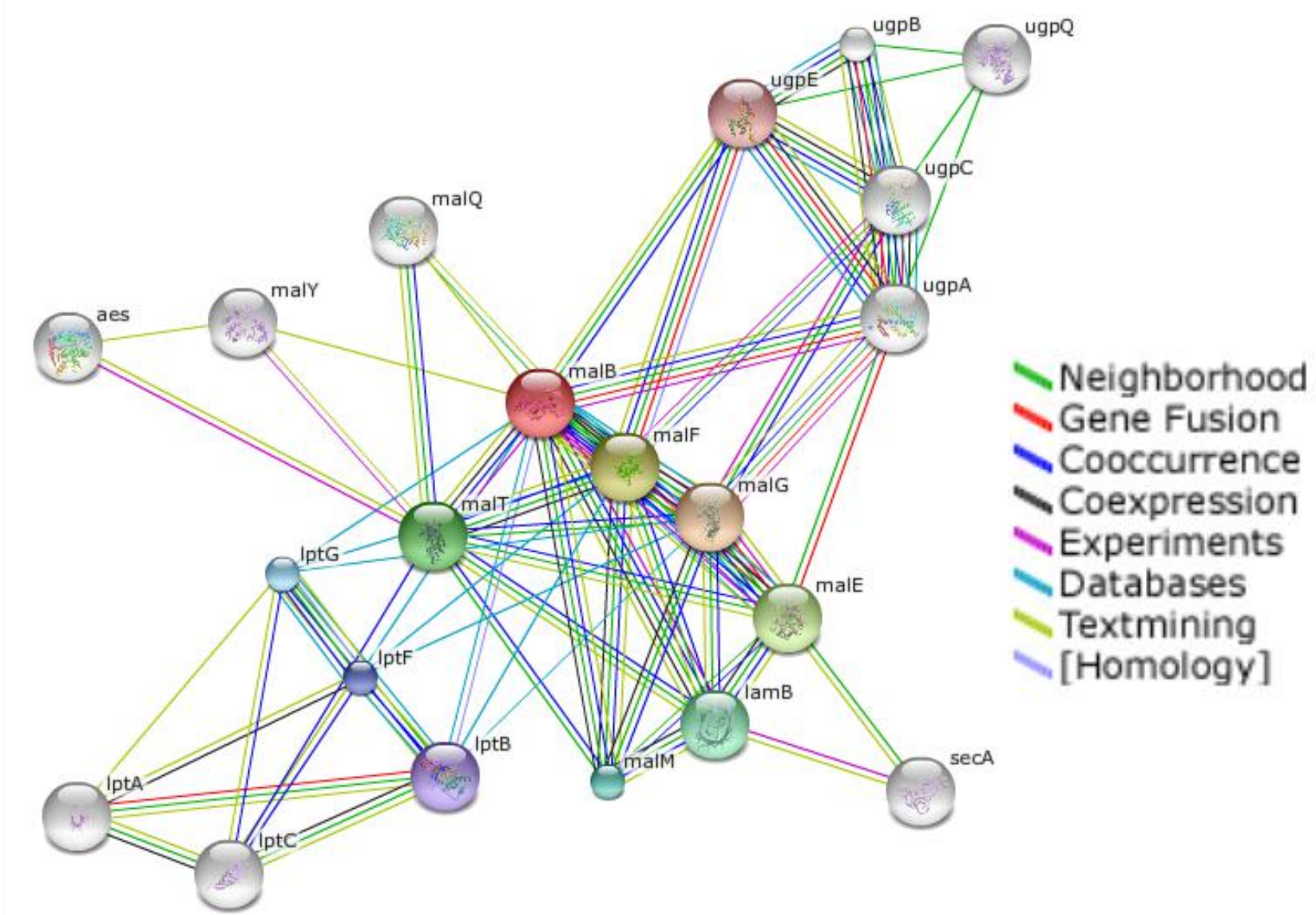
With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic 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

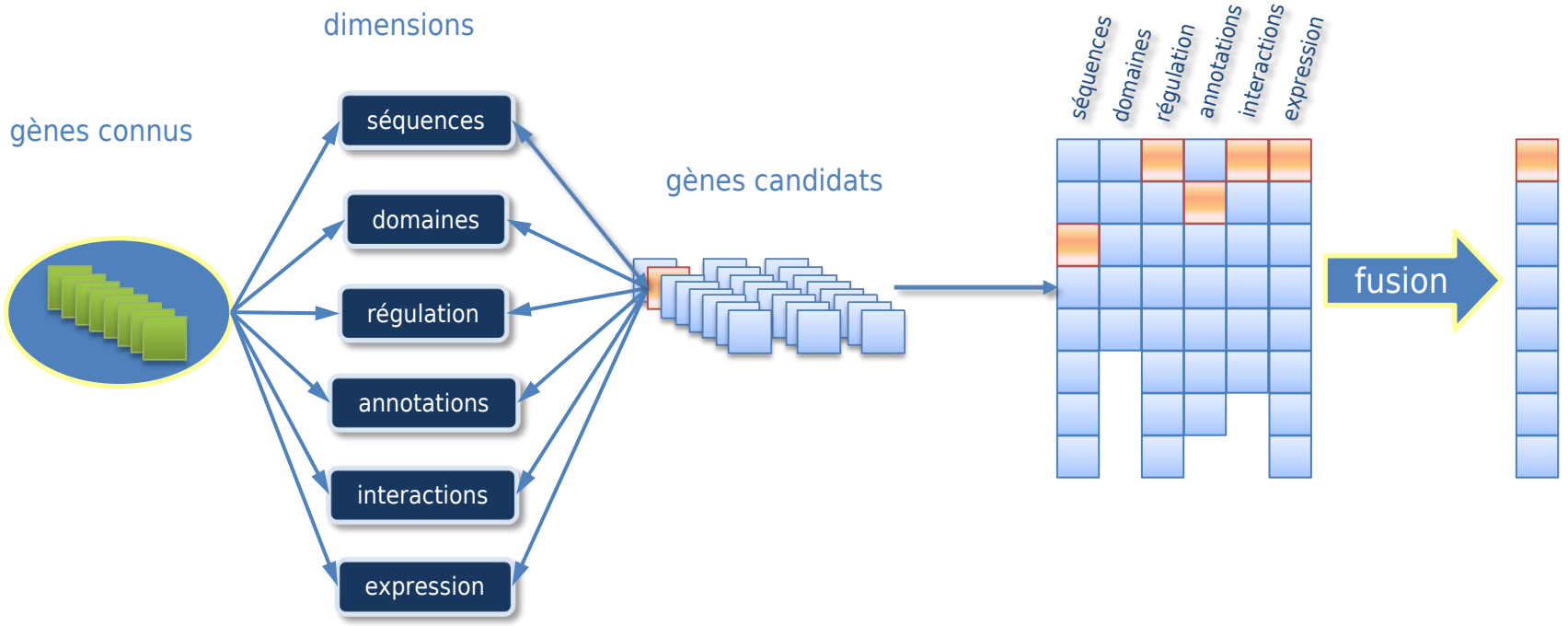
Gene Ontology

Fusion : approche basée sur les graphes



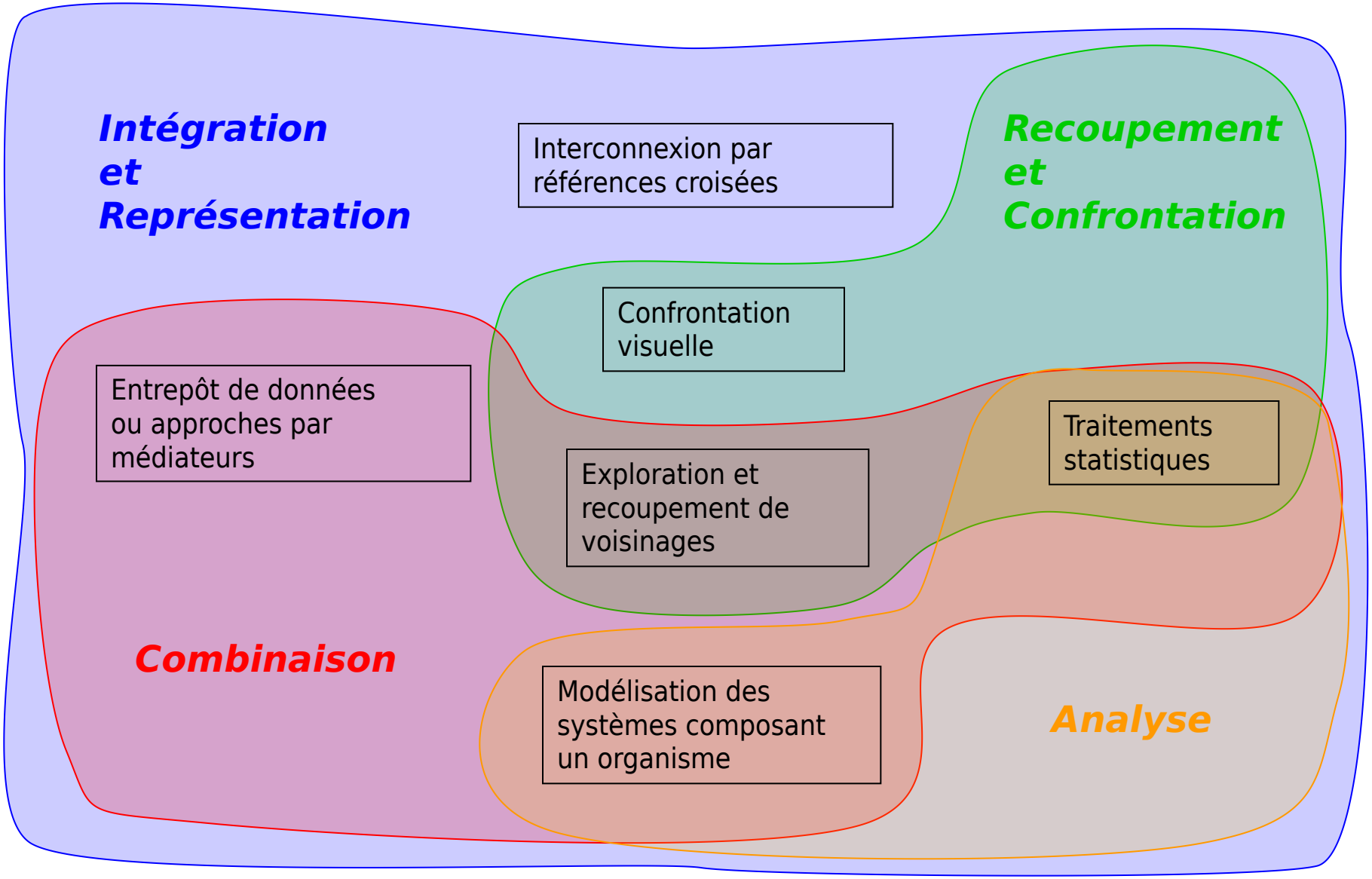
STRINGdb [von Mering *et al.*, 2003]

Approche apprentissage automatique / fusion : dimensions & dissimilarités. Priorisation et clustering



[Aerts *et al.*, 2006]

Vue générale



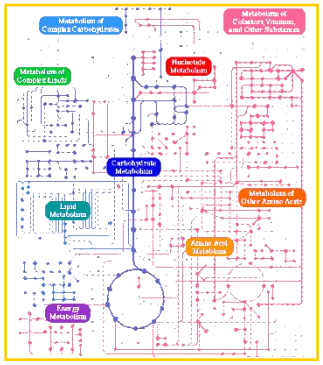
Gestion de données non structurées - Applications Post-Génomiques

Enrichissement

Master 2

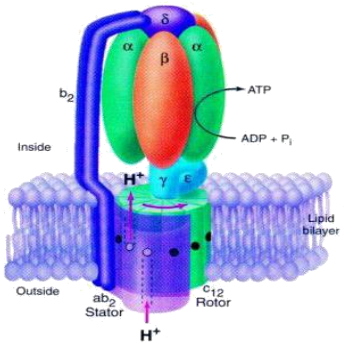
Bioinformatique et Biologie des Systèmes

Recoupement de voisinages : approche ensembliste

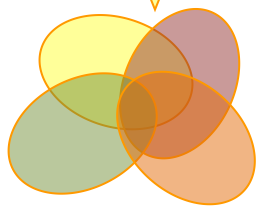


voies métaboliques

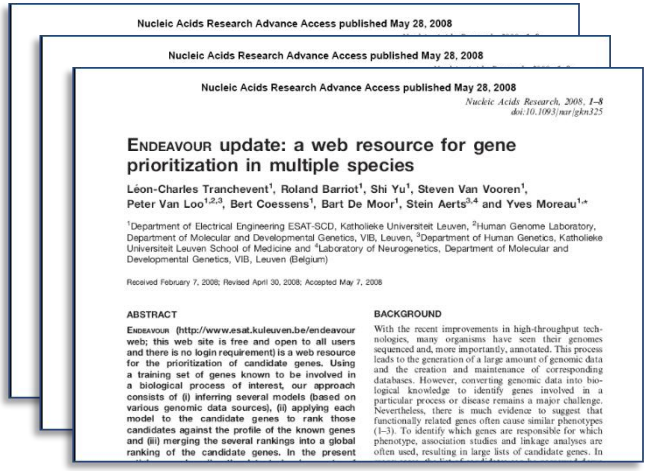
localisation chromosomique



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/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

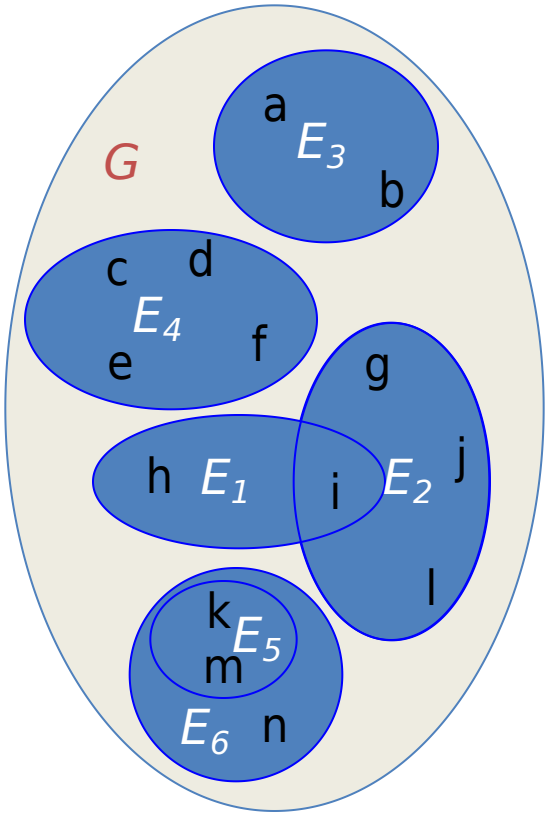
co-citation

domaines protéiques

Gene Ontology

Définitions

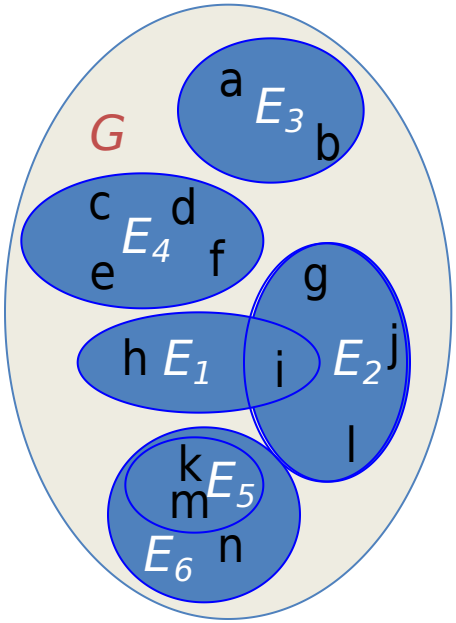
- (Identifiants de) gène → ARNm → protéine
- *G* : ensemble des gènes d'un organisme
- *Fonction de regroupement* : relation entre gènes basée sur un indice de similarité.
- *Ensemble de (gènes) voisins* : ensemble de gènes *E* *G* regroupés par une fonction de regroupement.
- *Voisinage* : sous-ensemble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



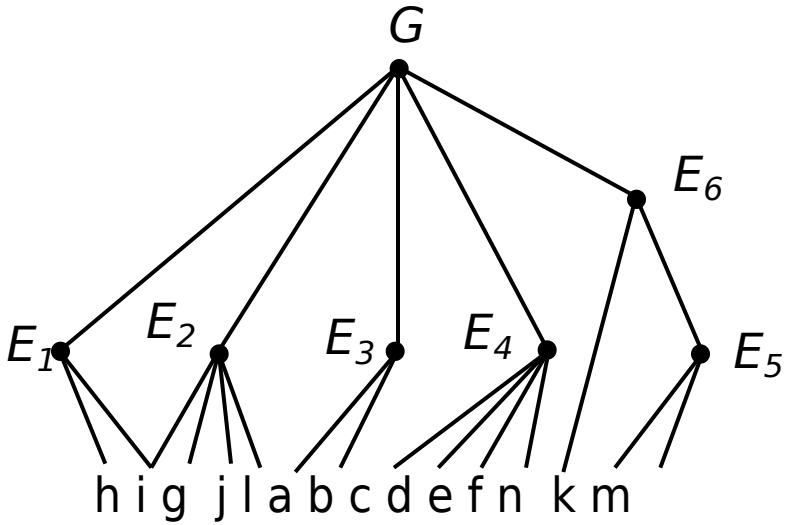
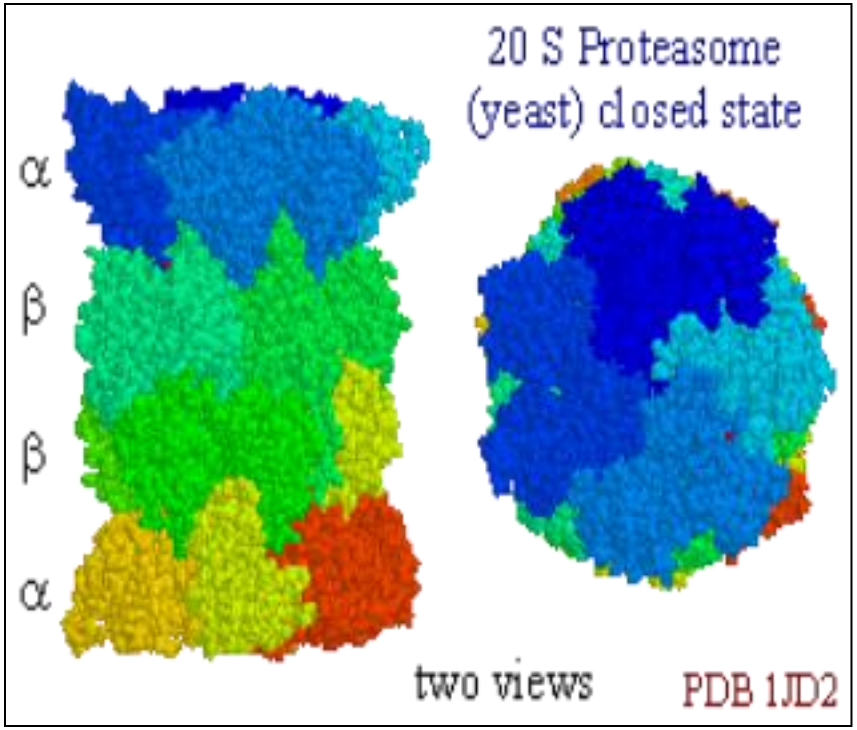
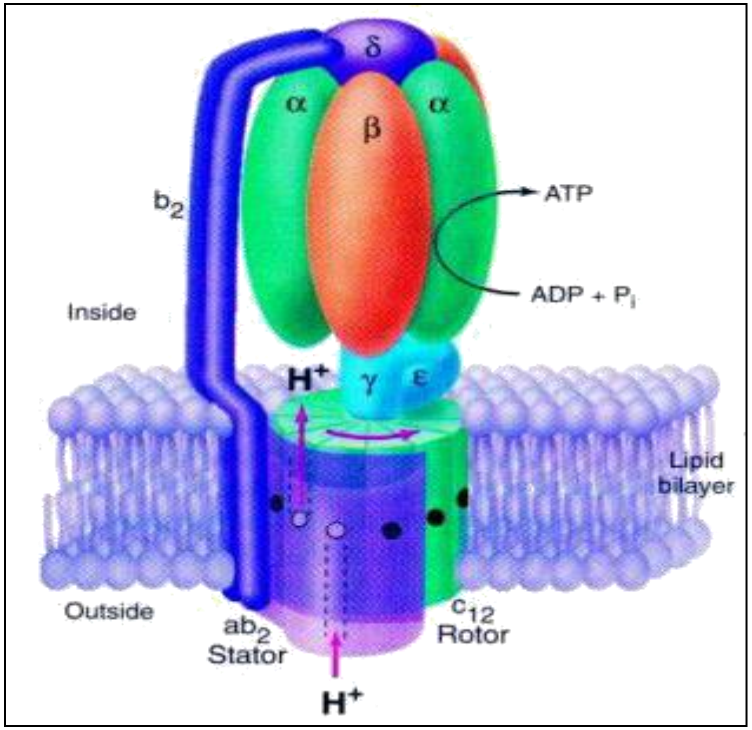


diagramme de Hasse de V

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

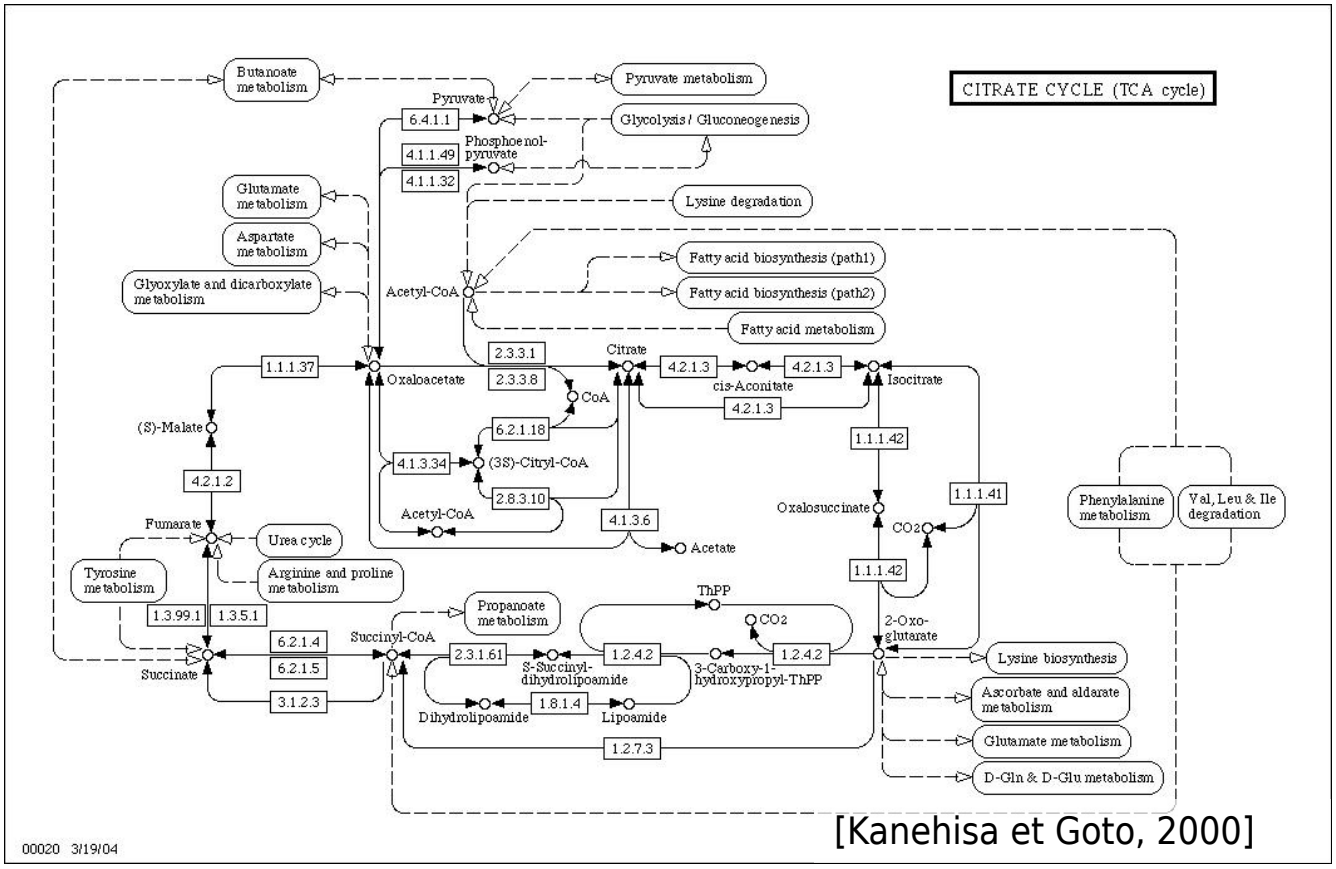
Exemple de fonction de regroupement : complexes protéiques





un complexe → un ensemble de protéines

Exemple de critère de regroupement : voies métaboliques

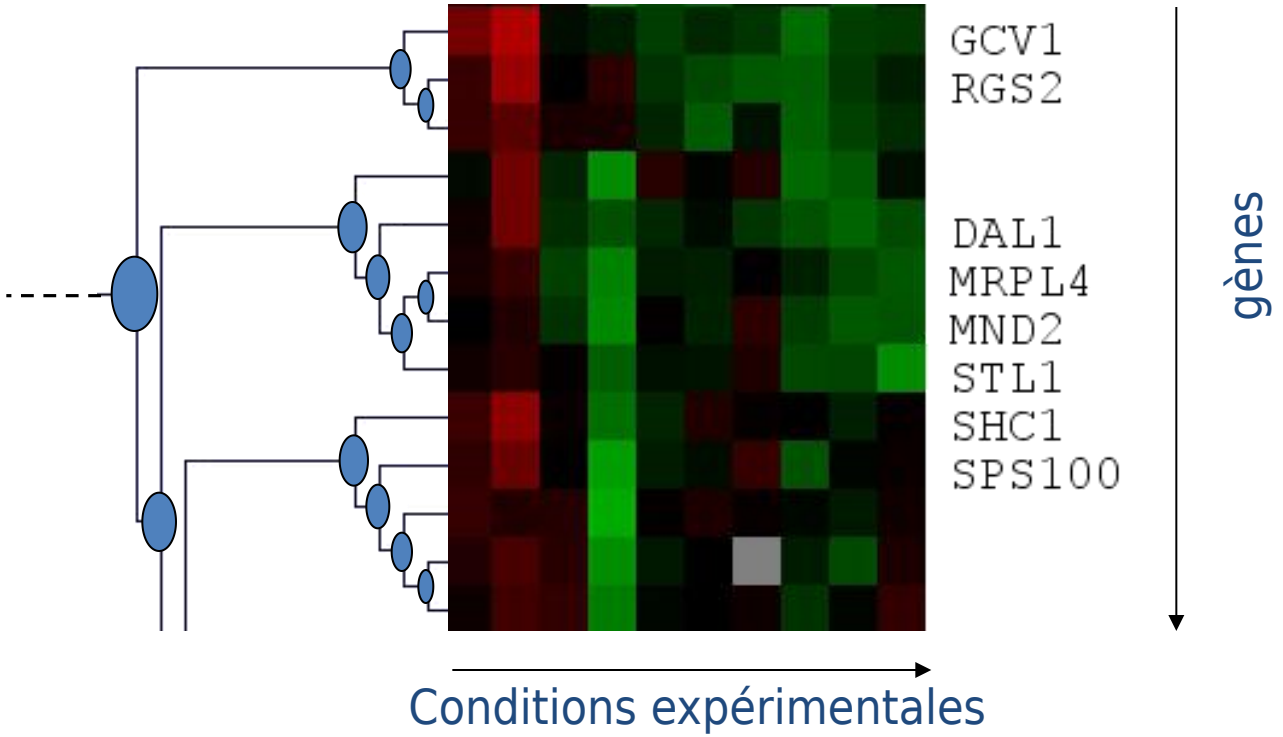


[Kanehisa et Goto, 2000]

une voie métabolique → un ensemble de protéines

Exemple de critère de regroupement : données d'expression

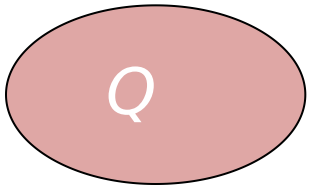
clustering hiérarchique
des profils



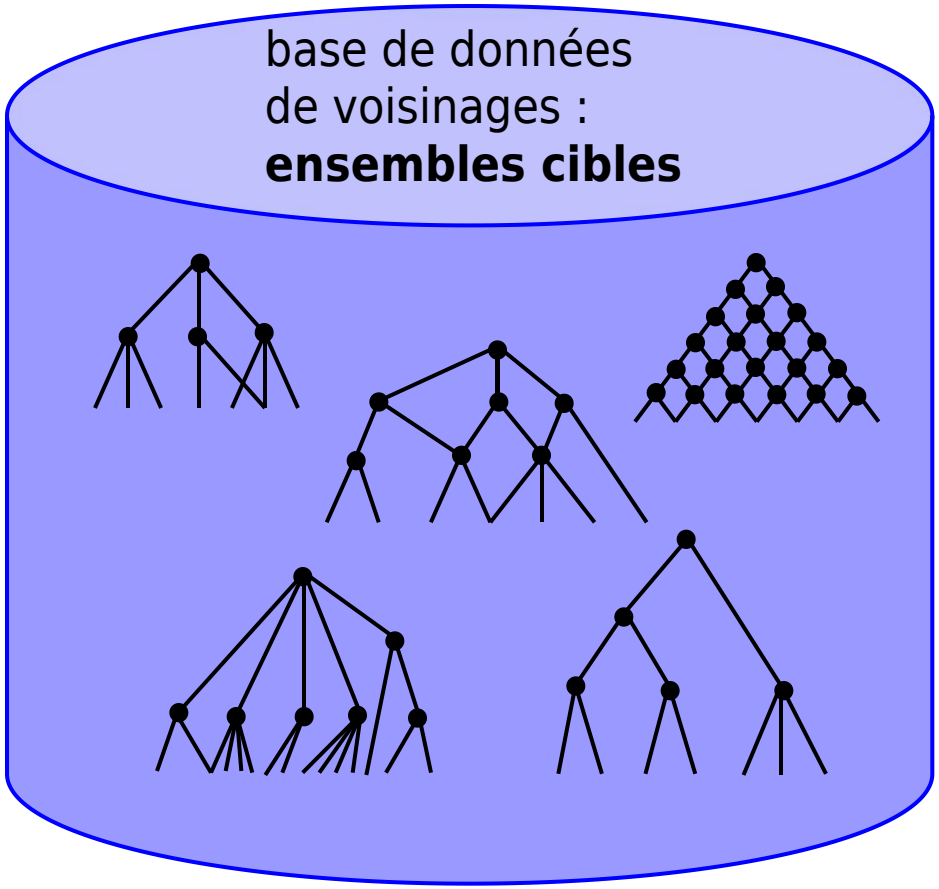
un cluster → un ensemble de gènes

- Recherche d'ensembles similaires

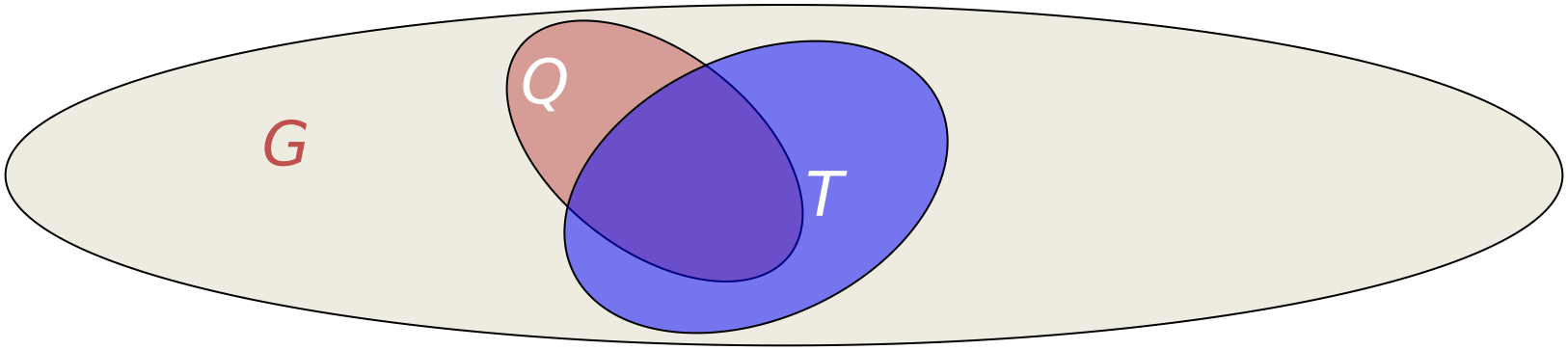
ensemble requête
$Q \subseteq G$



Quels sont les ensembles cibles qui lui sont similaires ?



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

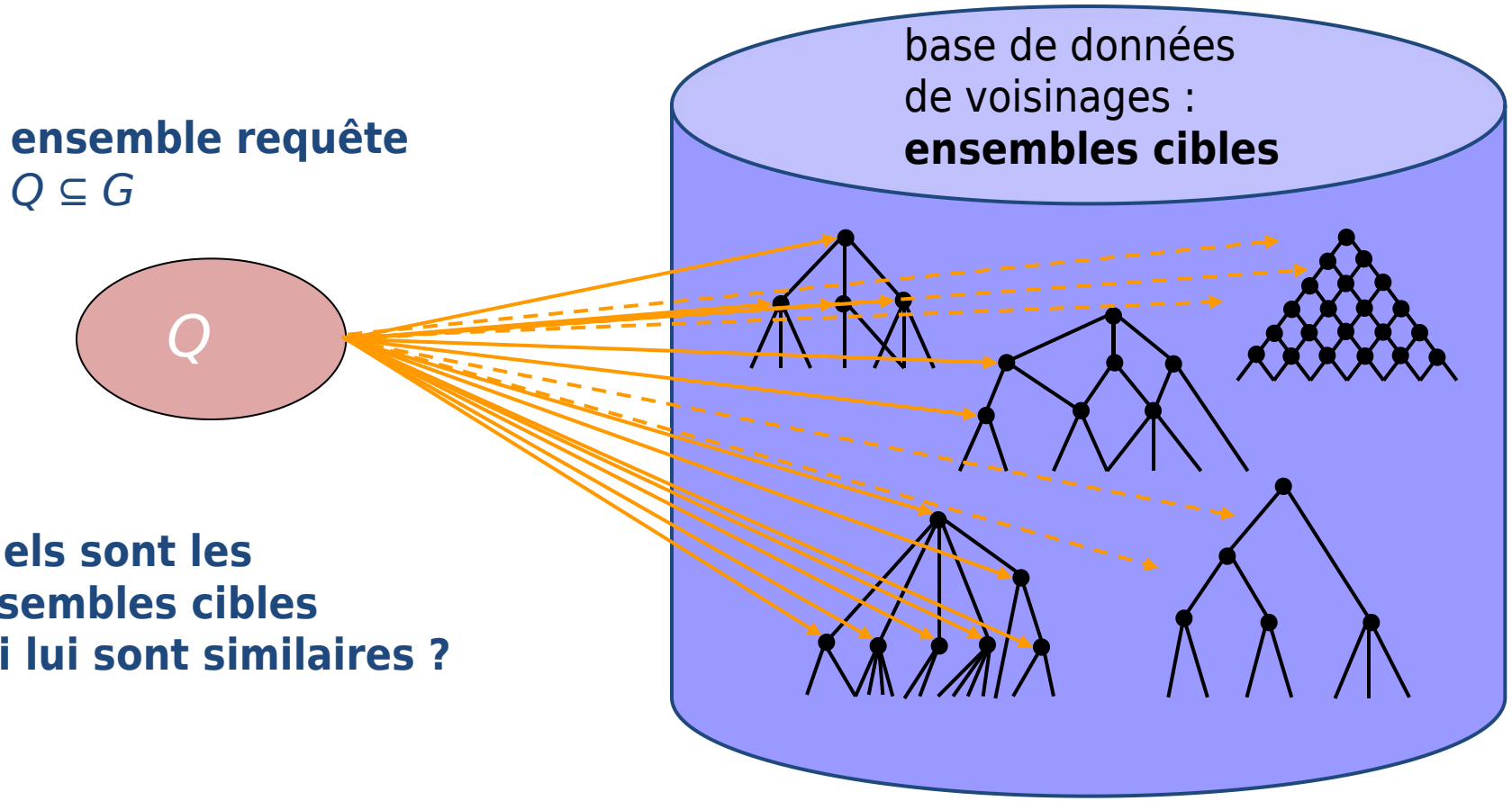
$$p-valeur(c,t,q,g) = \sum_{k=c}^{\min(q,t)} \frac{\binom{t}{k}\binom{g-t}{q-k}}{\binom{g}{q}}$$

avec

- $g = |G|$: taille de la population
- $q = |Q|$: taille de l'ensemble requête
- $t = |T|$: taille de l'ensemble cible
- $c = |Q \cap T|$: nombre d'éléments communs

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

- Recherche d'ensembles similaires

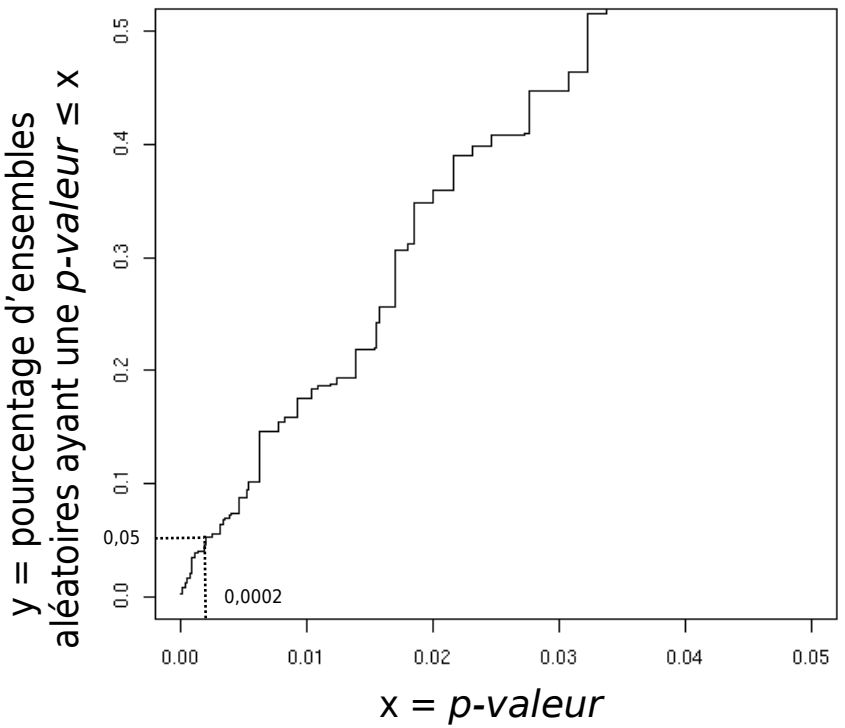


Tests multiples et significativité des p-valeurs

- 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

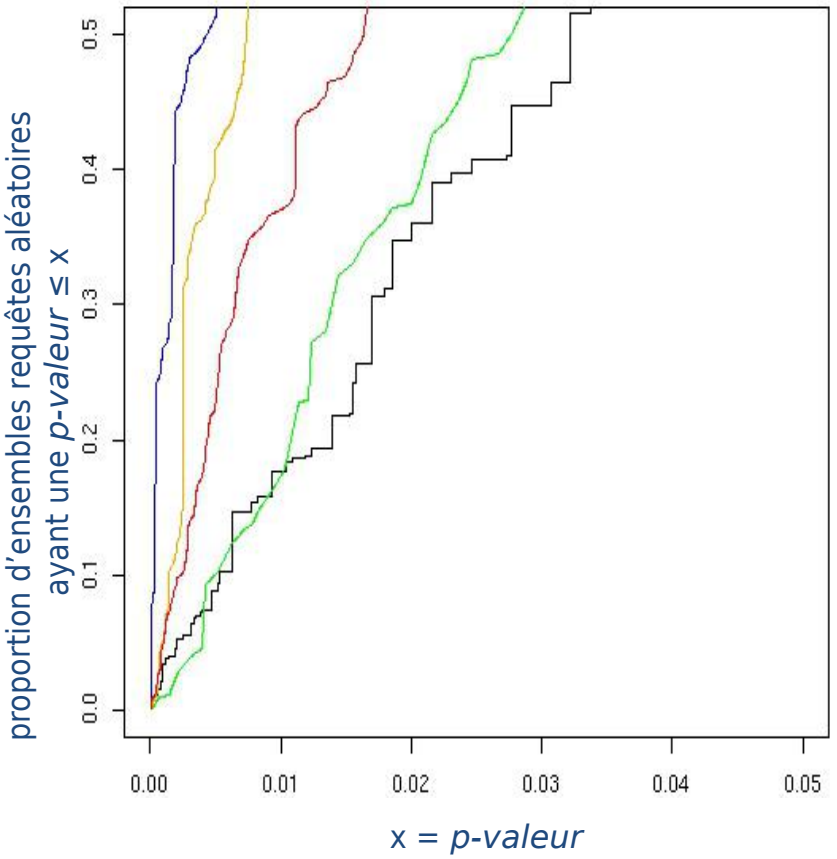
Étant donnée une p-valeur *p*
Combien ont un meilleur score ?



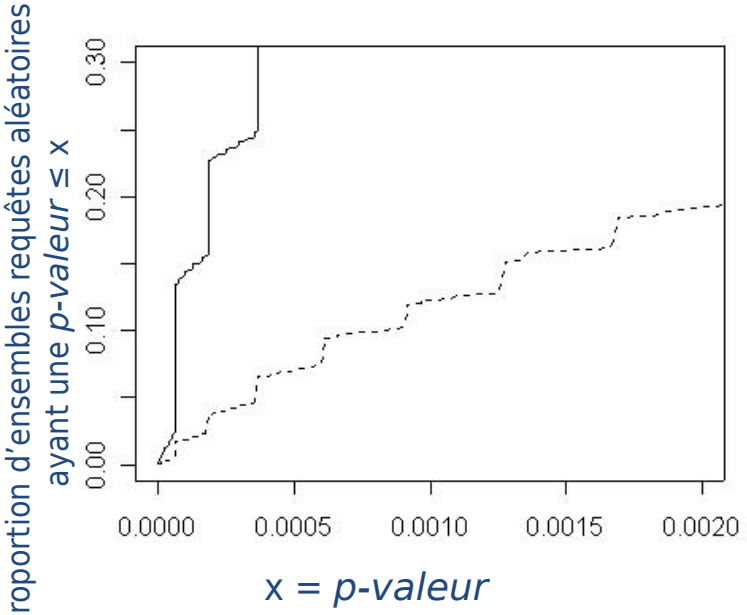
levure *Saccharomyces cerevisiae*
n=500, q=9, g=5786, KEGG Pathways

[Barriot *et al.* 2004]

Significativité des p-valeurs obtenues



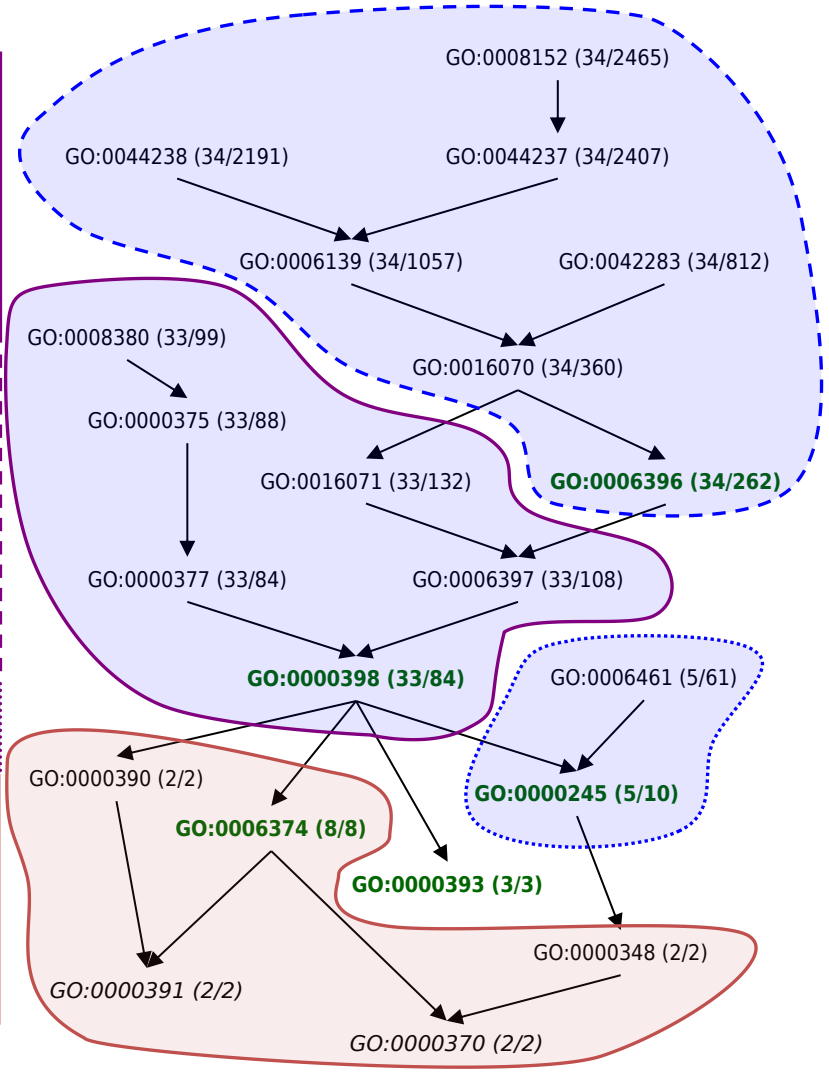
Saccharomyces cerevisiae
n=500, q=6-9-200-500-1000,
g=5786, KEGG Pathways



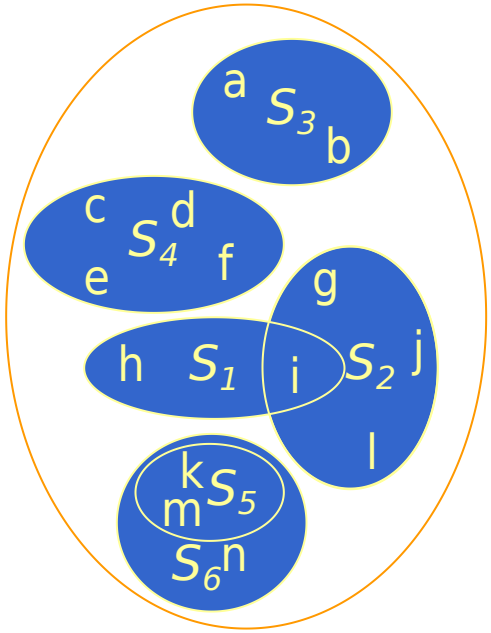
Saccharomyces cerevisiae
n=500, q=50, g=5786,
——— GO molecular function,
- - - Ferea *et al.*, 1999

Complex 440.30.10 mRNA splicing

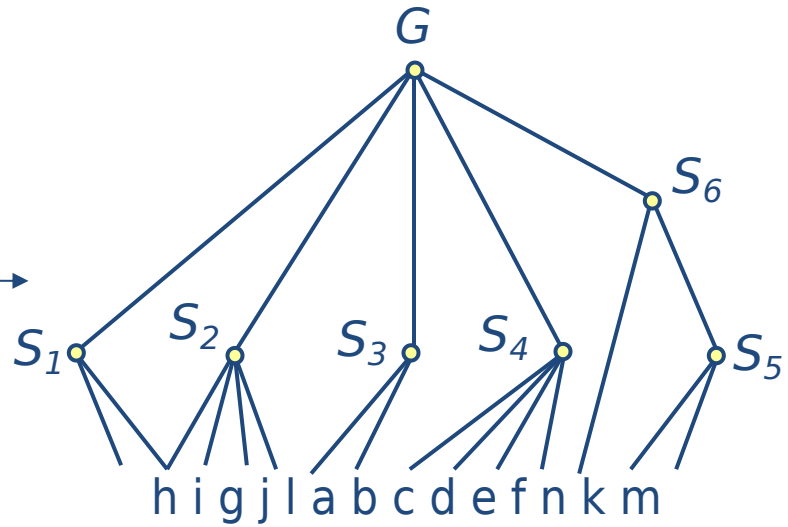
GO Term	Description	Target size	Common elements
GO:0000398	**nuclear mRNA splicing, via spliceosome**	**84**	**33**
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	84	33
GO:0000375	RNA splicing, via transesterification reactions	88	33
GO:0008380	RNA splicing	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	*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**



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

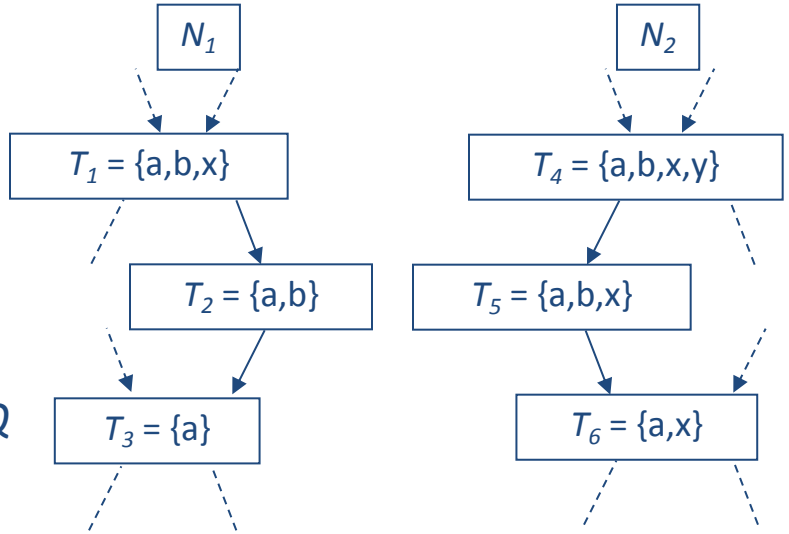
$$Q \cap T \neq \emptyset$$

and

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

and

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

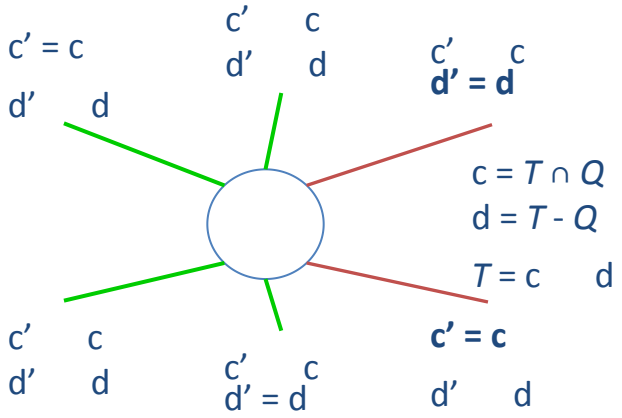


Local decision

$|c| > 0$

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

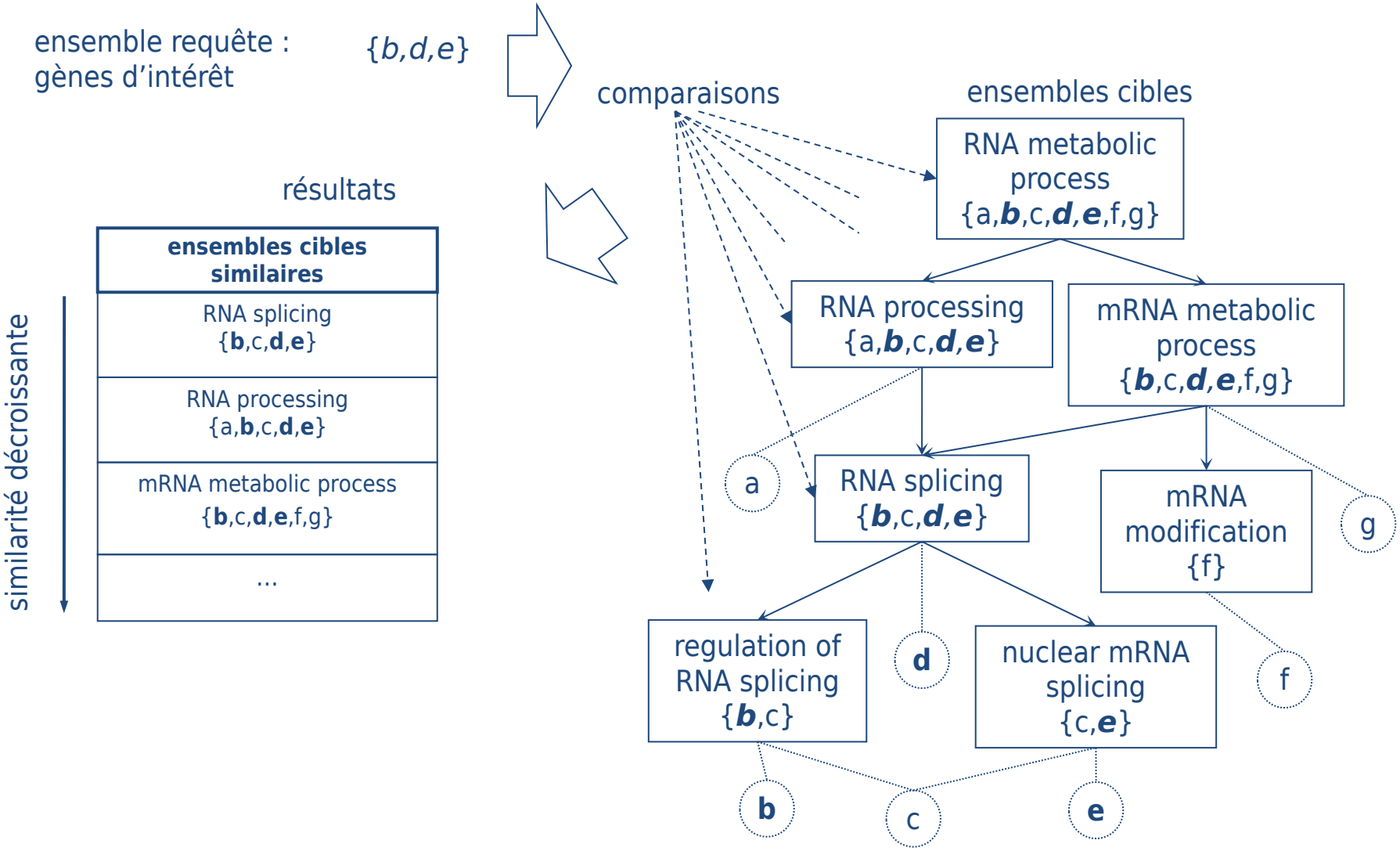
$|c| > \max(\{c_{children}\})$



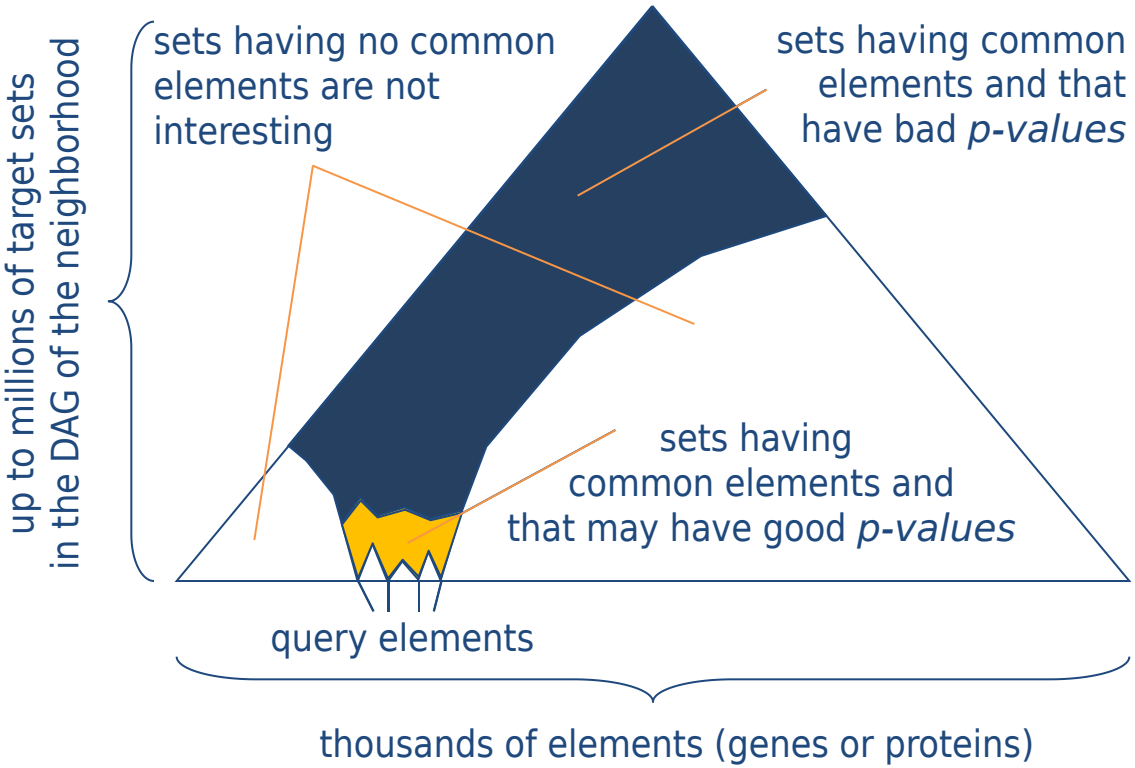
[Barriot, Dutour, Sherman, 2007, *BMC Bioinformatics*]

Illustration

- Pertinence des comparaisons & redondance des résultats

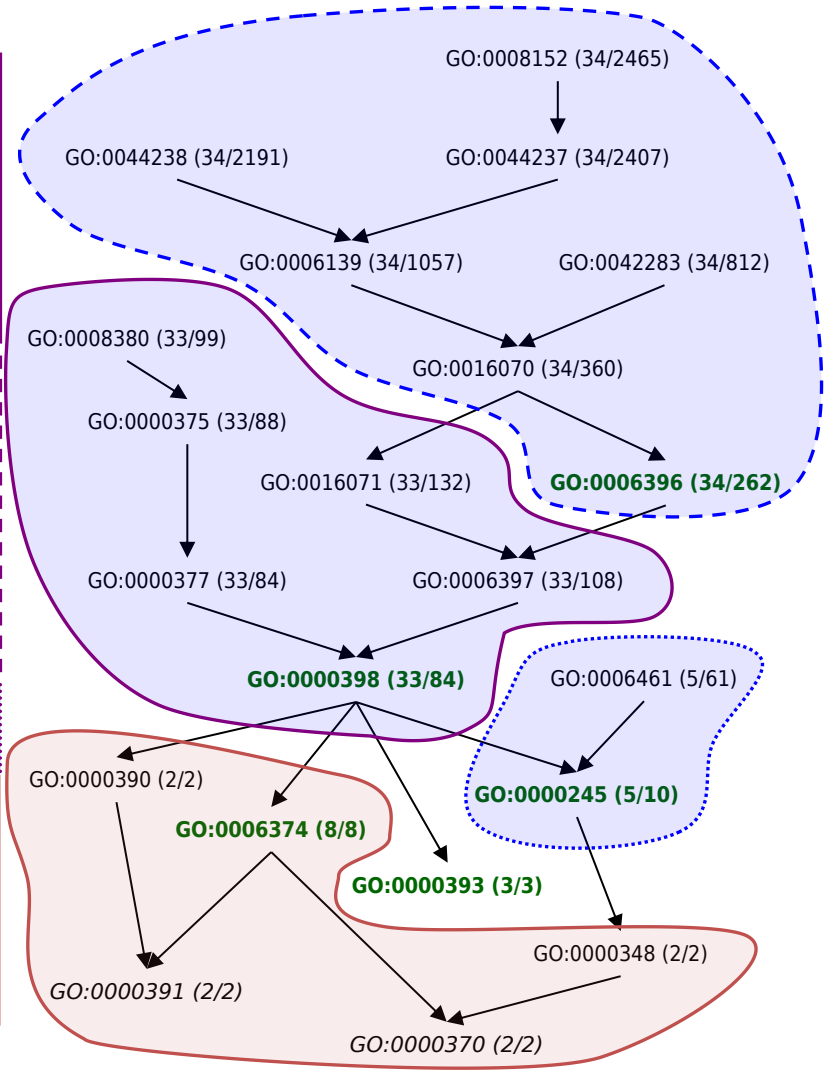


A small portion of the DAG is searched



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	*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**



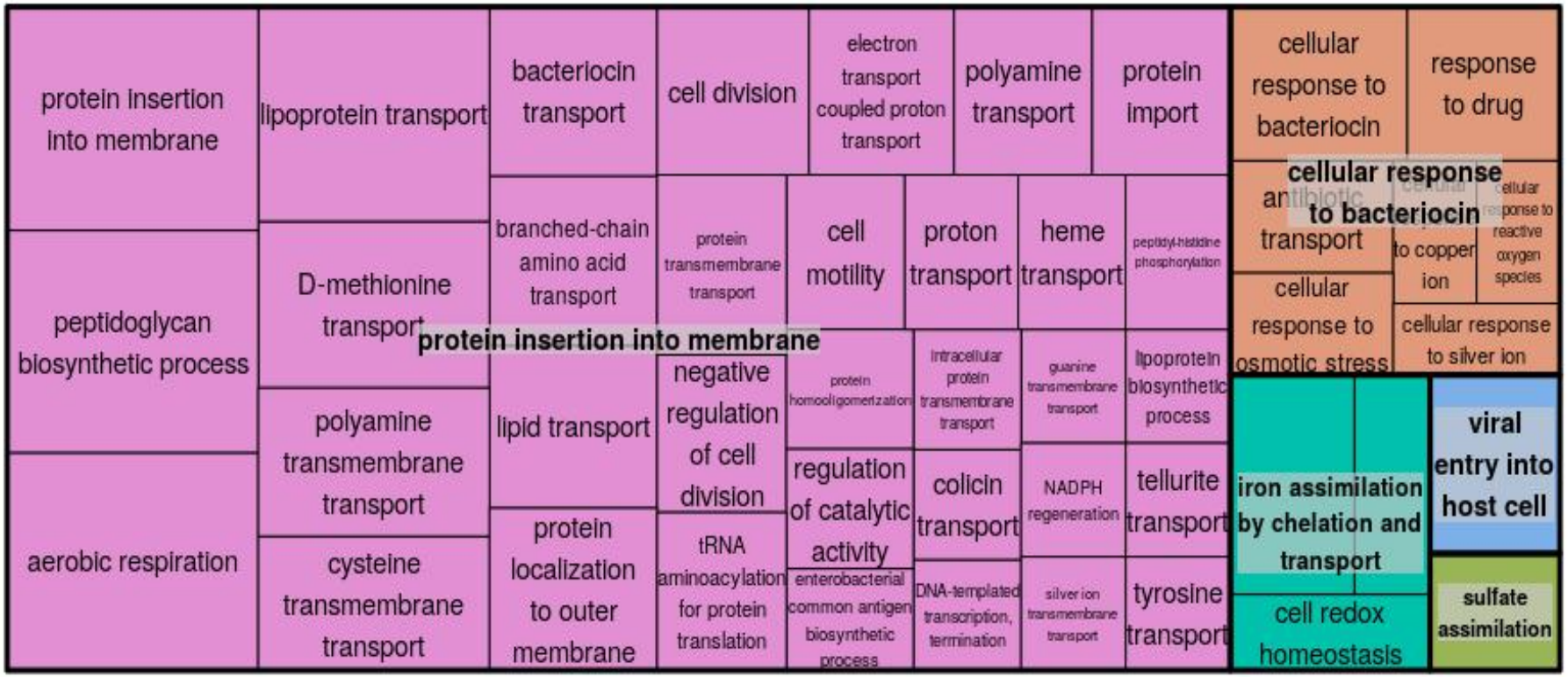
Autre approche pour la visualisation et l'interprétation des résultats

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

GOTerm	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: amino acid transmembrane transport	4	29
GO:0006865	0.052296588628331	BP: amino acid 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 amino acid 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 bacteriocin	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: colicin transport	1	2
GO:1903712	0.002401639427418	BP: cysteine transmembrane transport	3	5
GO:0042883	0.001277327190964	BP: cysteine transport	3	4
GO:0048473	0.000559823205814	BP: D-methionine transport	3	3
GO:0006353	0.099101977694913	BP: DNA-templated transcription, termination	1	2
GO:0015990	0.008780962317314	BP: electron transport coupled proton transport	3	8
GO:0009246	0.080891352640116	BP: enterobacterial common antigen biosynthetic process	2	9
GO:1903716	0.099101977694913	BP: guanine transmembrane transport	1	2
GO:0015886	0.039728901461296	BP: heme transport	2	6

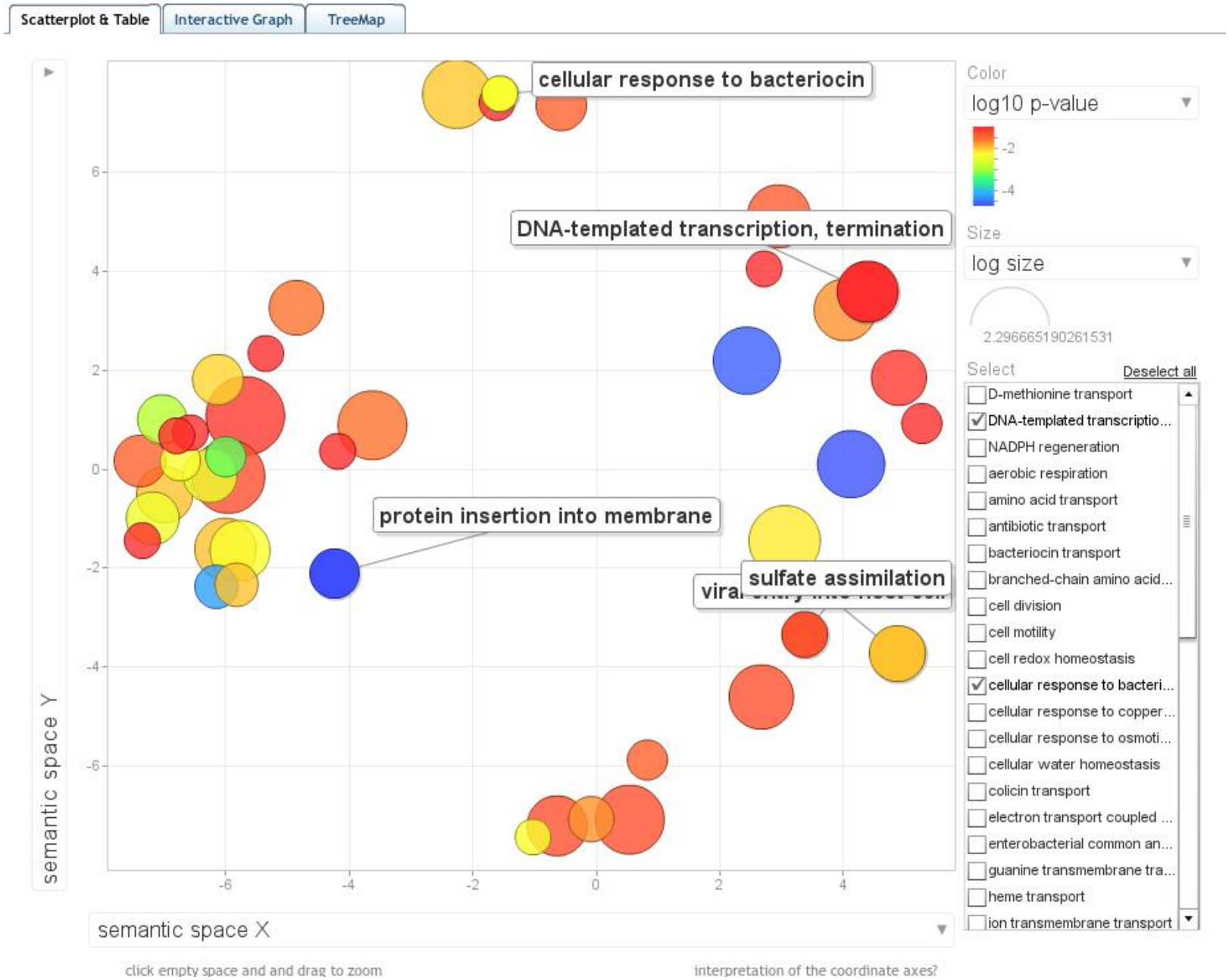
Treemap visualisation

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



REVIGO [Supek *et al.*, 2011]

REVIGO scatterplot



Similarité entre Termes GO

- Node/term information content

$$IC(term) = -\log p(term)$$ with p(term)= freq(term)

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

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

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

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

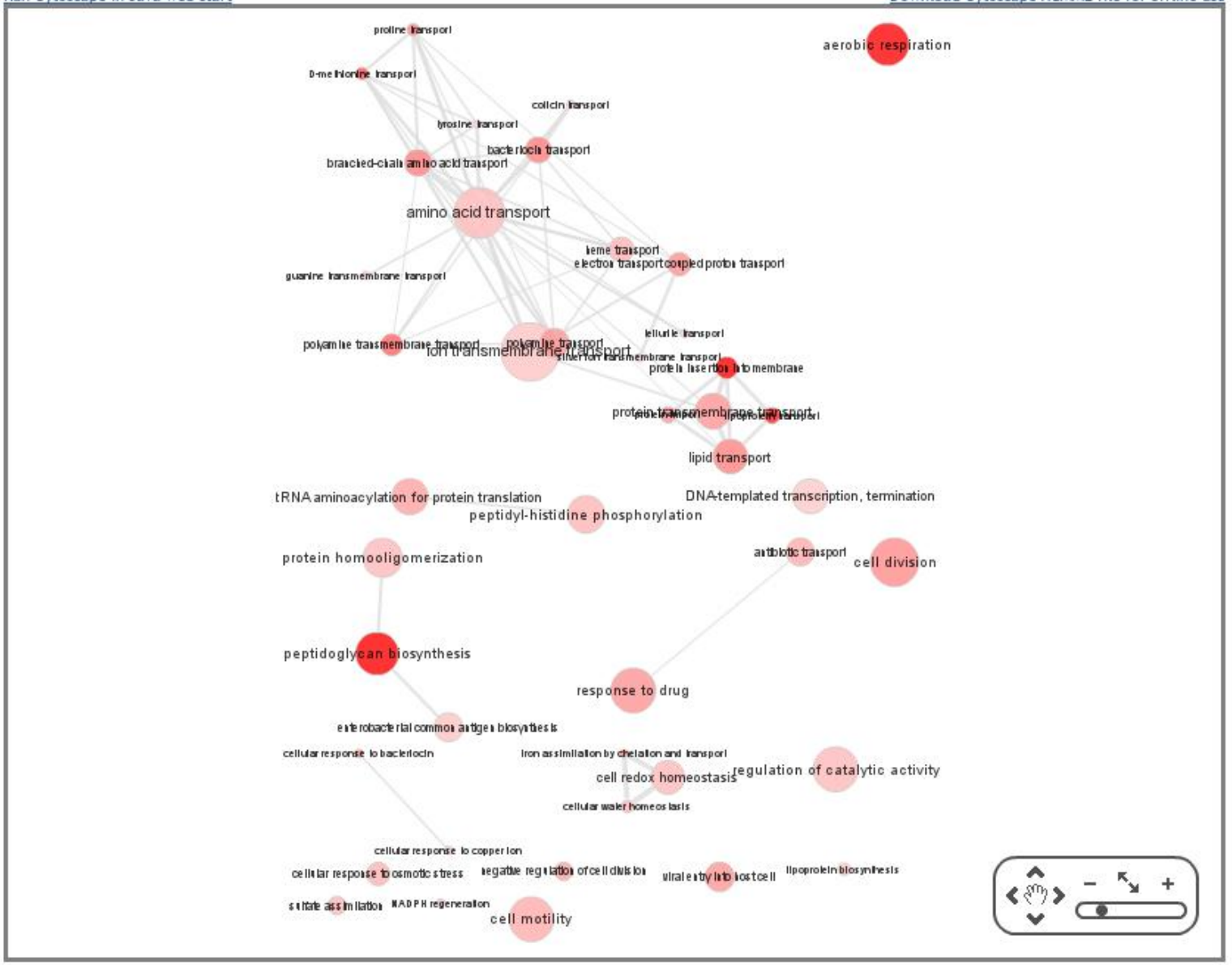
- $$sim_{gic}(t_1,t_2) = \frac{\sum_{t \in \{GO(t_1) \cap GO(t_2)\}} IC(t)}{\sum_{t \in \{GO(t_1) \cup GO(t_2)\}} IC(t)}$$ [Pesquita *et al.*, 2008]

REVIGO interactive graph

Scatterplot & Table | Interactive Graph | TreeMap

Run Cytoscape in Java web start

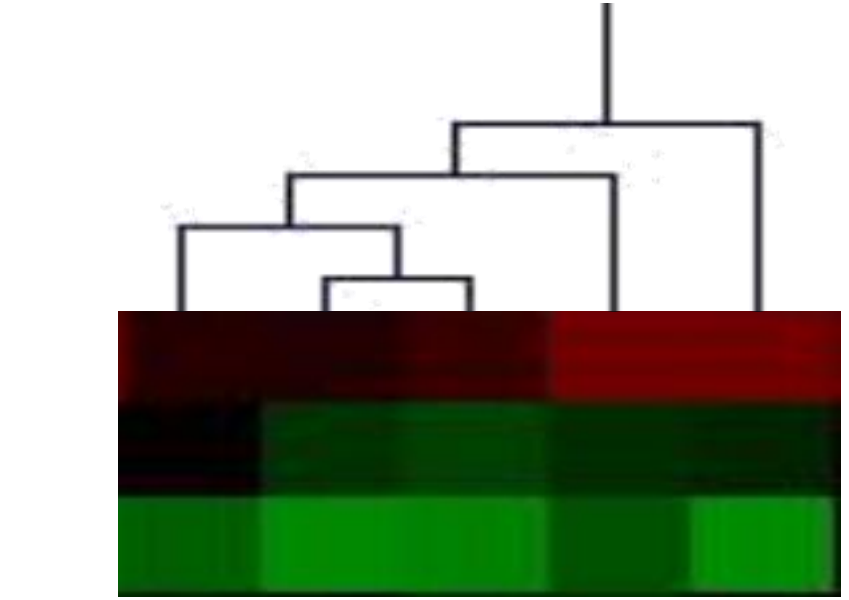
Download Cytoscape XGMML file for offline use



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

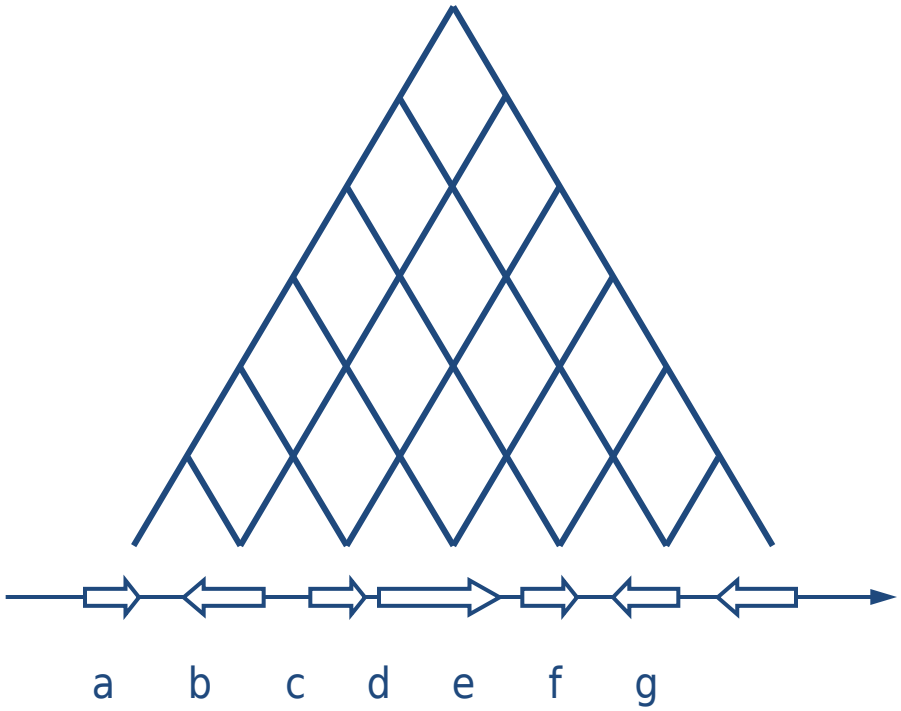


(((a (b c)) d) e)

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

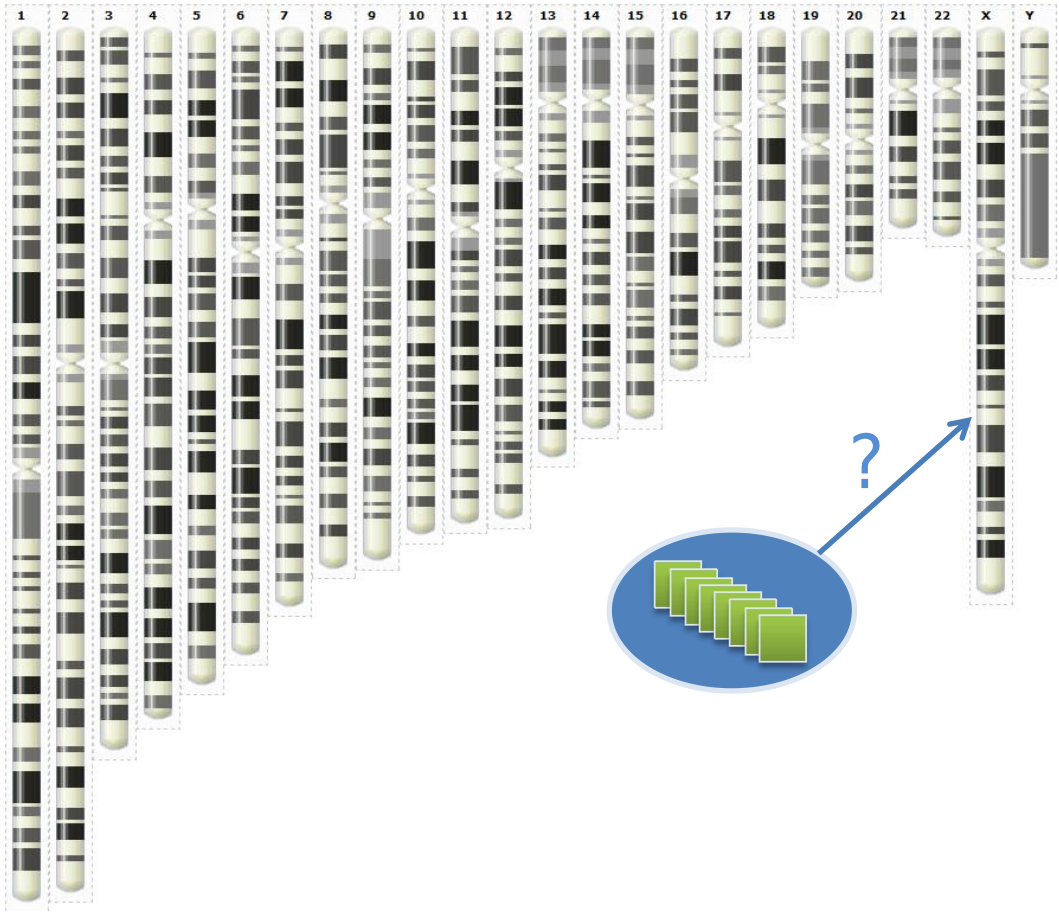


implicit

Set of genes of interest

Examples

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



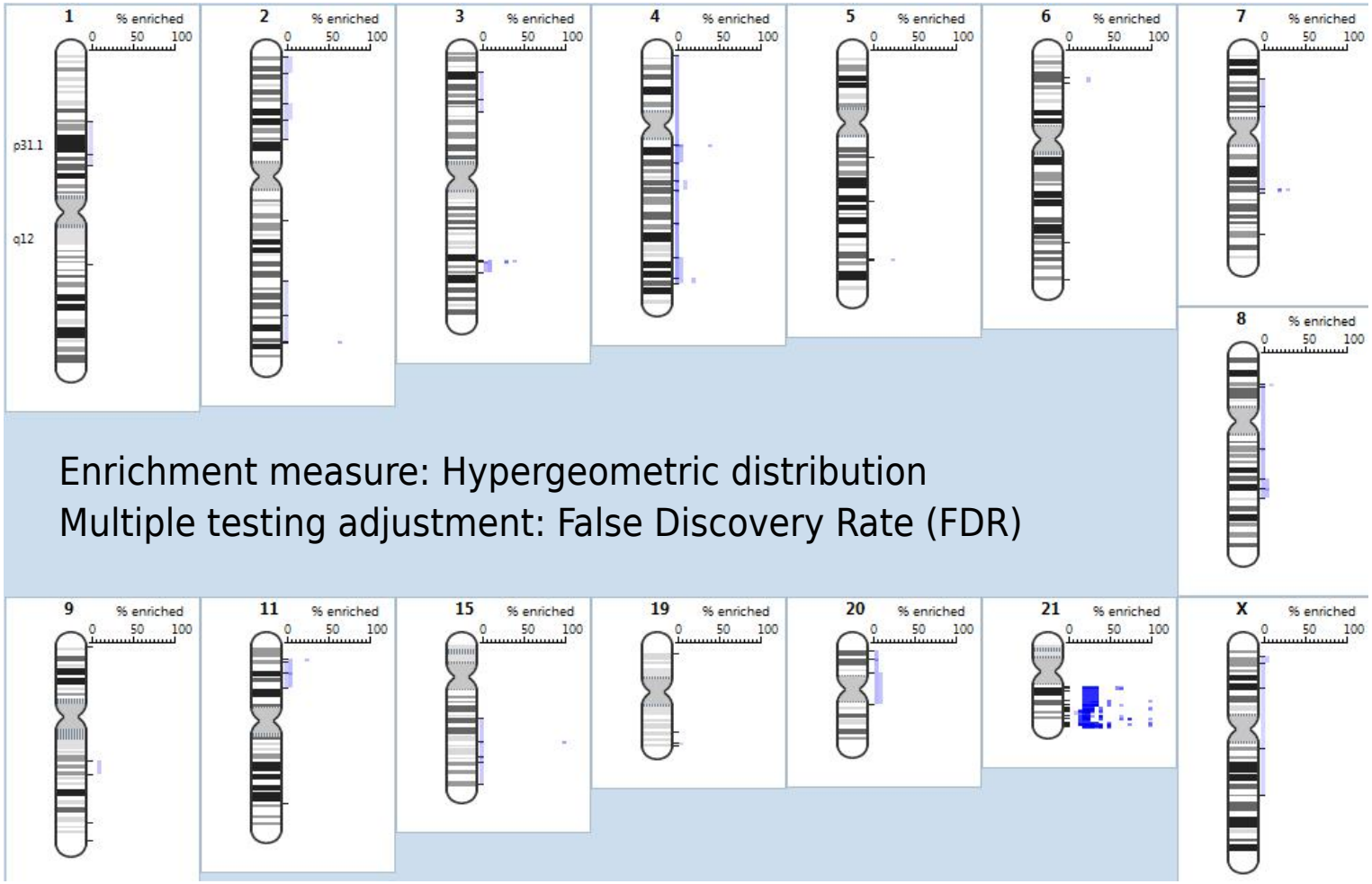
→ Question: Do those genes surprisingly cluster in the genome?

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.



Enrichment measure: Hypergeometric distribution

Multiple testing adjustment: False Discovery Rate (FDR)

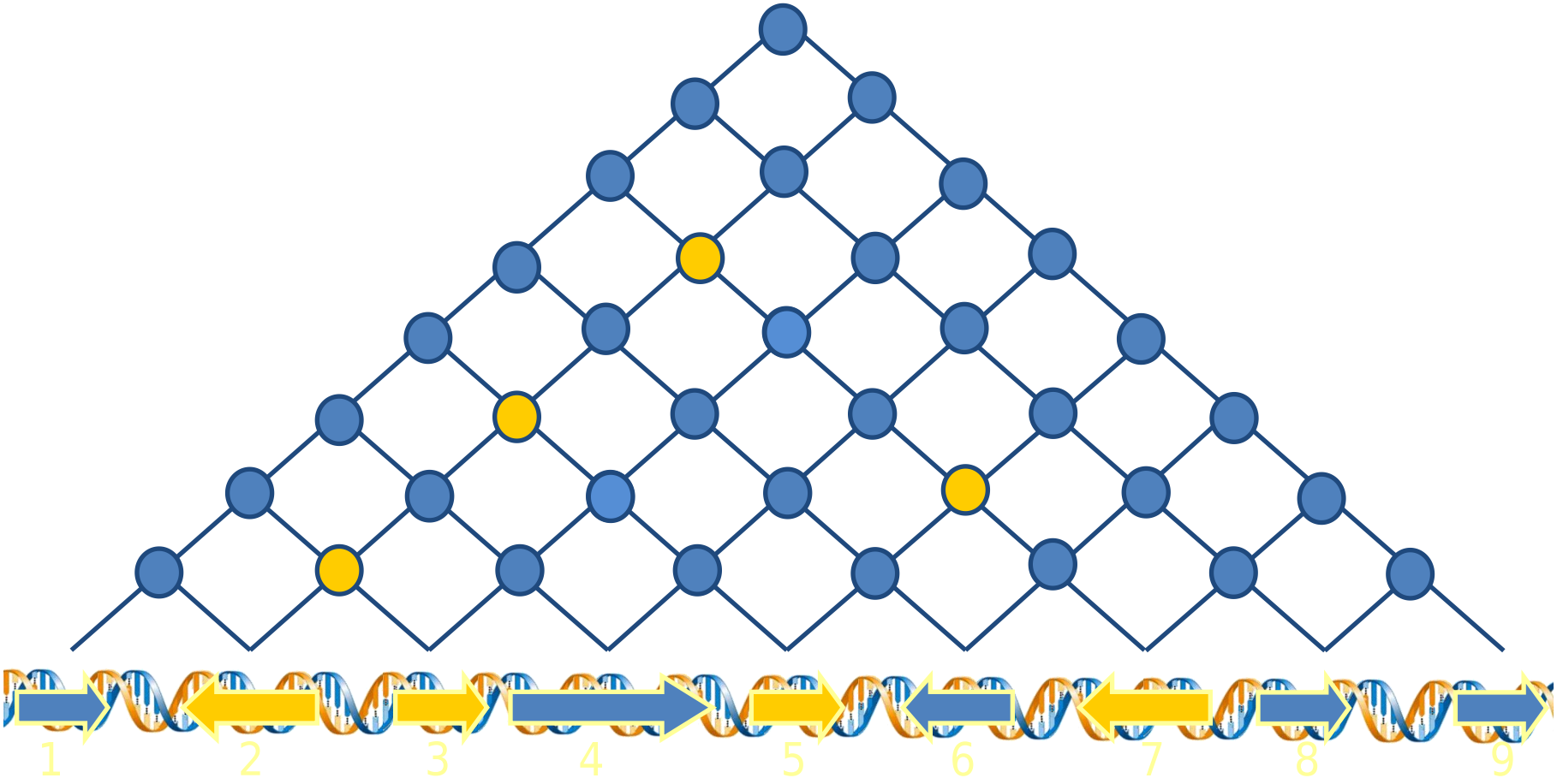
Issues:

- Number of regions to test
- False positives
- Redundancy

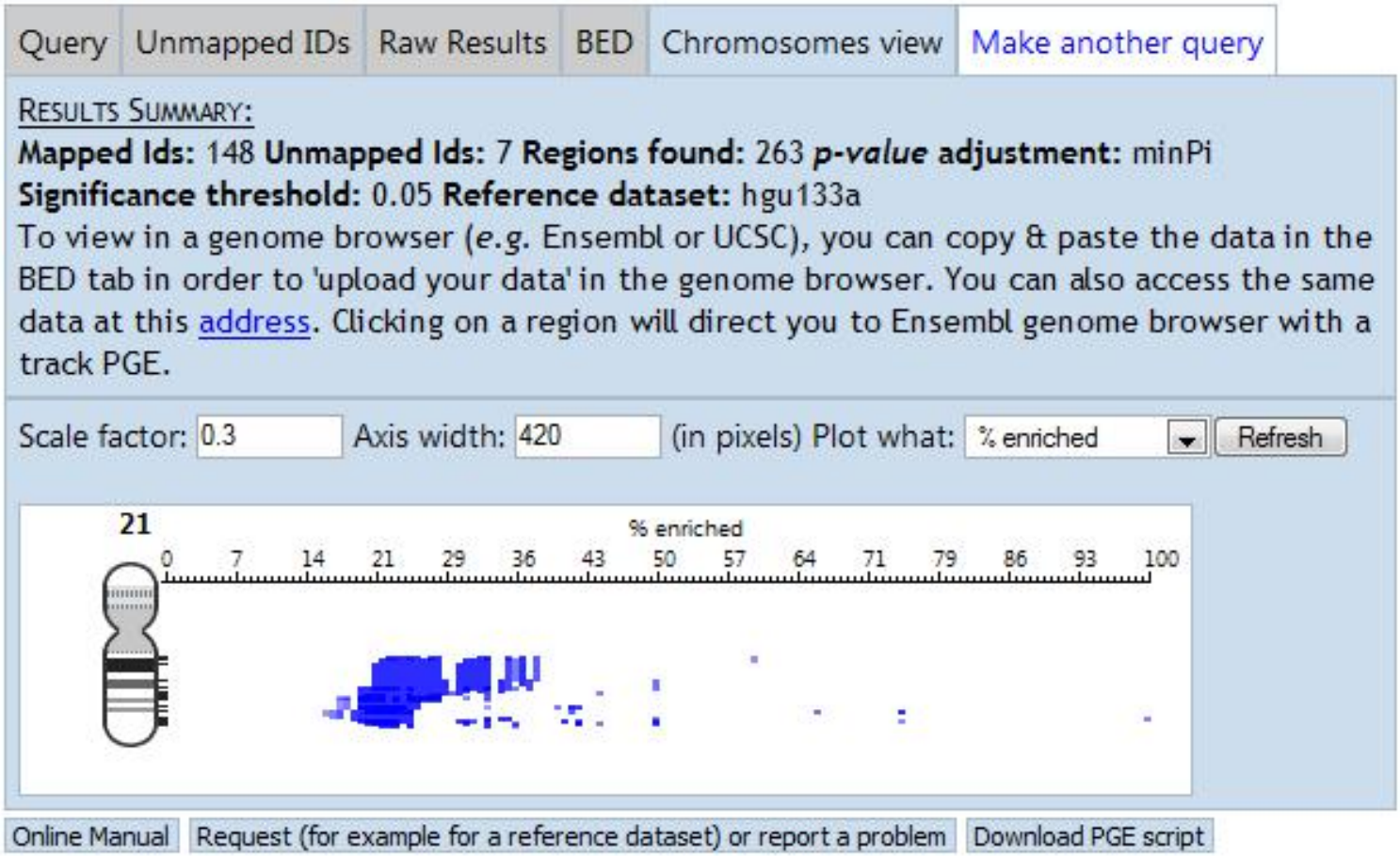
Pertinent Regions

A region is pertinent if it is:

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



Down Syndrome (minP_i)



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

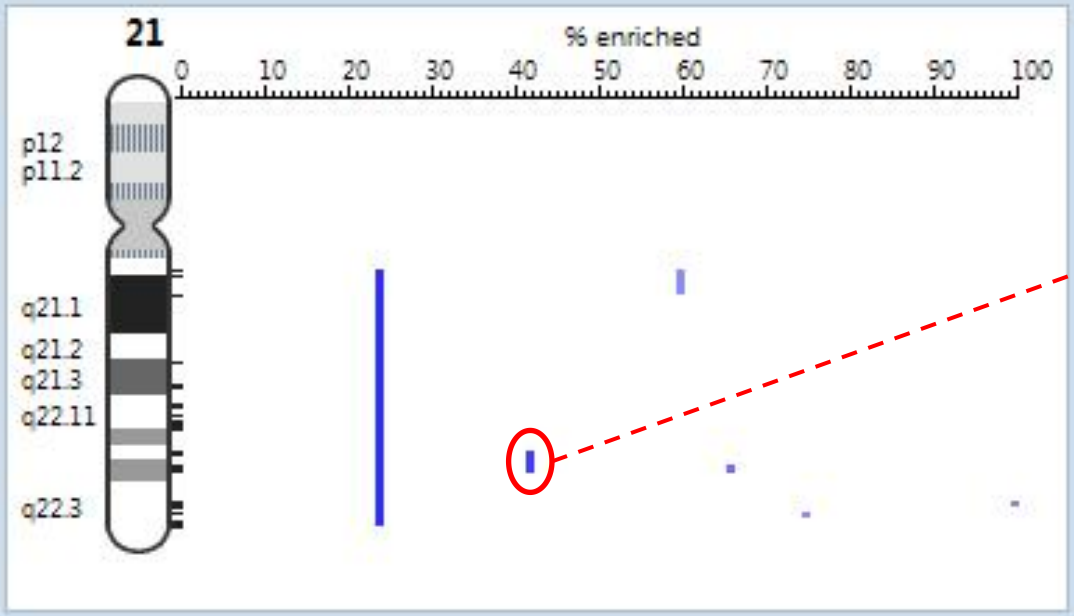
RESULTS SUMMARY:

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

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

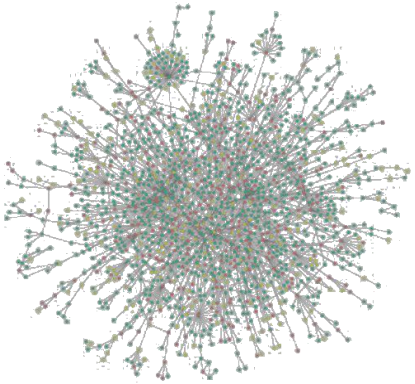
Scale factor: 1 Axis width: 300 (in pixels) Plot what: % enriched

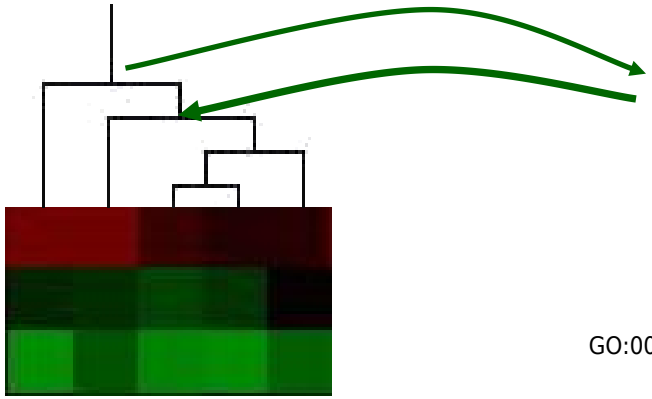
Refresh

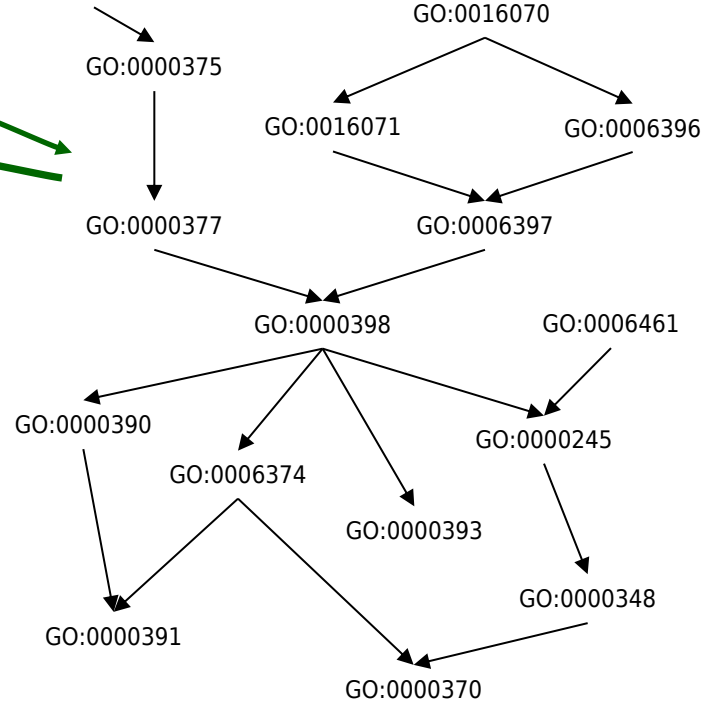


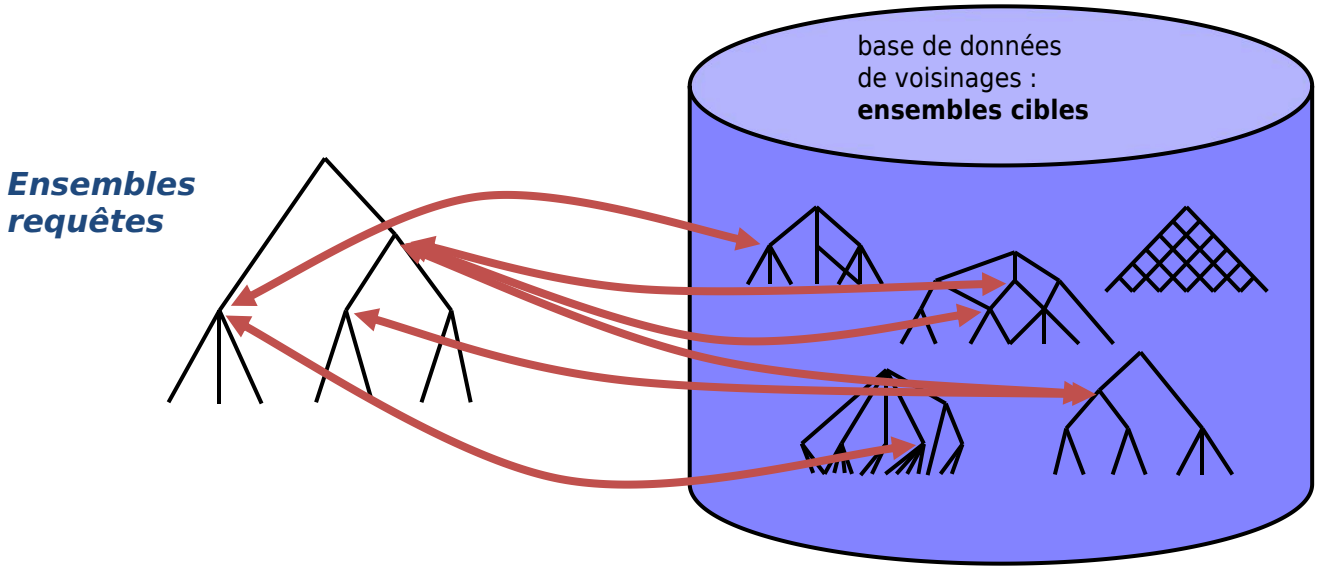
p-value: 7.87E-10
$p\text{-value}_{adj}$: <0.002
Score: 91.039
$Score_{adj}$: INF
Common elements: 6 / 14 (43%)
Overlaping regions: 1
Start of region: 37 359 546 bp
End of region: 40 223 183 bp
Genes:
DSCR2, DYRK1A, PCP4, PIGP, SH3BGR
WRB

Défis actuels

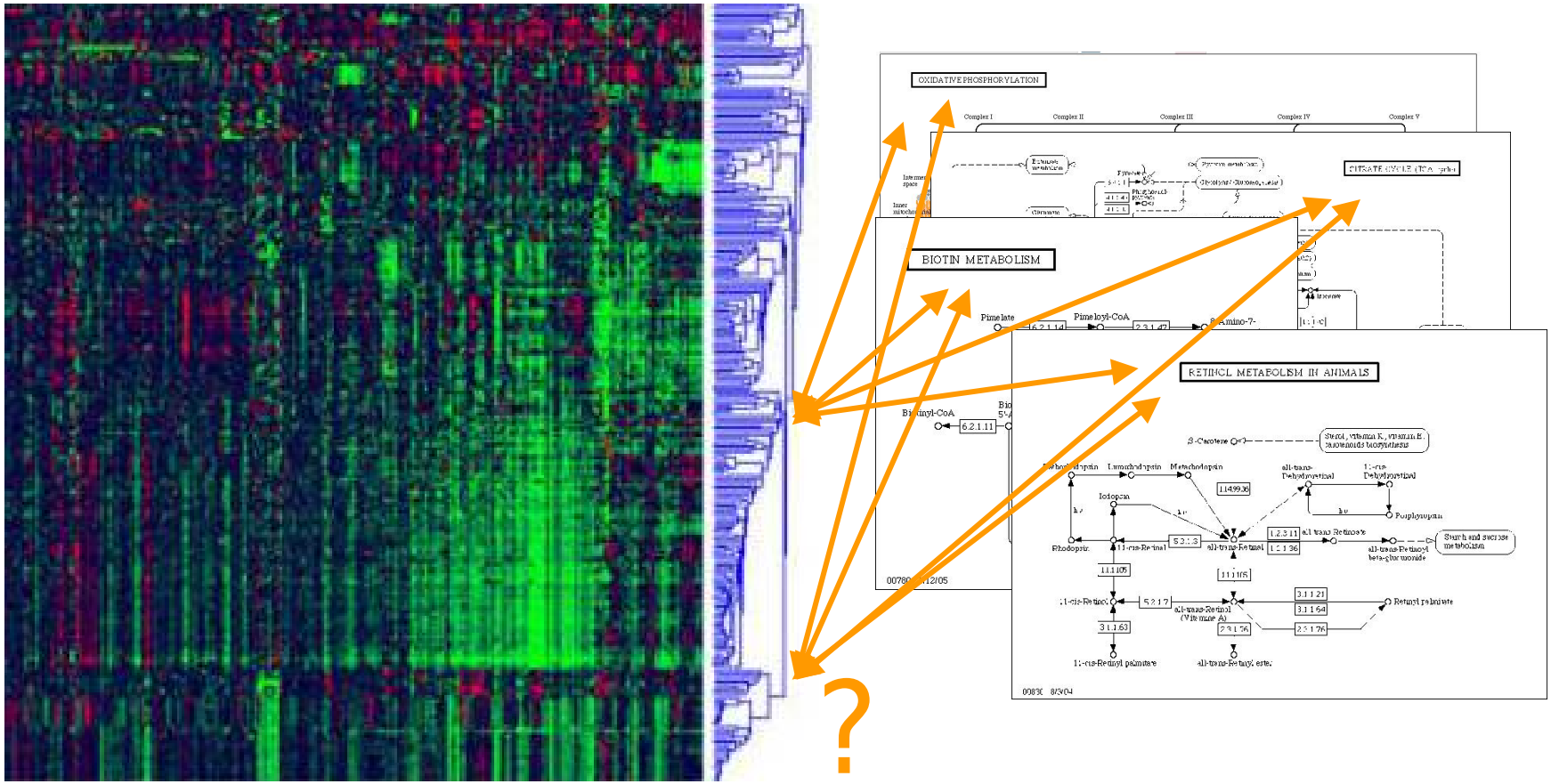








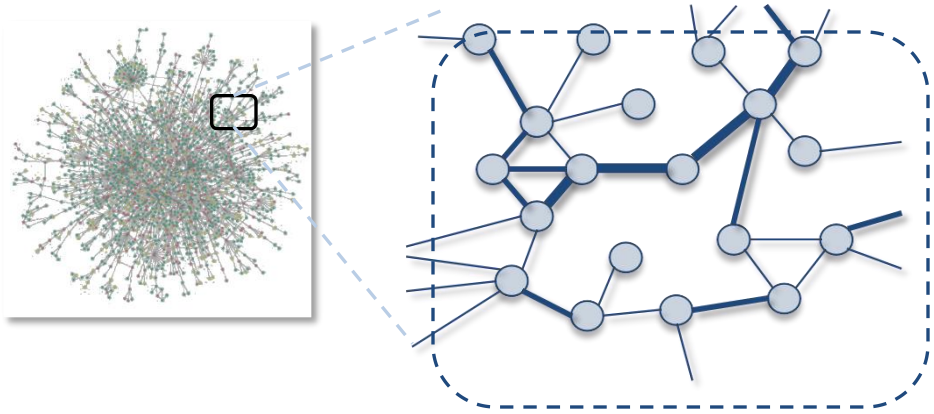
Analyse de données d'expression



[Ferea *et al.*, 1999]

[Kanehisa & Goto, 2000]

Extraction de sous-graphe pertinent & visualisation



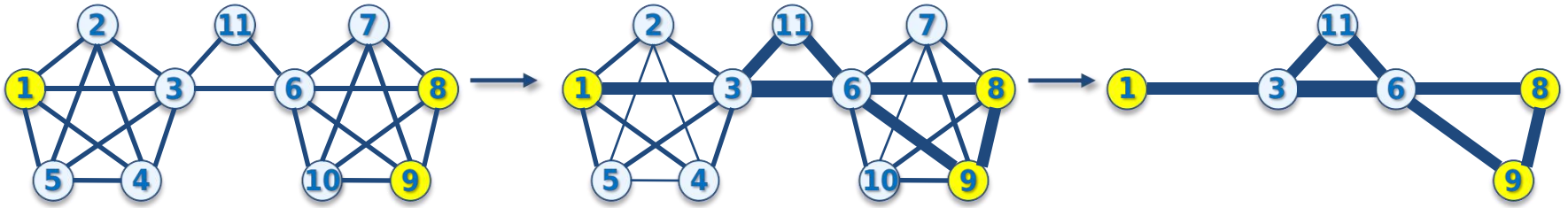
Idée :

- 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

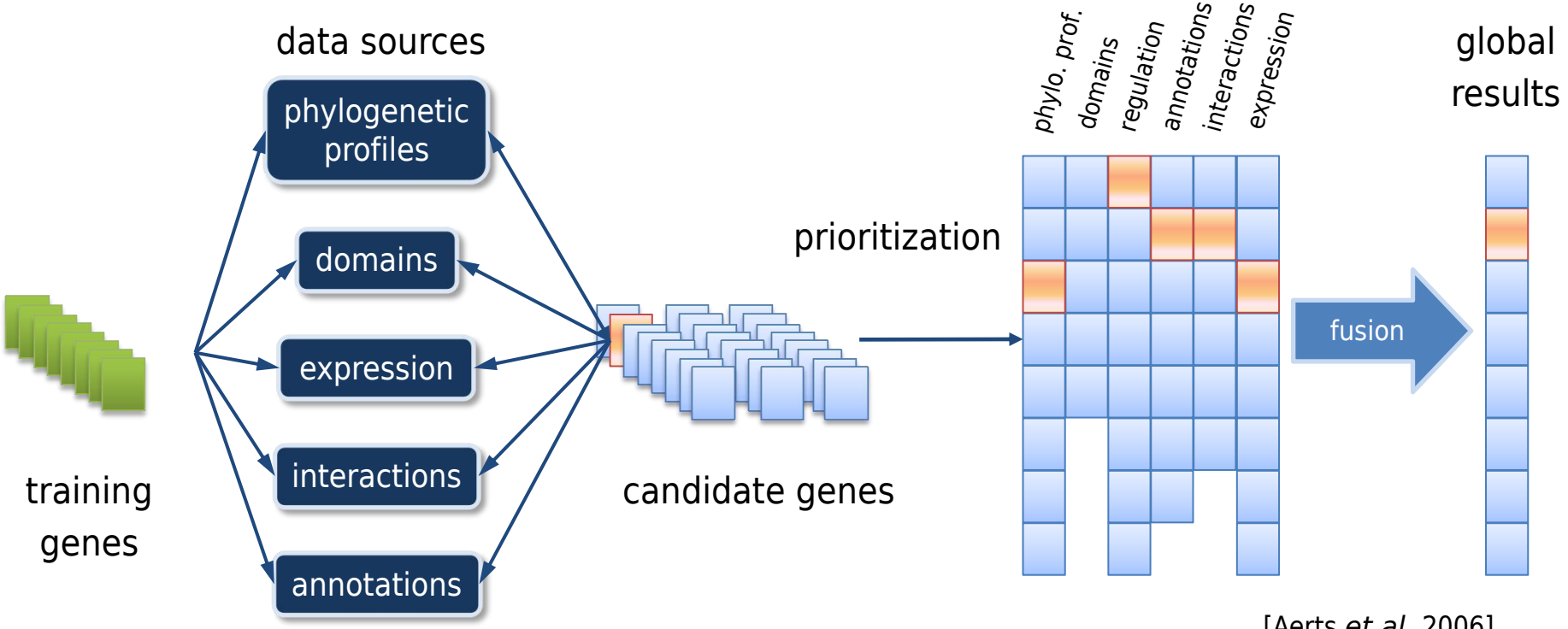
Intégration de données hétérogènes
Priorisation

Master 2

Bioinformatique et Biologie des Systèmes

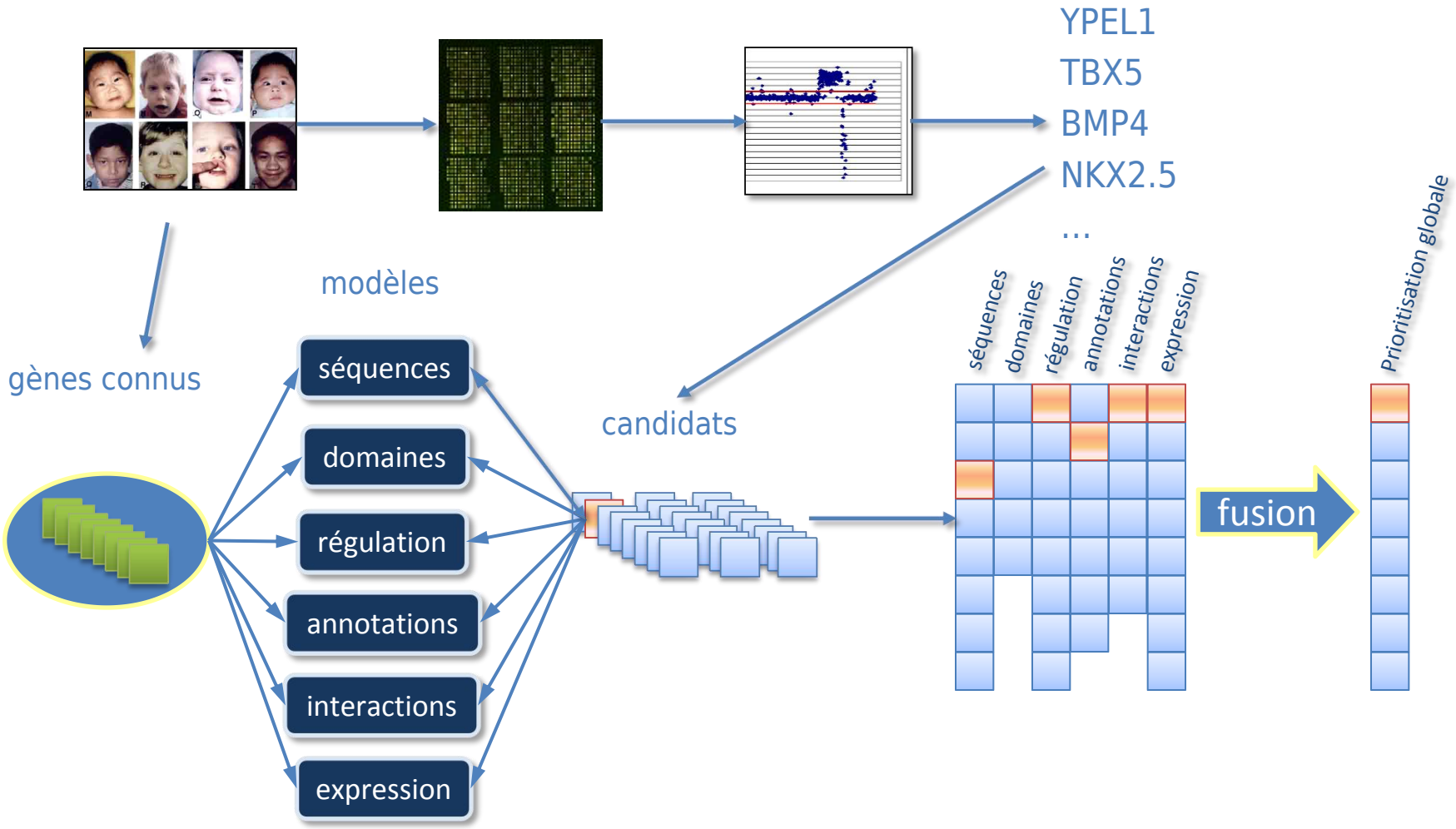
Candidate gene prioritization by genomic data fusion

- 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]

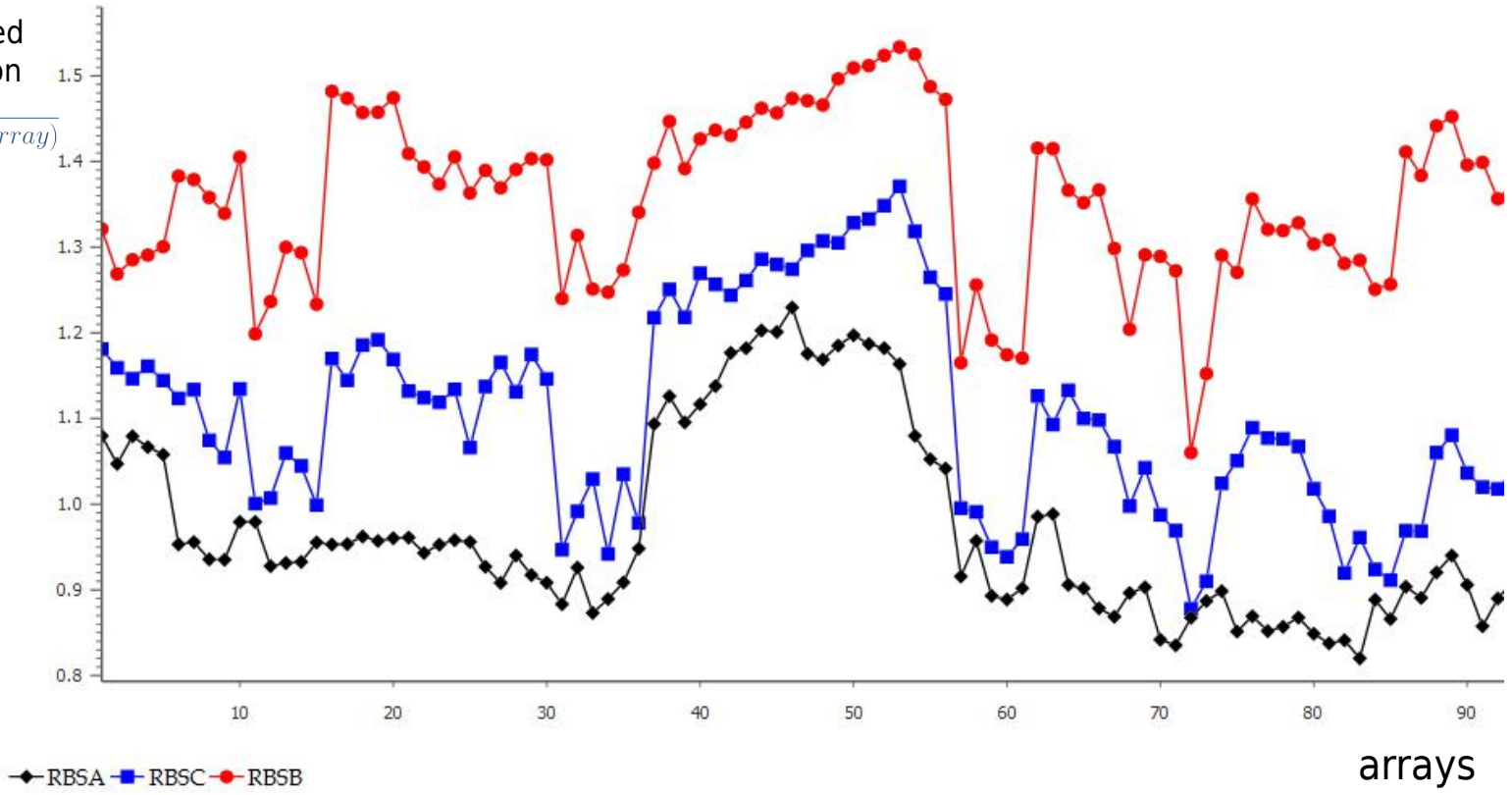
Exploitation des données disponibles



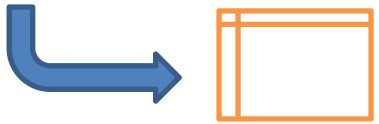
Gene expression

normalized expression

$$y' = \frac{y}{median(array)}$$



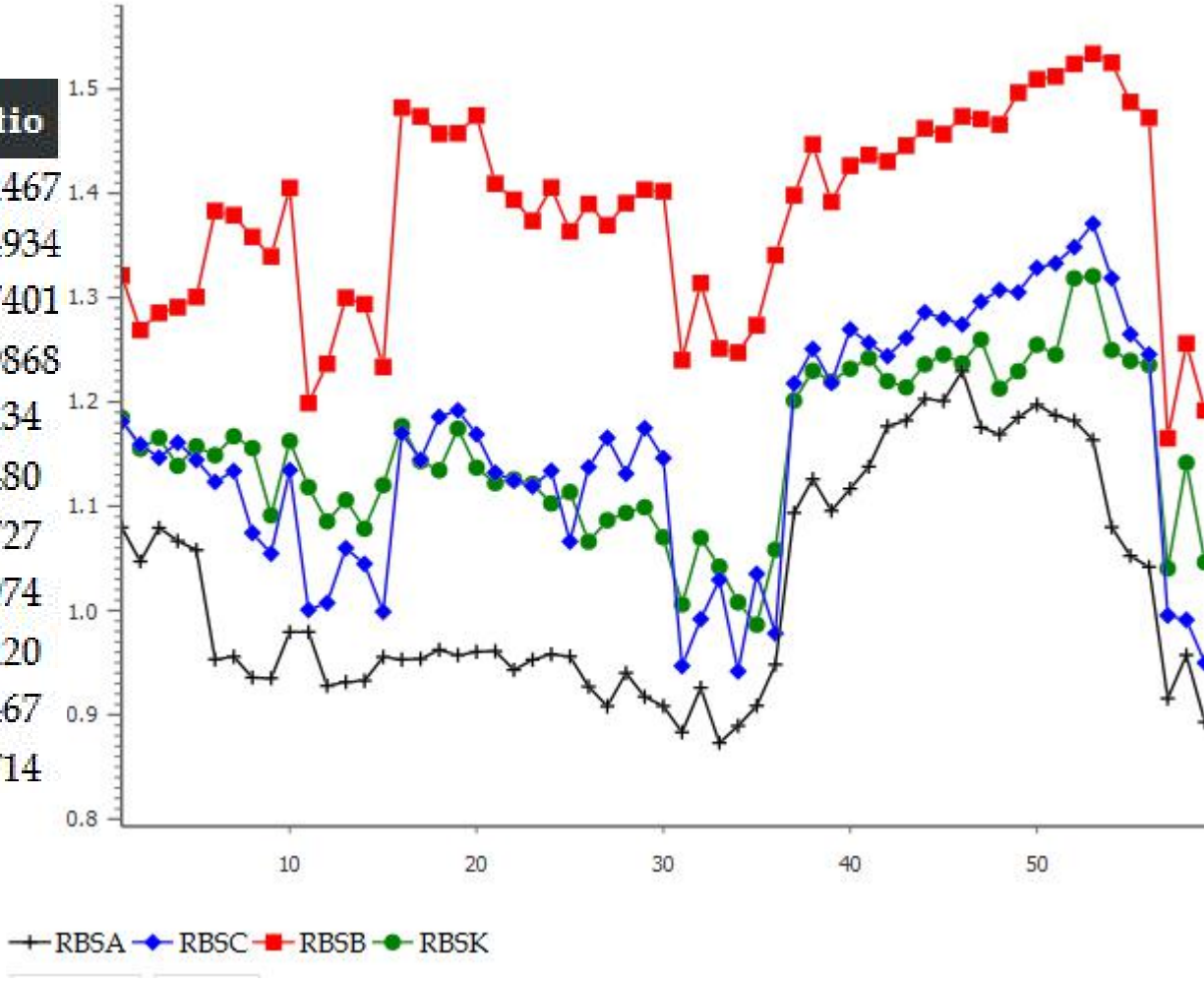
- a gene: set of expression values in various experimental conditions
- a pair of genes: dissimilarity index based on Pearson's correlation coefficient
- score : average dissimilarity



gene pairwise dissimilarity matrix

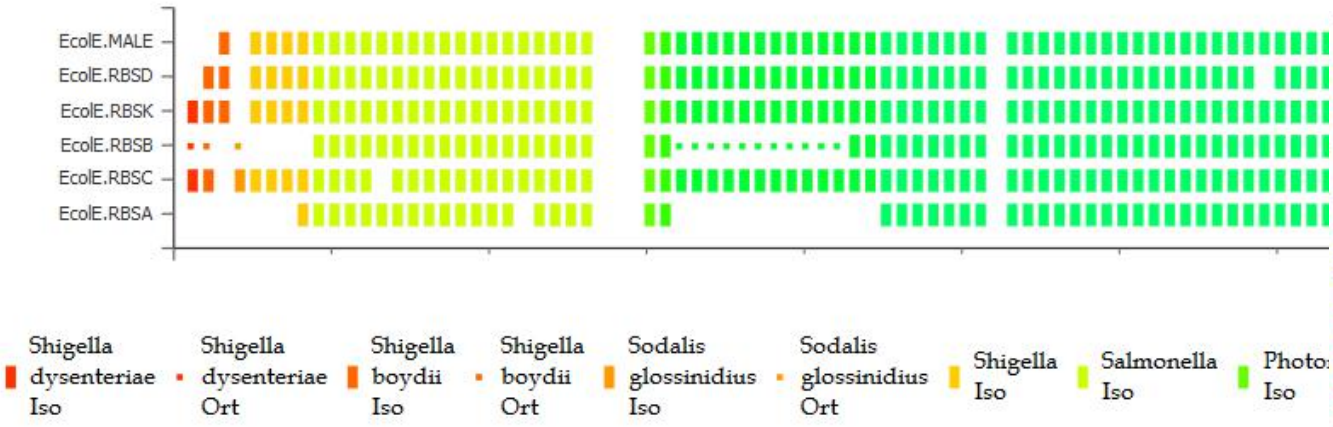
- training: rbsA, rbsB, rbsC in *E. coli* K-12

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



Phylogenetic profiles

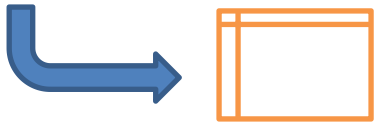
training: rbsA, rbsB, rbsC



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

sinia

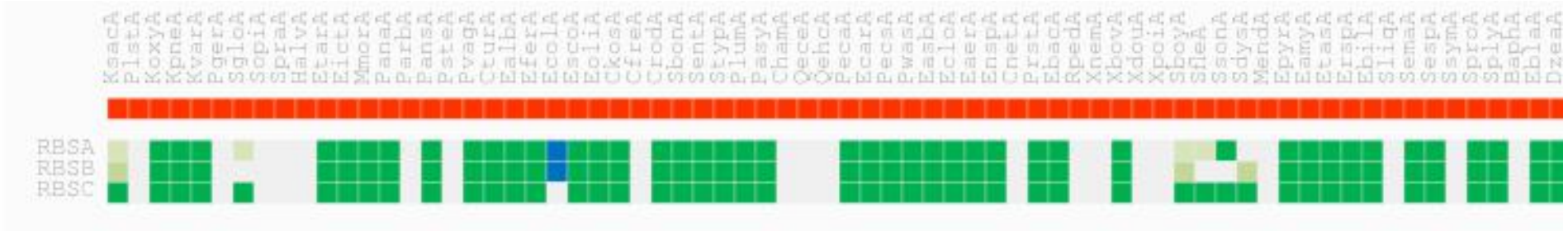
- a gene: presence/absence of orthologs 1:1 in other genomes
- pair of genes: dissimilarity index based on the Jaccard index
- score: average dissimilarity



gene pairwise dissimilarity matrix

Phylogenetic data

- Phylogenetic profiles



- gene pairwise distance matrix computation
  - Hypothesis: genes located near each other in a set of genomes are likely to be functionaly related
  - g_1' and g_2' orthologs 1:1 of $gene_1$ and $gene_2$ in another genome *i*
  - Probability that the distance D_i is smaller than the observed distance d_i

$$p_i = Pr(D_i \leq d_i) = \frac{2d_i}{N_i - 1}$$

  - For a set of *M* other genomes

$$d = Pr(D_1 \leq d_1, \ldots, D_M \leq d_M) = \prod_{i=1}^{M} p_i$$

  - *M* depends on the pair of genes considered
  - *d* not comparable between genes (*e.g.* $0.1^6=10^{-6}$ *vs.* $0.5^{20}=9.5\ 10^{-7}$)
  - normalization: log transformation, z-score, average of distance matrix and its transpose

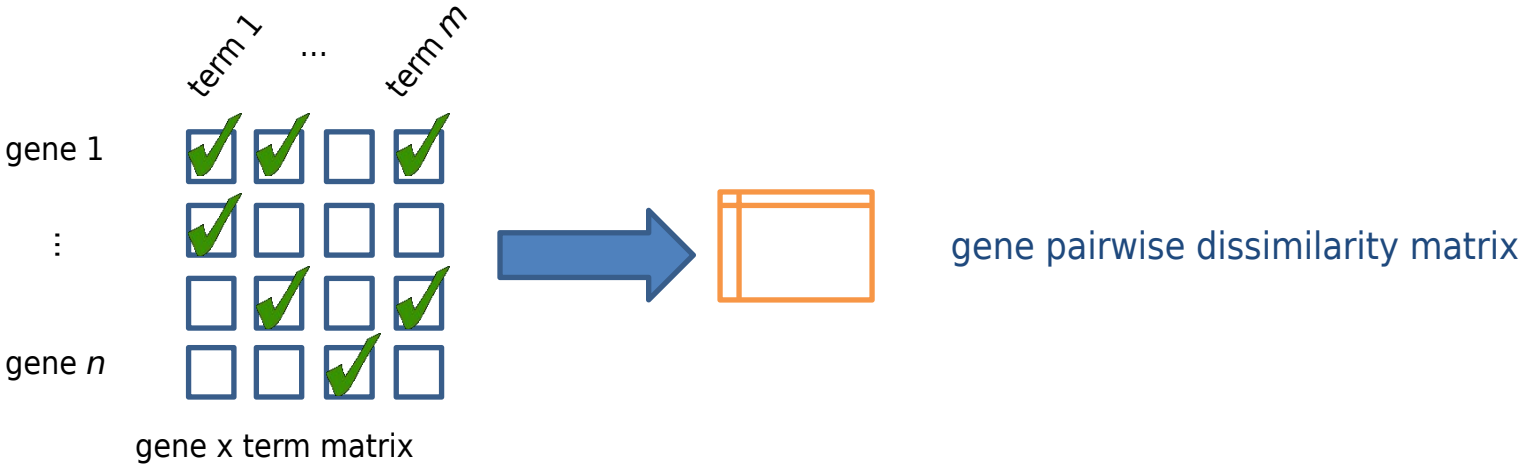
Phylogenetic data: genome selection

- Reference genomes should not be too evolutionary close to the genome of interest
- Reference genomes should not be redundant in order not to introduce biases
- Need to estimate the relevance of a genome with respect to
  - A genome of interest: is it not too closely related? is it informative?
  - A set of already selected reference genomes: redundancy *vs.* additional signal
- Parameters
  - Rearrangements
    - Significance of genes proximity on the chromosome
  - Core genome size
    - Maximize the coverage of the genome of interest

Approaches:

- gene-term matrix: distance between rows
  - manhattan/euclidean, Jaccard, ...

  but: same weight for each GO-term
- based on GO-term similarity
  - adapt weight to information content



- Node/term information content

$$IC(term) = -\log p(term)$$ with p(term)= freq(term)

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

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

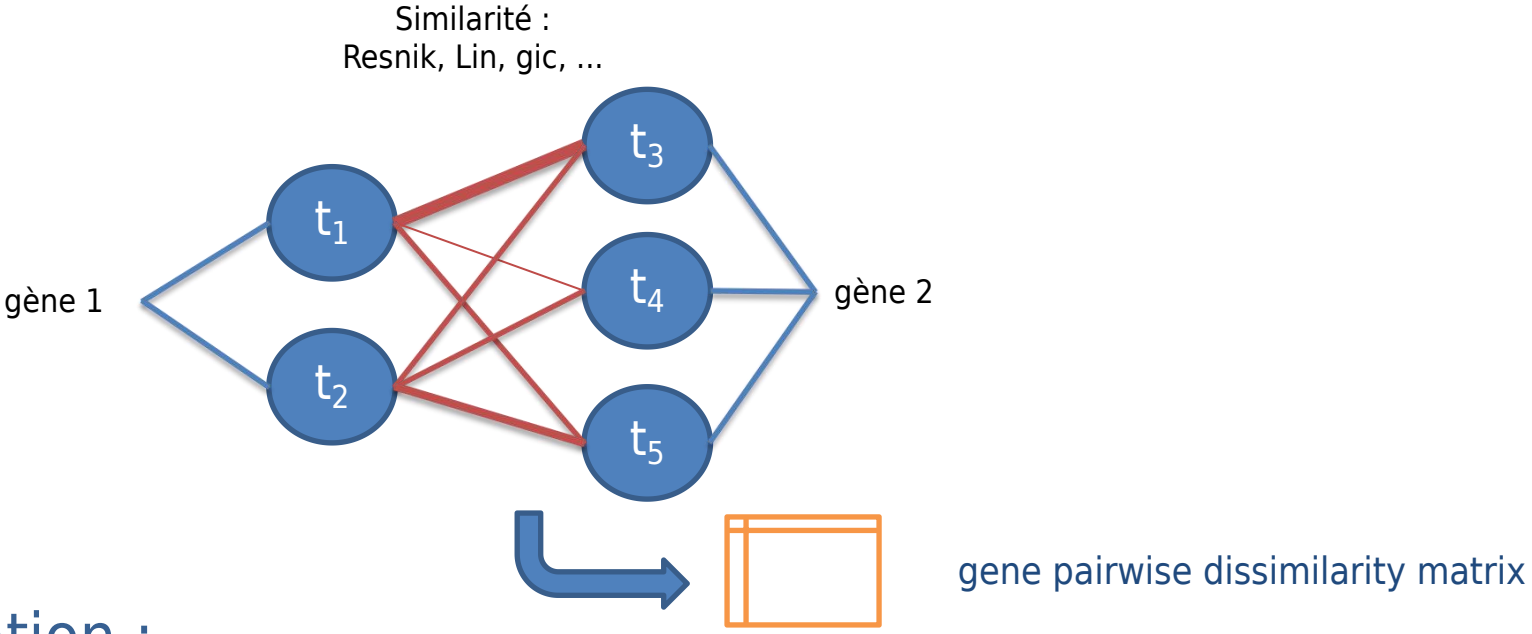
- $sim_{res}(t_1,t_2) = IC(MICA(t_1,t_2))$ [Resnik, 1995]

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

Possibilités :

- Similarité moyenne des termes communs au 2 gènes
- Similarité maximale, ex : t_1-t_3
- Best Match Average (bma), ex : ave(t_1-t_3, t_2-t_5)

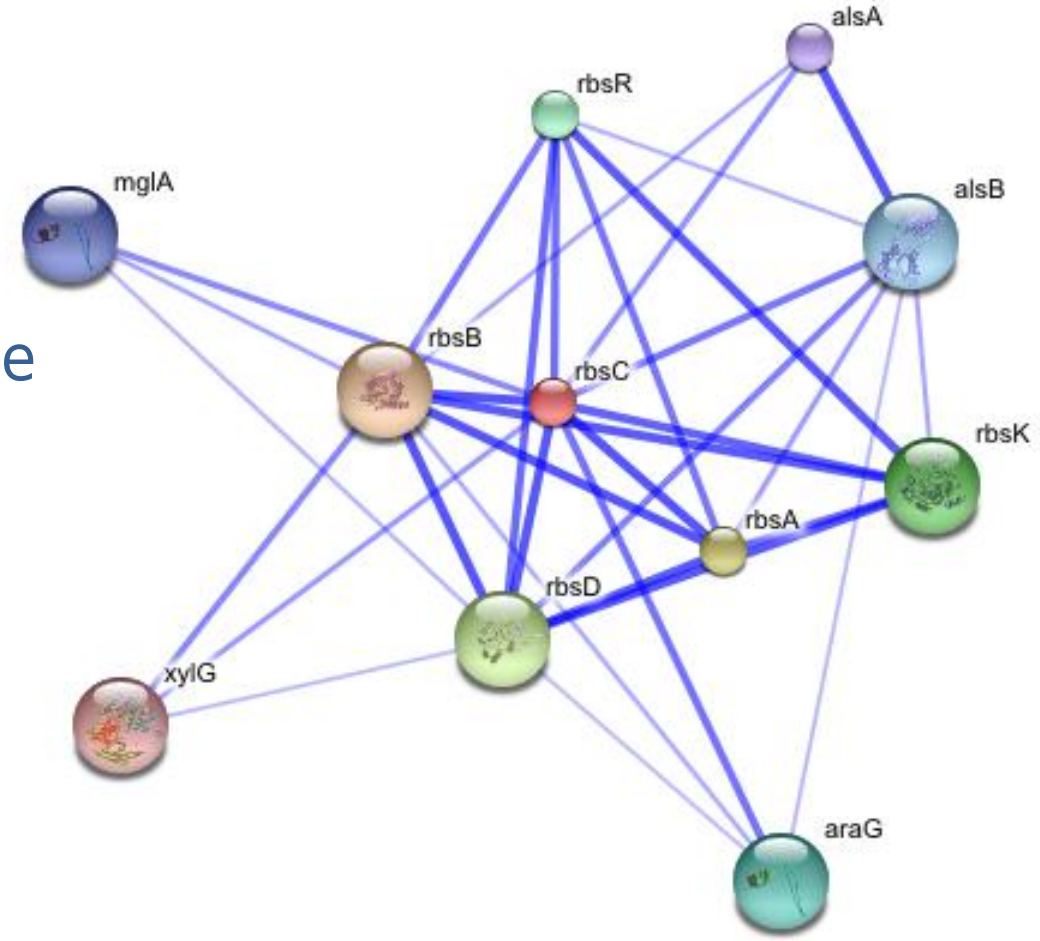


Application :

- Performances légèrement meilleures obtenues que les autres avec la combinaison Resnik + similarité maximale
- à confirmer sur d'autres jeux de données ou d'autres contextes

Interactions

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



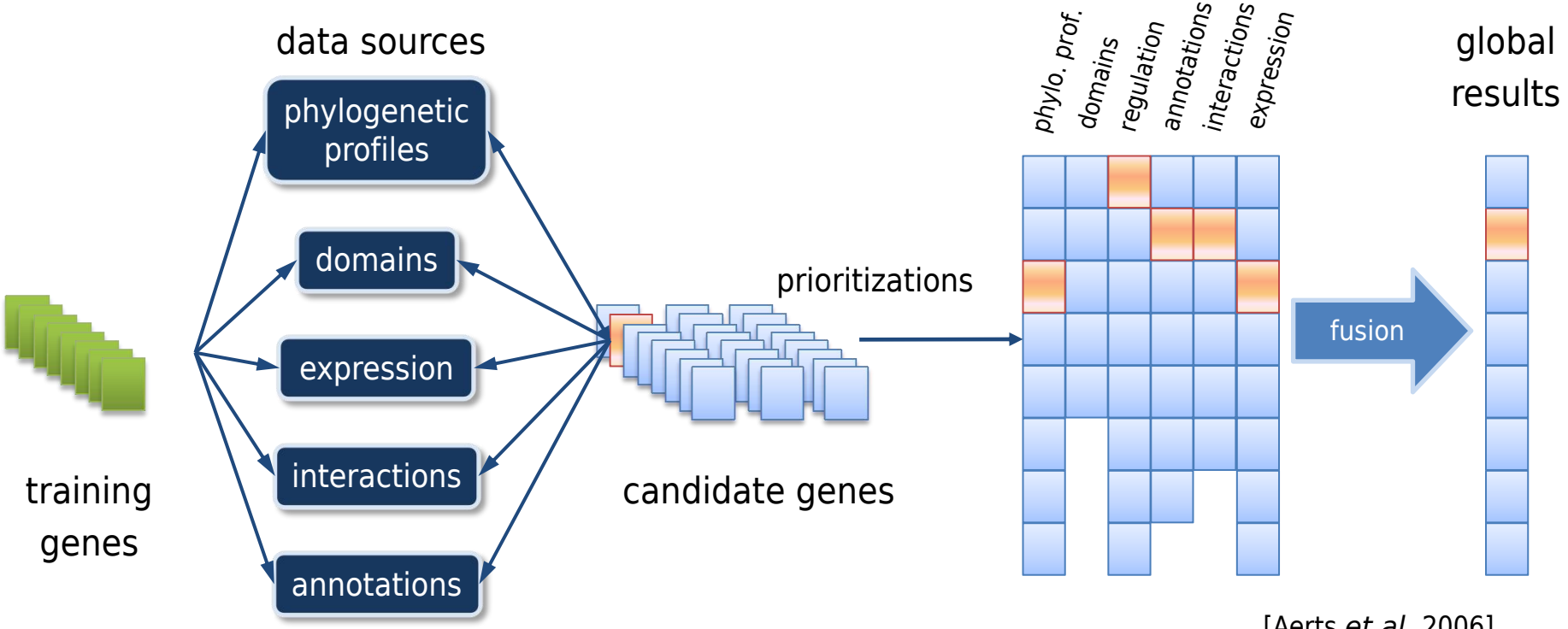
from STRING
http://string-db.org

training: rbsA, rbsB, rbsC

candidate	score	rank	rank ratio
RBSK	1.000	2	0.0005136
RBSD	1.000	2	0.0005136
RBSR	1.000	2	0.0005136
ALSB	1.333	5	0.001284
ALSC	1.333	5	0.001284
YPHD	1.333	5	0.001284
MGLC	1.667	10.5	0.002696
XYLG	1.667	10.5	0.002696
ALSA	1.667	10.5	0.002696
YTFT	1.667	10.5	0.002696

Candidate gene prioritization by genomic data fusion

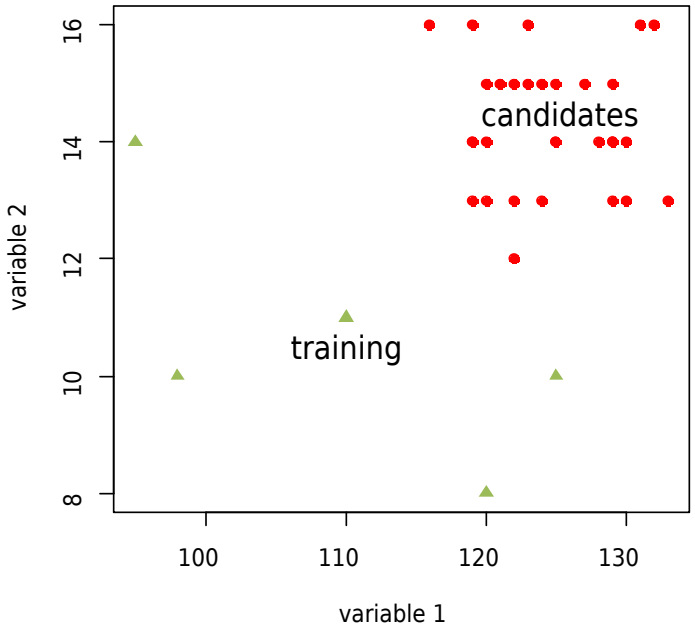
- Observation: more and more post-genomic data available
- Paradox: more difficult to select the best candidates
- Goal: objective and comprehensive evaluation of candidates



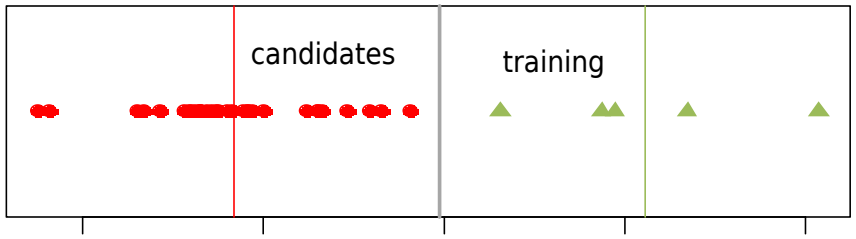
[Aerts *et al.* 2006]

Weighted fusion through linear discriminant analysis

- Principles
  - 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)

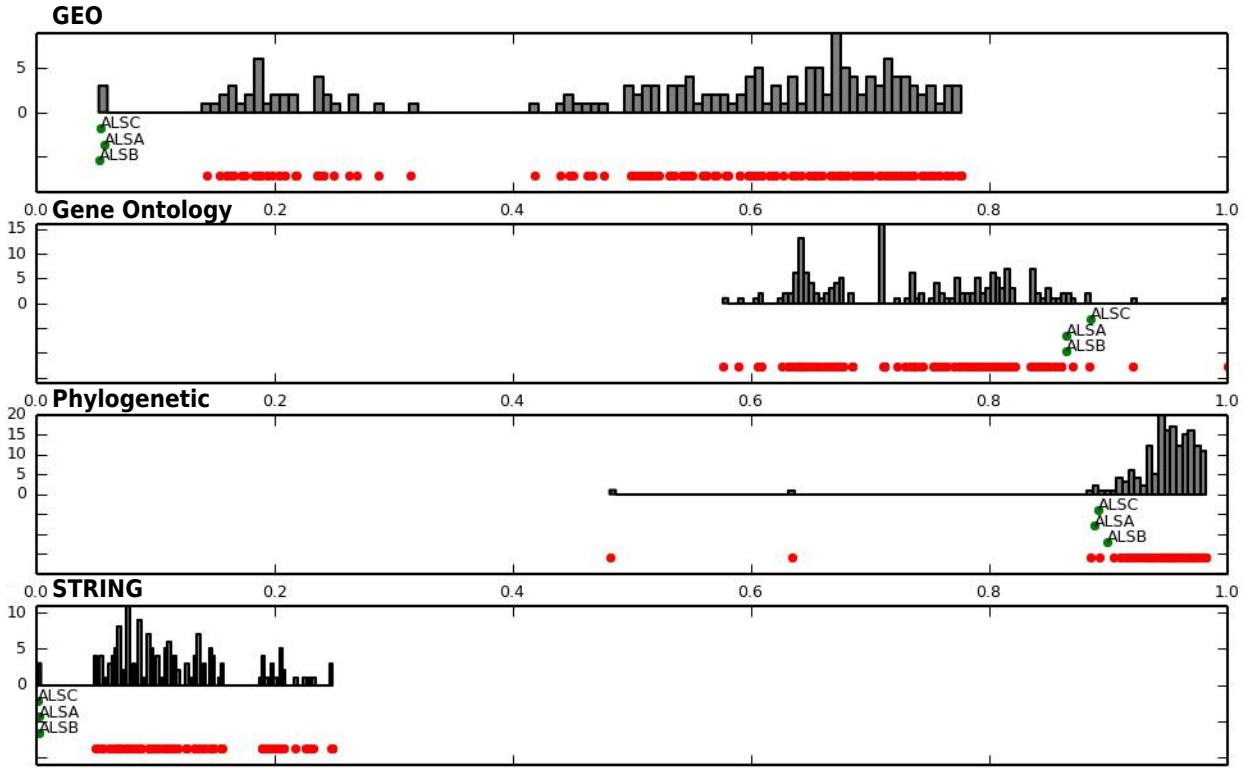


dimension	weight
variable 1	-0.1307346
variable 2	-0.7031850



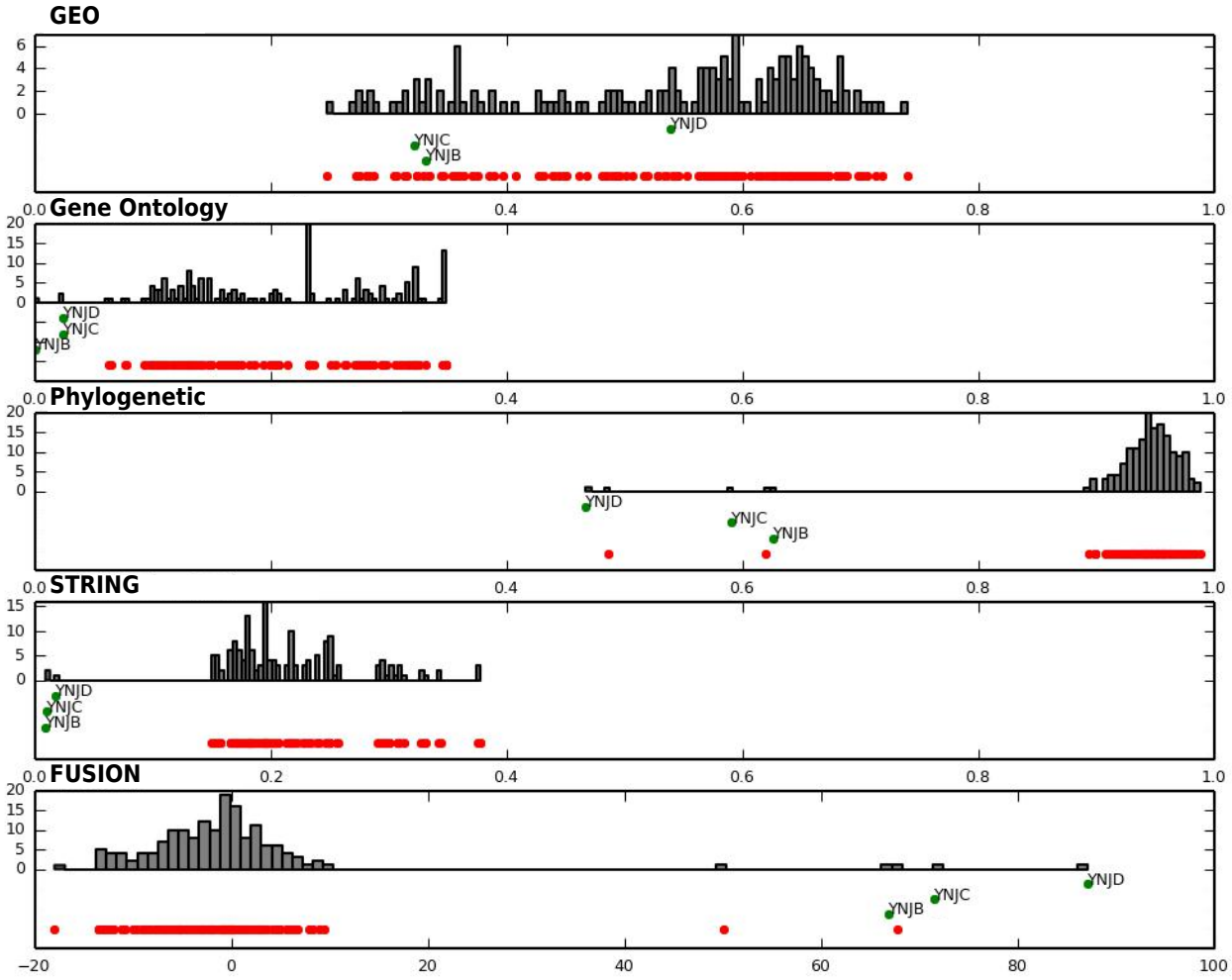
Application to *E. coli* alsA system: *alsA, alsB, alsC*

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



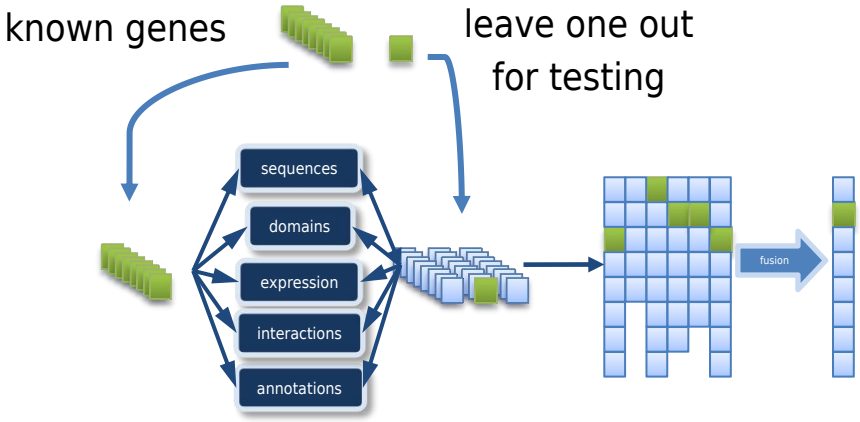
Application to *E. coli* ynjD system: *ynjB, ynjC, ynjD*

Data source	Weight
Expression (geo)	1.3
Annotations (Gene Ontology)	3.4
Phylogenetic	17.4
Interactions (string)	6.6



Evaluation methodology

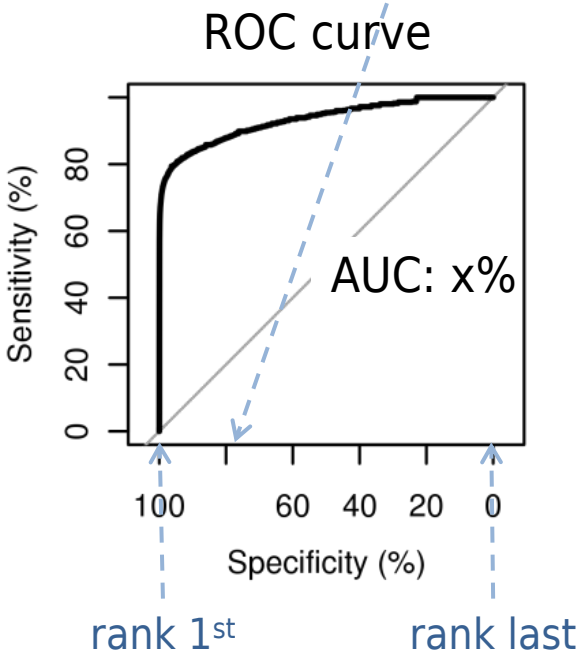
- Leave-one-out cross validation (LOOCV)



- 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)

How well does it rank?

e.g. rank ratio = 2/8 = 0.25



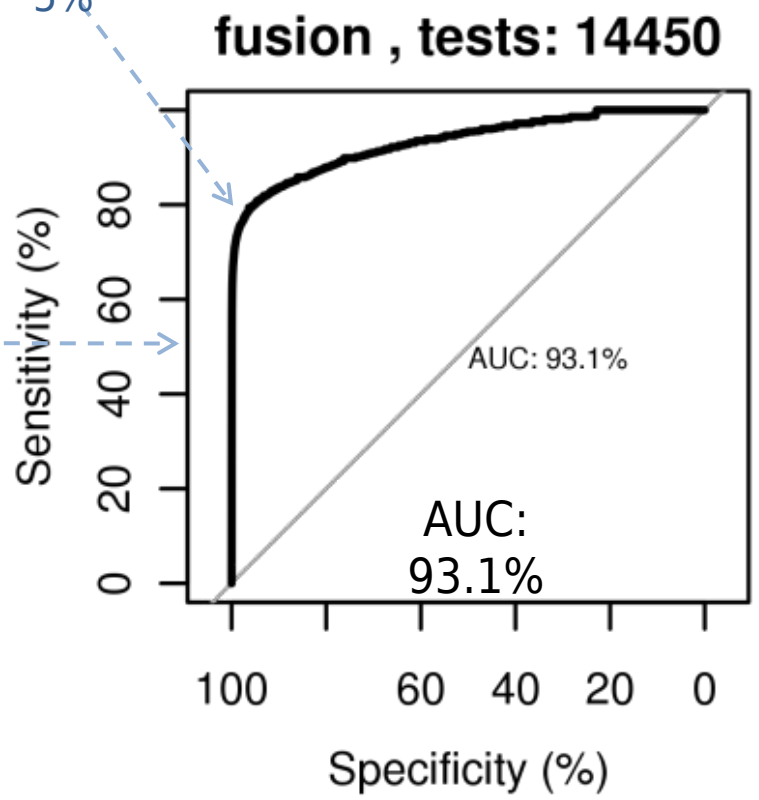
Evaluation: global results

Gold standard

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

53% of the left out genes rank 1st

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



Prioritization for functional inference

Organism **Hide** **Hide**

Organism	**Escherichia coli** (strain K12)
External Links	[UNIPROT] [NCBI]
Taxonomic Lineage	> Bacteria > Proteobacteria > Gammaproteobacteria > Enterobacteriales > Enterobacteriaceae > Escherichia > Escherichia coli > EcolE
Strain Name	K12
ABCdb identifier	EcolE
Chromosomes	EcolE01

Assembly **Hide** **Hide**

Assembly	NBD	MSD	SBP	Class
EcolE01.RBSB	EcolE01.RBSA	EcolE01.RBSC	EcolE01.RBSB	A_1a

Proteins **Hide** **Hide**

Protein	Domain	Subfamily	TCdb
EcolE01.RBSB	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))
EcolE01.RBSC	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))
EcolE01.RBSA	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 for functional inference

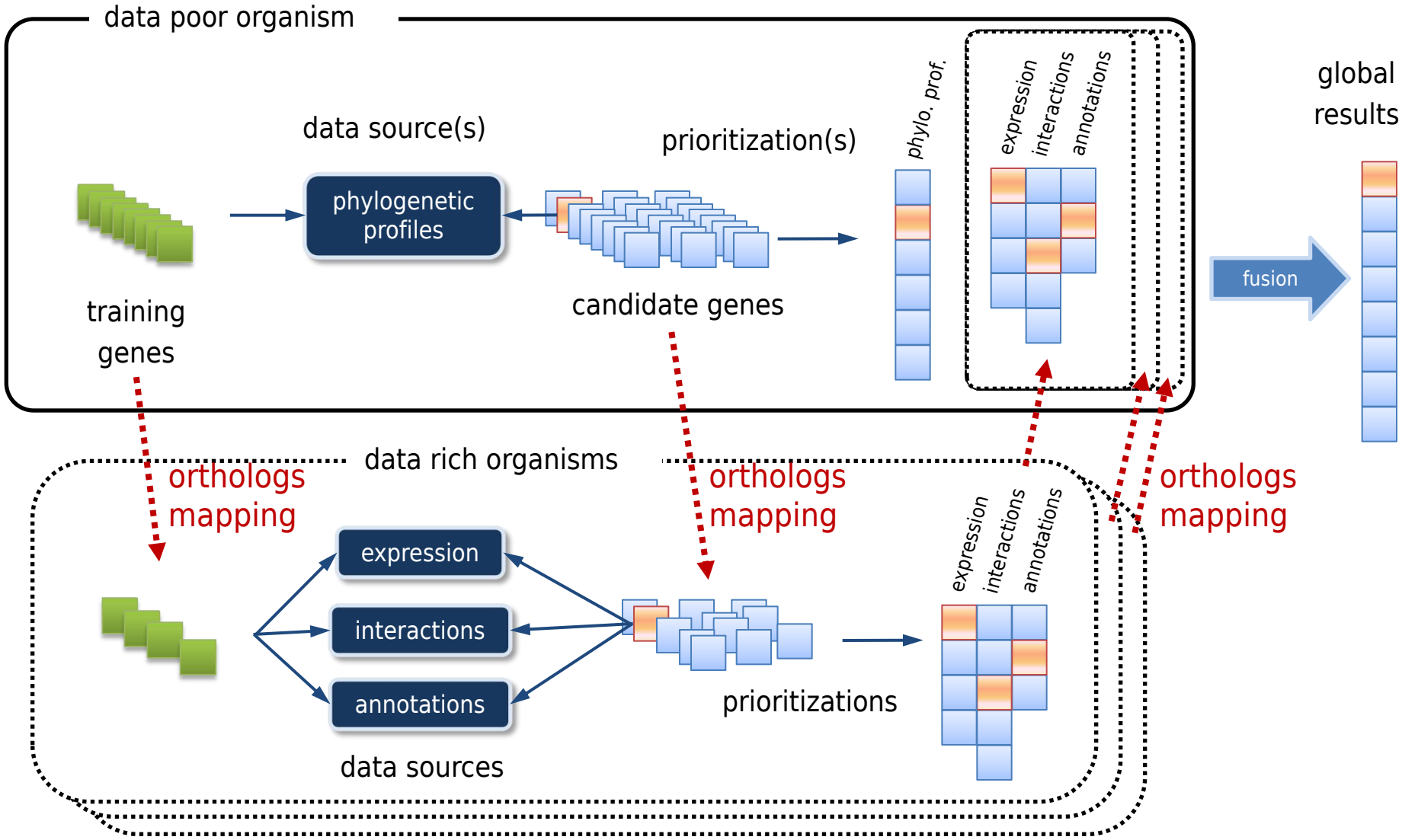
Prioritization **Hide** **Hide**

Run prioritization.

Show 10 entries Search:

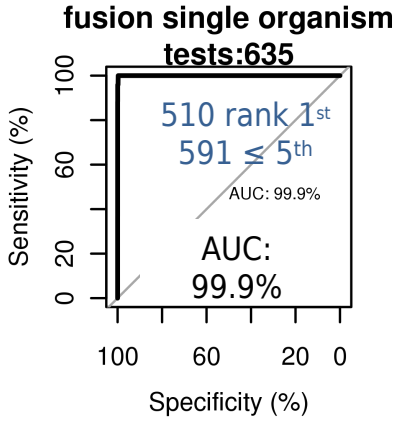
rank	Global results	pathways (fusion)	string (fusion)	transcriptome (fusion)	phylogenetic_profiles EcolE	go (fusion)	interactome EcolE
1	RBSD (1) S: 0, RR: 0	D-ribose pyranase					
2	RBSK (2) S: 0, RR: 0	Ribokinase					
3	MALE (3) S: 0, RR: 0.001	SBP of maltose/maltodextrin/maltoologisaccharide ABC transporter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

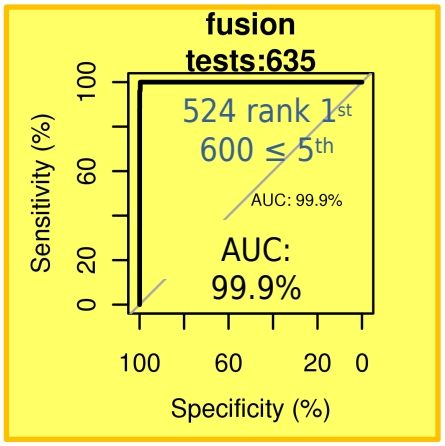
Extension of the method to data poor organisms

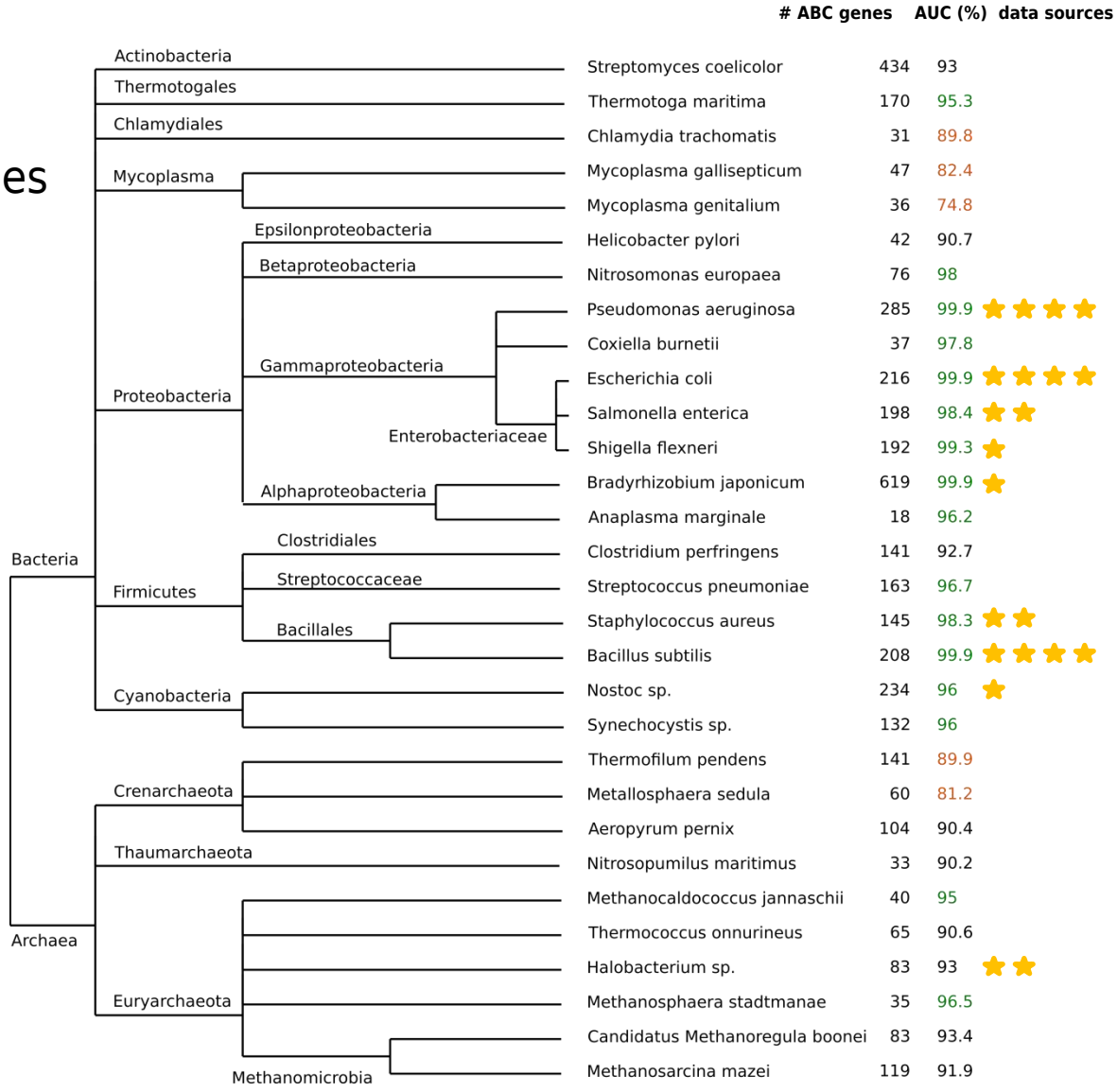


Performances: using other organisms data through orthology

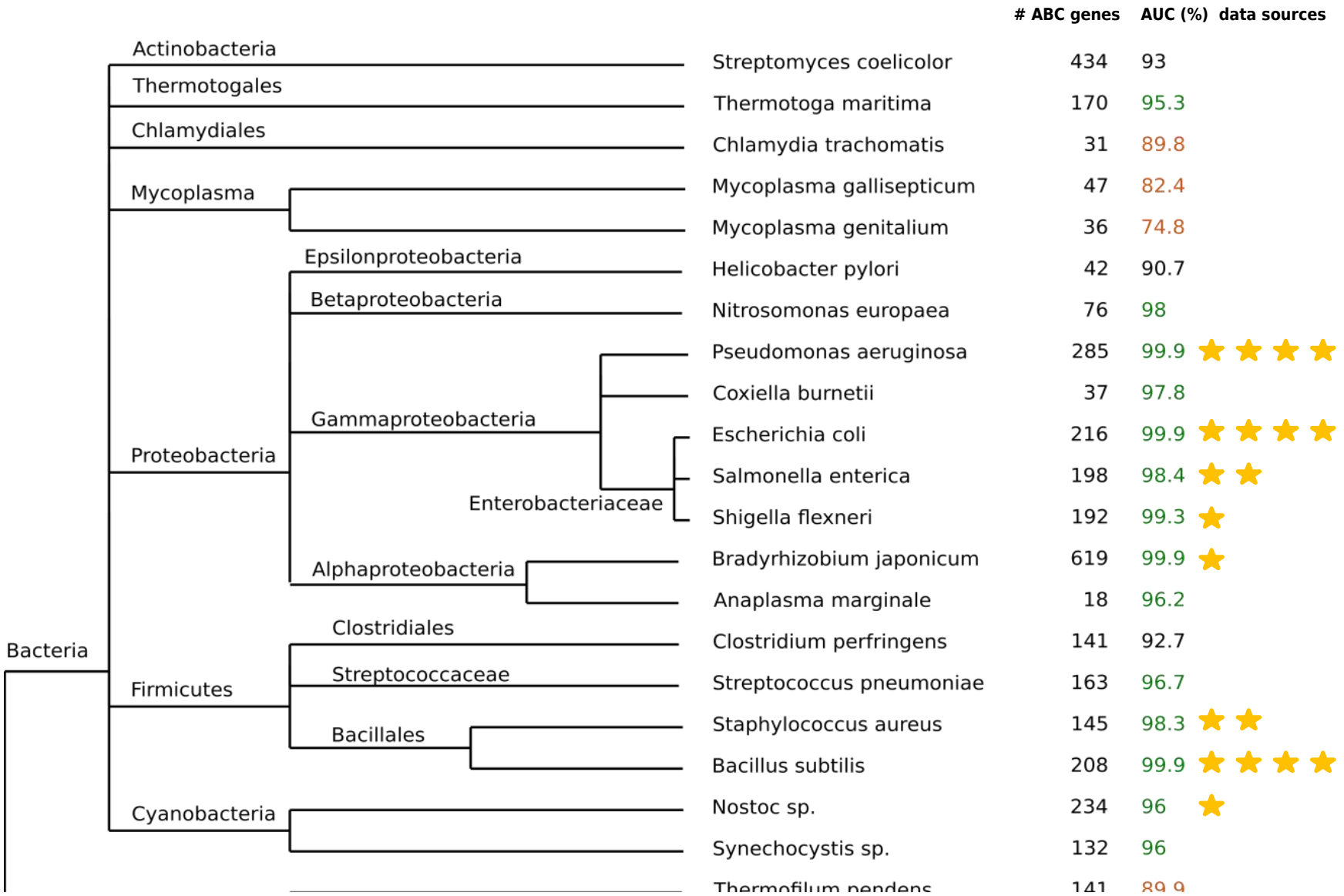
Organisms:
B. subtilis, E. coli, P. aeruginosa
192 ABC systems, 635 genes





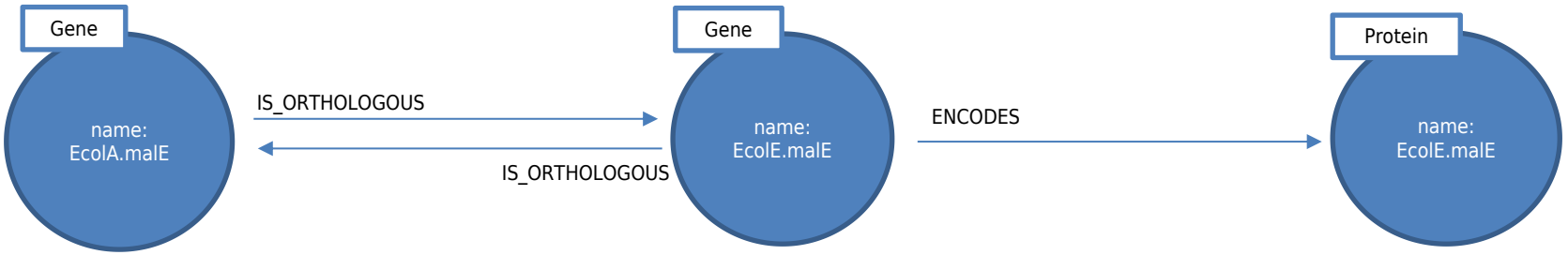


Performances: using other organisms data through orthology



Base de données orientée graphe (*Graph databases*)

- Principes
  - Représenter les données en tant qu'objets reliés par des relations
  - Chaque objet ou relation peut avoir des attributs qui lui sont propres
  - Développement d'un langage de manipulation et de requête

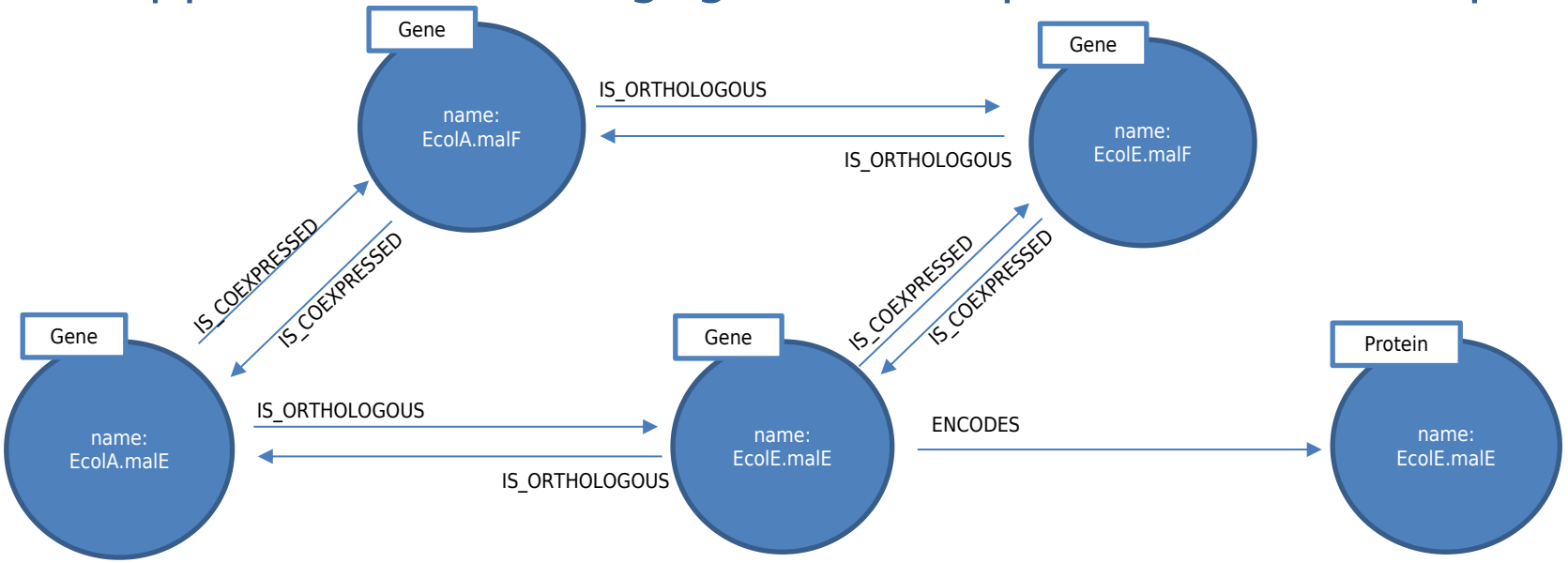


Labeled Property Graph

`(EcolA.malE:Gene)<-[:IS_ORTHOLOGOUS]->(EcolE.malE:`*`Gene`*`)-[:ENCODES]->(EolE.malE:`*`Protein`*`)`

Base de données orientée graphe (*Graph databases*)

- Principes
  - Représenter les données en tant qu'objets reliés par des relations
  - Chaque objet ou relation peut avoir des attributs qui lui sont propres
  - Développement d'un langage de manipulation et de requête

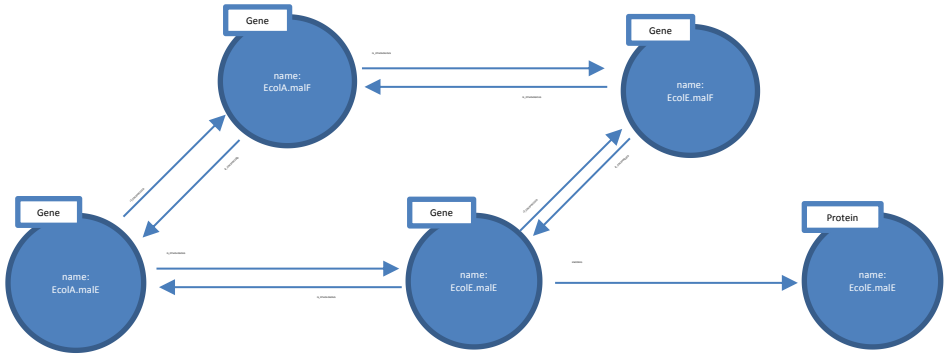


Labeled Property Graph

(**EcolE.malE:Gene**)<-[:IS_COEXPRESSED]->(EcolE.malF)<-[:IS_ORTHOLOGOUS]->
(EcolA.malF:Gene)<-[:IS_COEXPRESSED]->(EcolA.malE:Gene)<-[:IS_ORTHOLOGOUS]->(**EcolE.malE:Gene**)
-[:ENCODES]->(EolE.malE:Protein)

Labeled property graph

Un graphe avec propriétés étiquetées est constitué de sommets, relations, propriétés et étiquettes :



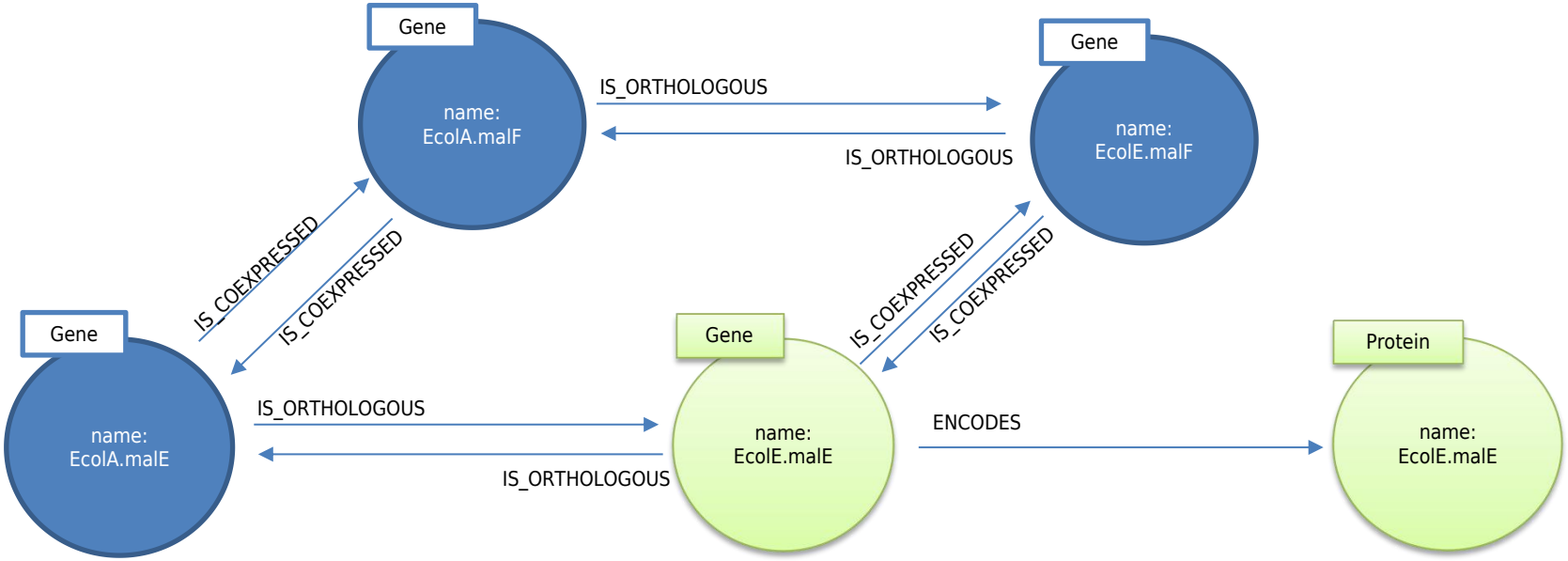
- Propriétés des sommets : de type clé/valeur

- Étiquettes des sommets : une ou plusieurs afin de les regrouper (`Gene, Protein`)

- Relations : orientées, peuvent avoir des propriétés comme les sommets.

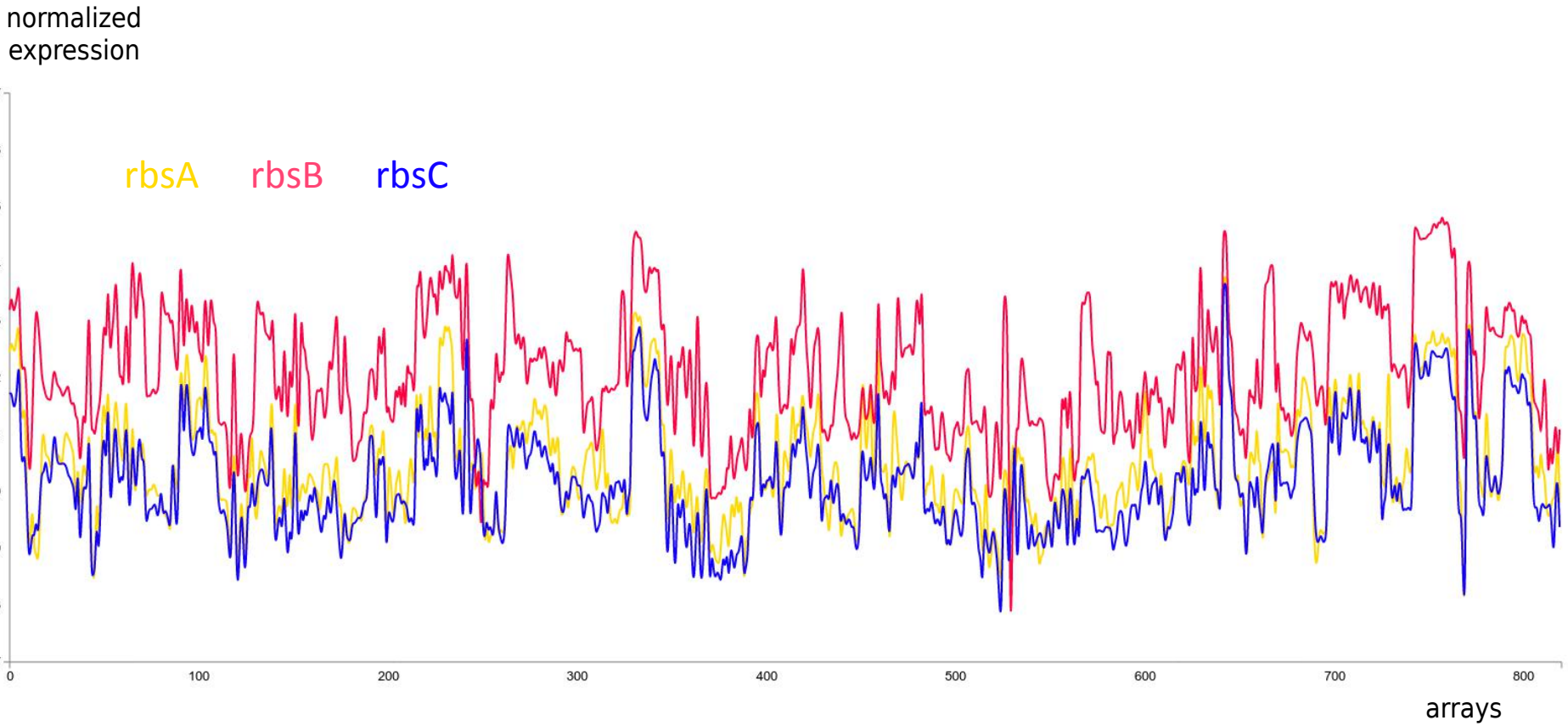
Langage de requête, exemple : Cypher

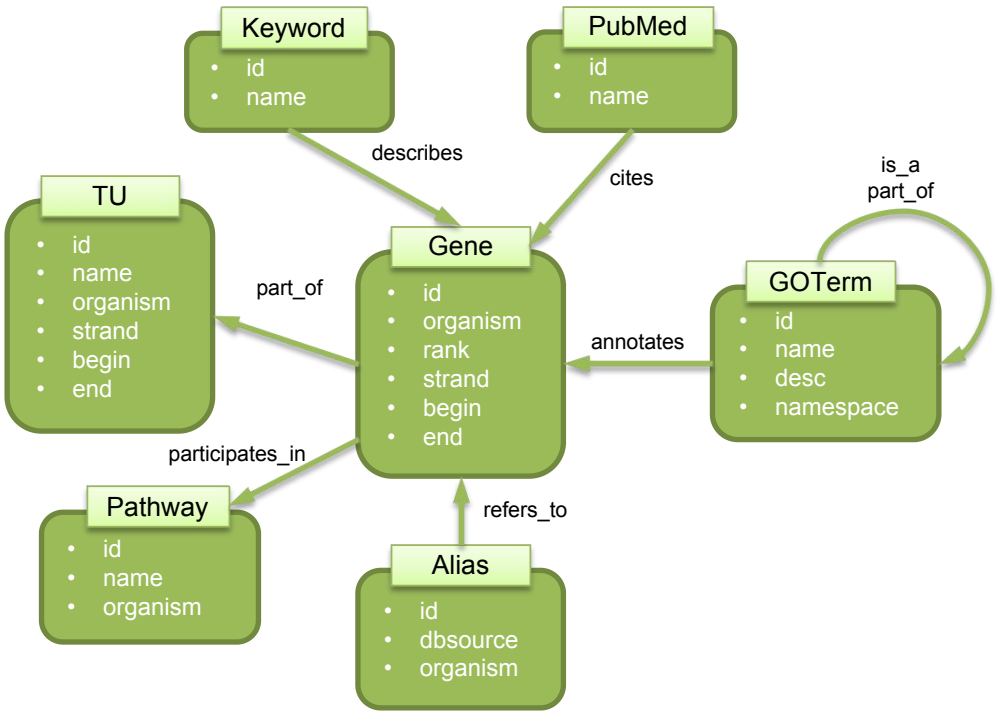
```
MATCH (g:Gene)-[:ENCODES]->(p:Protein)
WHERE g.name='EcolE.malE'
RETURN g,p
```



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




Keyword
id
name
PubMed
id
name
describes
cites
is_a
part_of
TU
id
name
organism
strand
begin
end
part_of
Gene
id
organism
rank
strand
begin
end
annotates
GOTerm
id
name
desc
namespace
participates_in
Pathway
id
name
organism
refers_to
Alias
id
dbsource
organism