Intégration de données hétérogènes

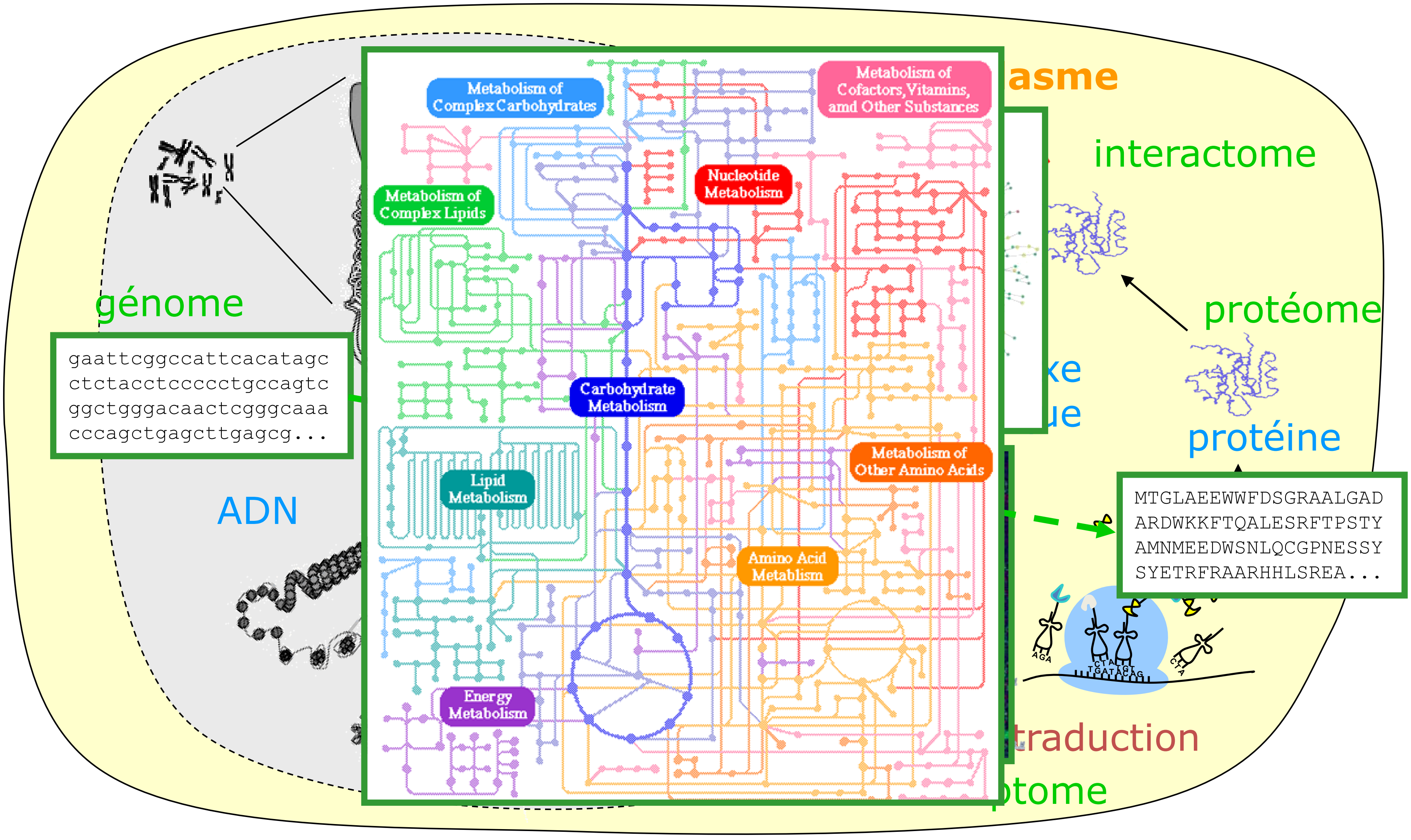
Master 2 MABS

Bioinformatique et Biologie des Systèmes

Introduction

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

Biologie, mesures, données et connaissances



```
gaattcggccattcacatagc
ctctacctccccctgccagtc
ggctgggacaactcgggcaaa
cccagctgagcttgagcg...
```

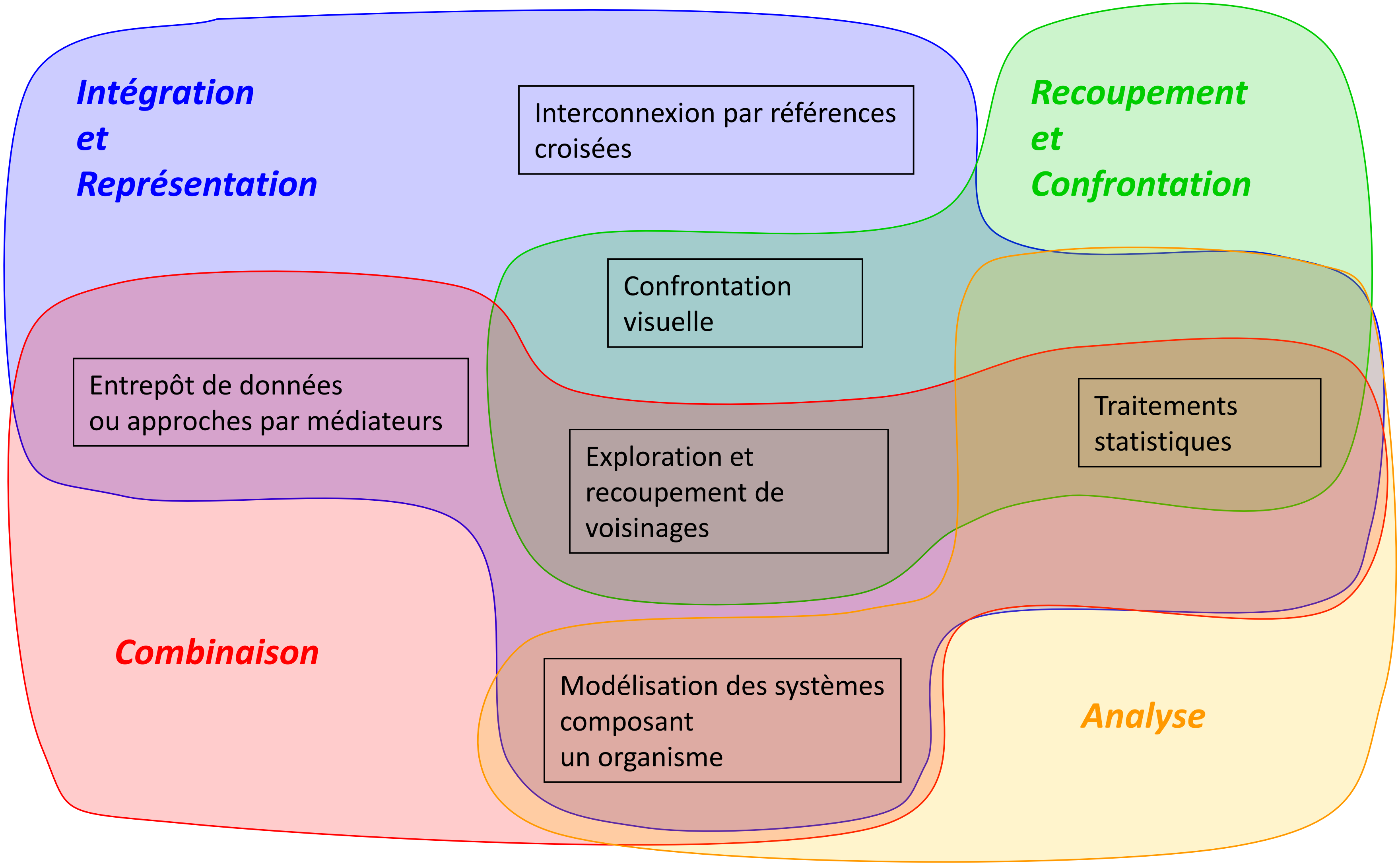
```
MTGLAEEWWFDSGRAALGAD
ARDWKKFTQALESRFTPSTY
AMNMEEDWSNLQCGPNESSY
SYETRFRAARHHLSREA...
```

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

Vue générale



Intégration et Représentation

Recoupement et Confrontation

Combinaison

Analyse

Interconnexion par références croisées

Confrontation visuelle

Entrepôt de données ou approches par médiateurs

Exploration et recoupement de voisinages

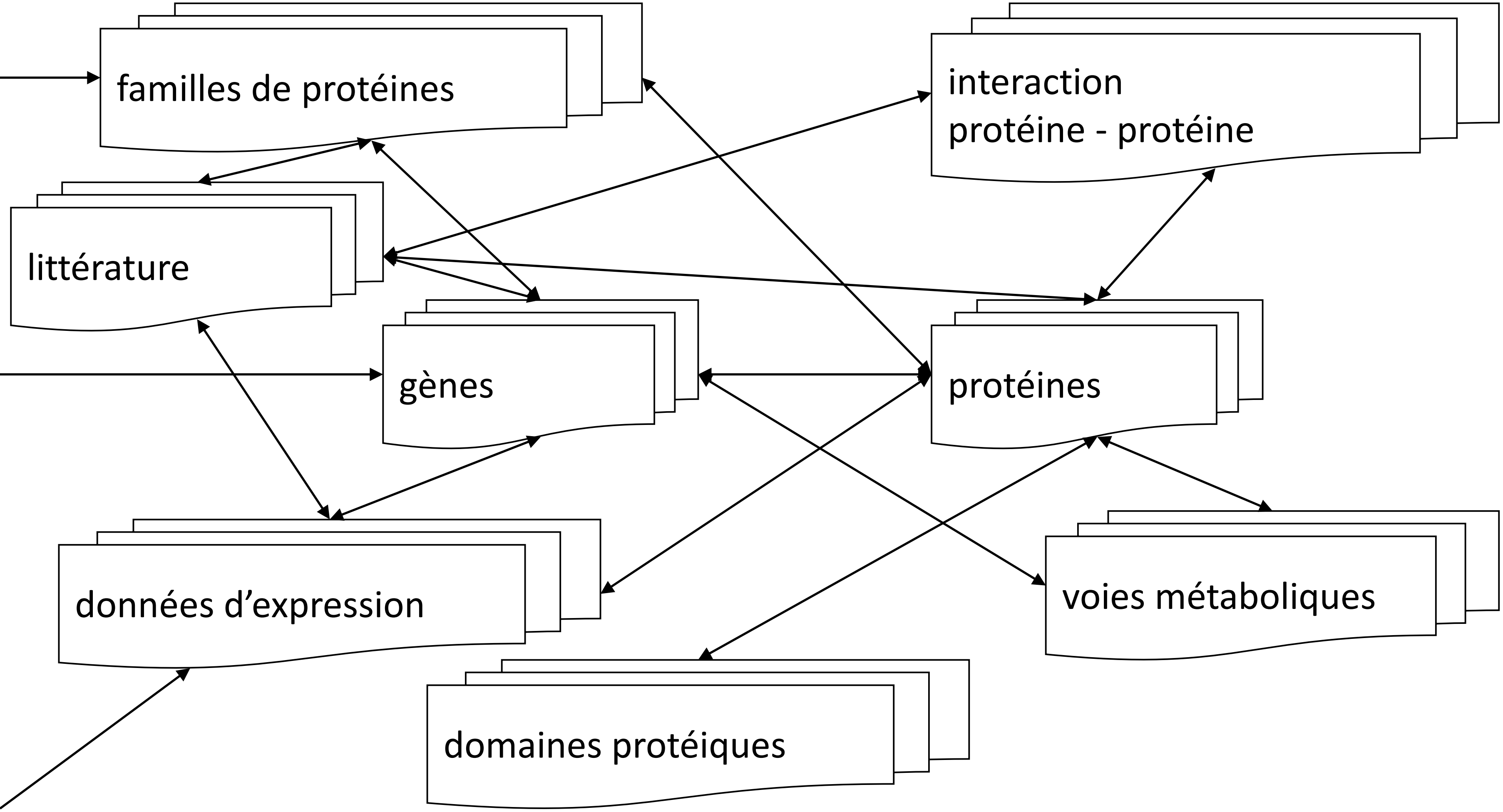
Traitements statistiques

Modélisation des systèmes composant un organisme

Intégration : vers une vue unifiée

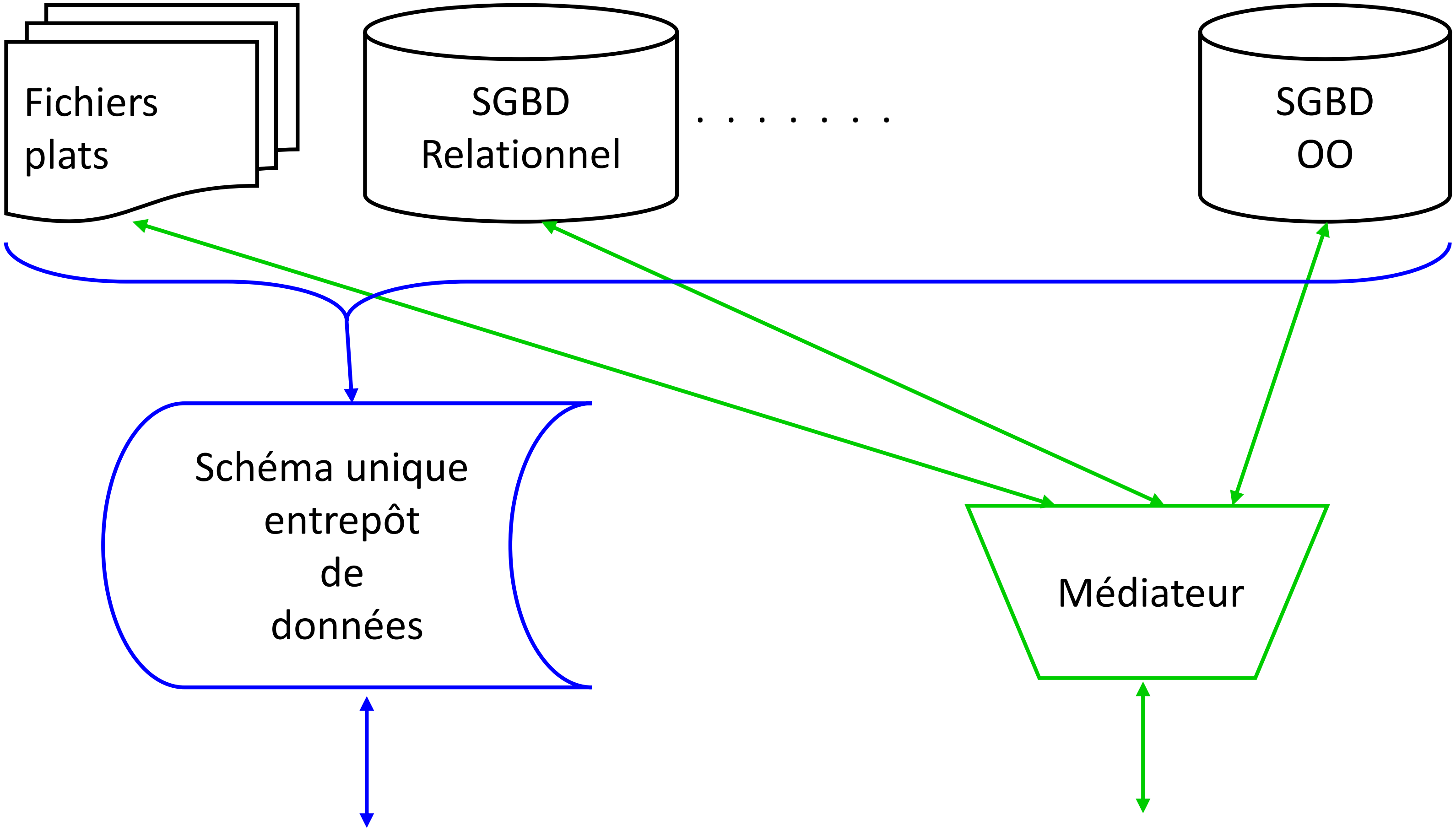
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virtuel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

Intégration par interconnexion : principe



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

Intégration par fusion ou par médiateurs

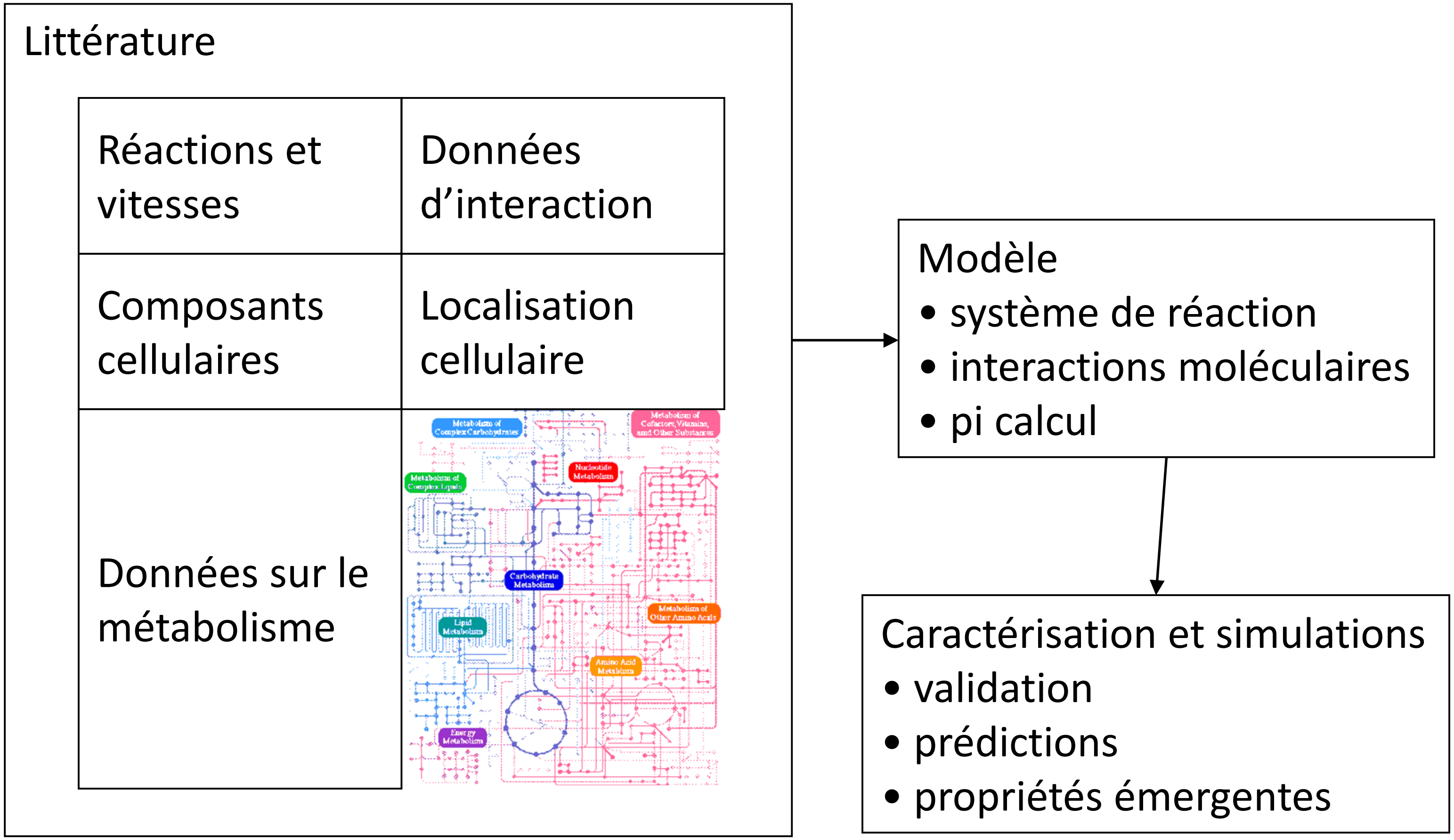


Integr8 [Kersey et al., 2005], BioMart [Kasprzyk et al., 2004],
WInGS [Abergel et al., 2004], BioKleisli [Davidson et al., 1997], …

Intégration : vers une vue unifiée

- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : 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

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

Modèle

- système de réaction
- interactions moléculaires
- pi calcul

Caractérisation et simulations

- validation
- prédictions
- propriétés émergentes

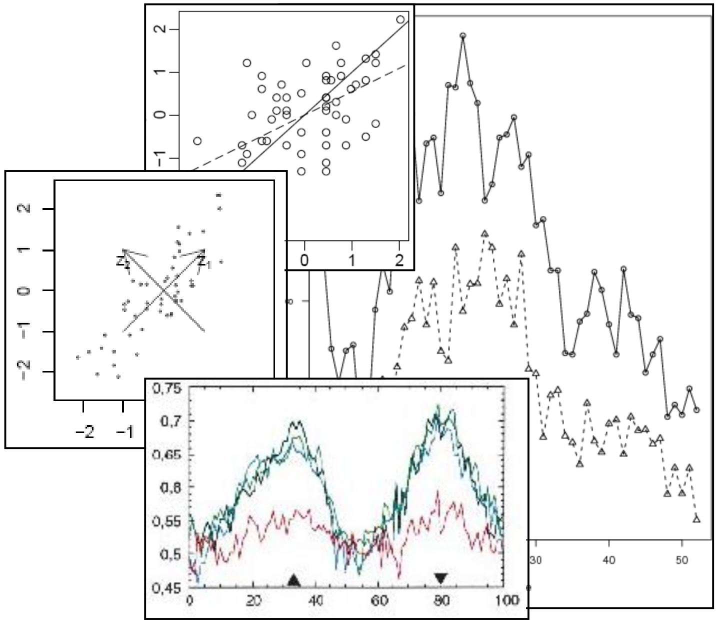
Virtual Cell [Loew et Schaff, 2001], E-CELL [Tomita et al., 1999], Cellerator [Shapiro et al., 2003], …

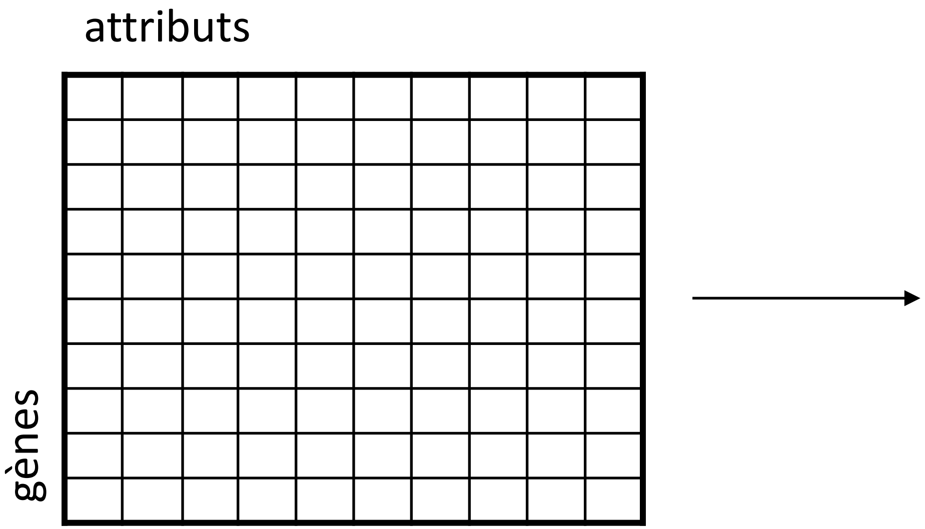
Intégration : vers une vue unifiée

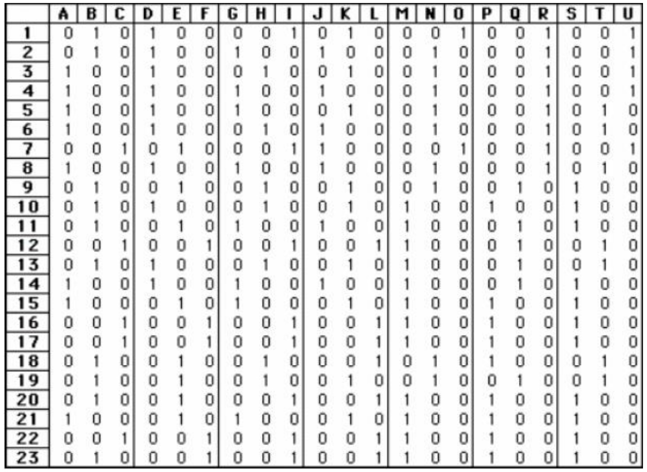
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

Statistiques

- Analyse de la variance
- Analyse en composantes principales
- Analyse factorielle des correspondances
- Analyse des correspondances multiples
- …





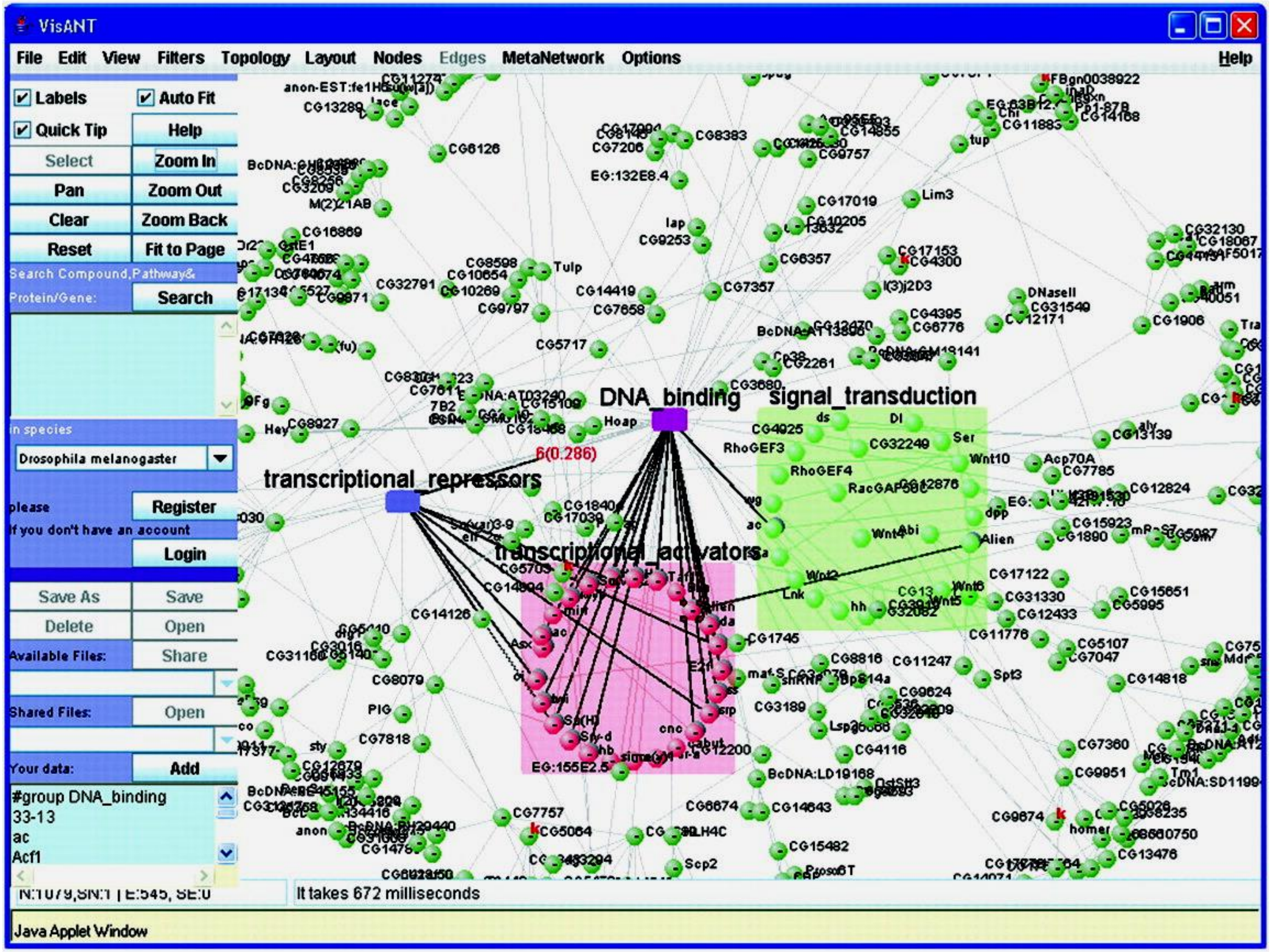


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0	1	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1
2	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1
3	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
4	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1
5	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0
6	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	1	0
7	0	0	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	1	0	0	1
8	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0
9	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	1	0	0
10	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0
11	0	1	0	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0
12	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	0	1	0	0	1	0
13	0	1	0	1	0	0	0	1	0	0	1	0	1	0	0	0	1	0	0	1	0
14	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	0
15	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0
16	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
17	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
18	0	1	0	0	1	0	0	1	0	0	0	1	0	1	0	1	0	0	0	1	0
19	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0
20	0	1	0	0	1	0	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
21	1	0	0	0	1	0	1	0	0	0	1	0	1	0	0	1	0	0	1	0	0
22	0	0	1	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0
23	0	1	0	0	0	1	0	0	1	0	0	1	1	0	0	1	0	0	1	0	0

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

Confrontation visuelle



Drosophila melanogaster

#group DNA_binding
33-13
ac
Acf1

N:1079,SN:1 | E:545, SE:0

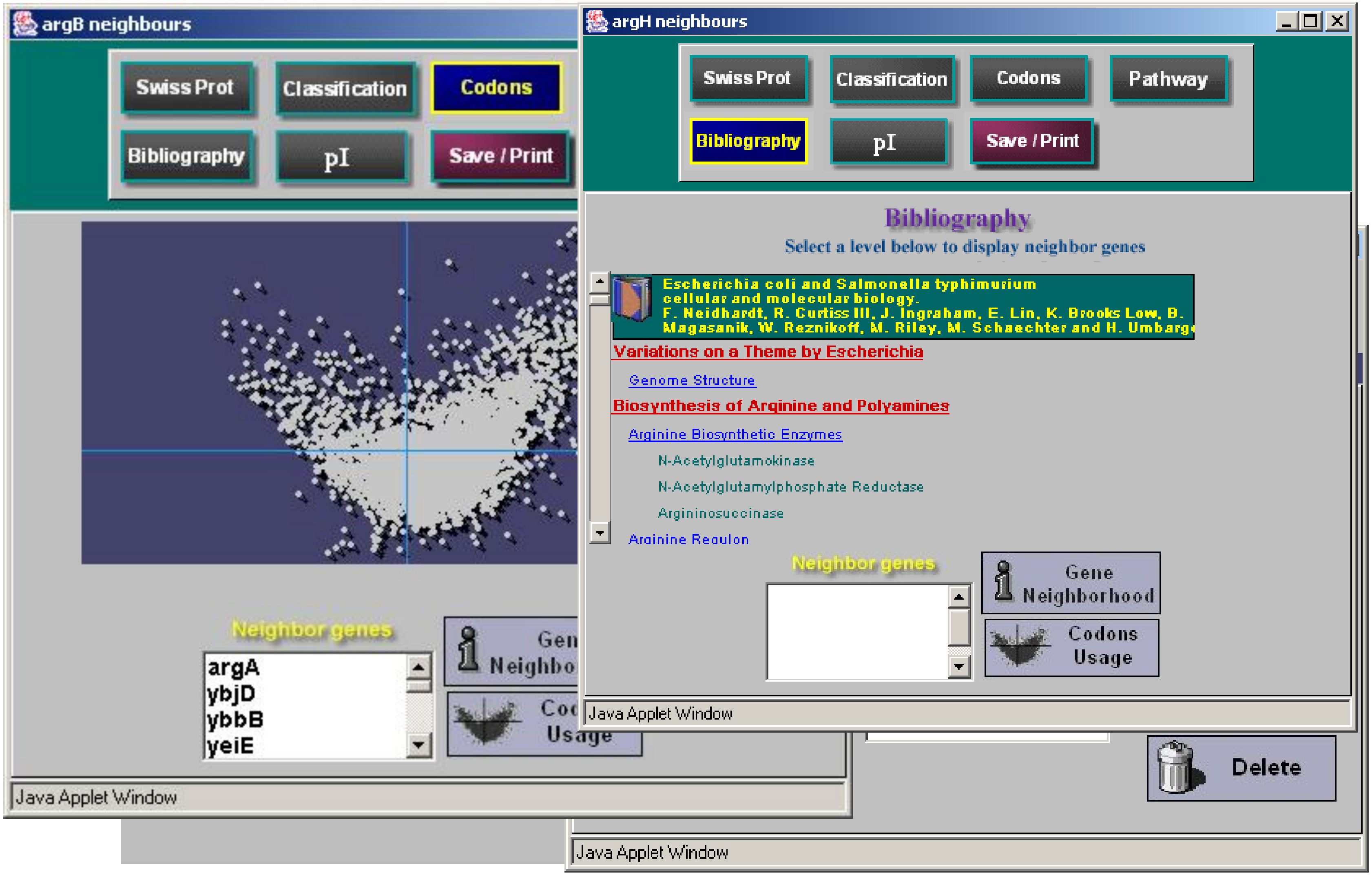
It takes 672 milliseconds

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



argA
ybjD
ybbB
yeiE

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. Umbarg

Variations on a Theme by Escherichia

Genome Structure

Biosynthesis of Arginine and Polyamines

Arginine Biosynthetic Enzymes

N-Acetylglutamokinase

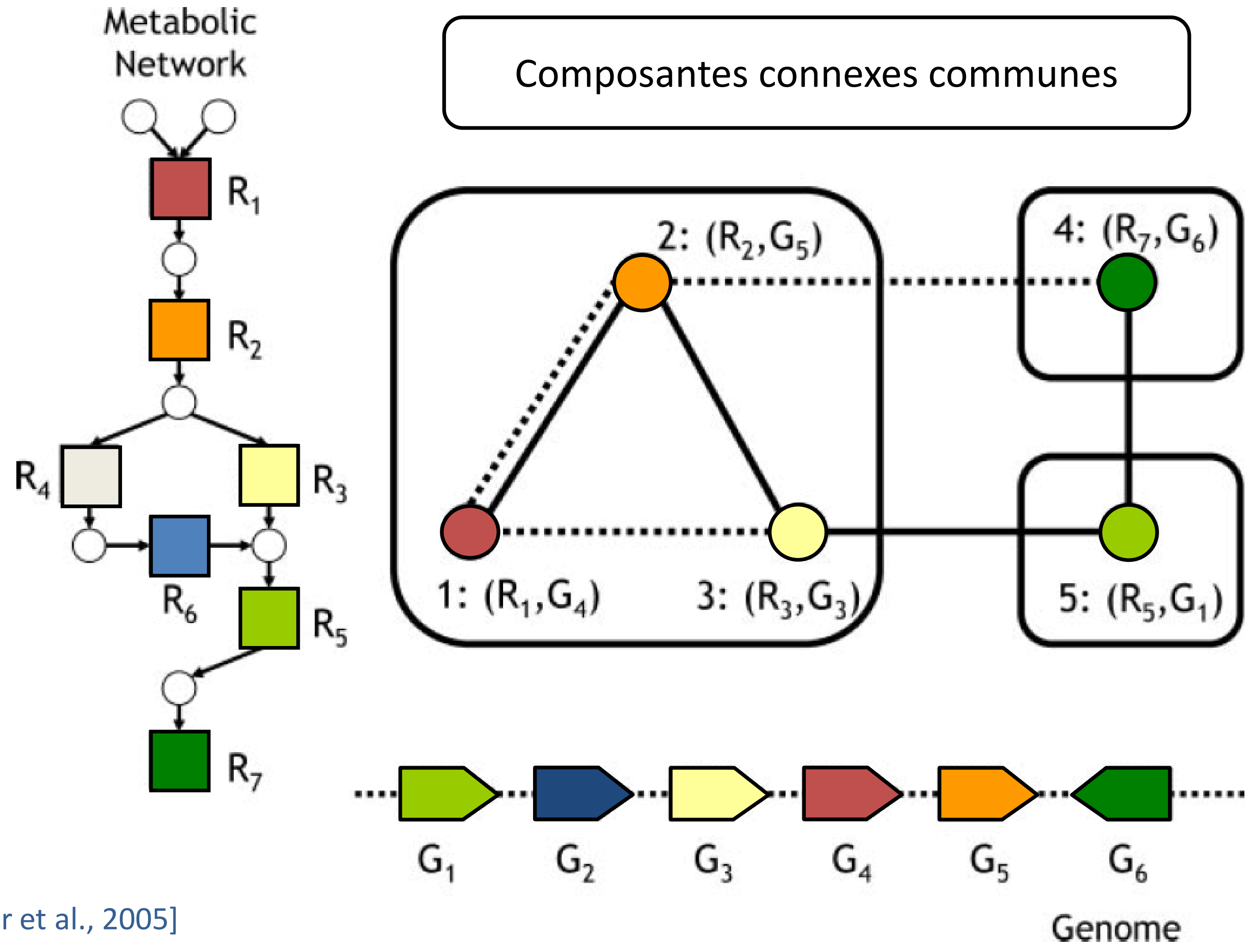
N-Acetylglutamylphosphate Reductase

Argininosuccinase

Arginine Regulon

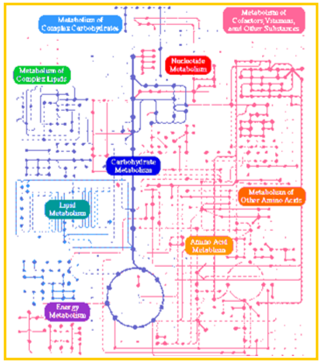
Recoupement de voisinages : approche graphique

Composantes connexes communes

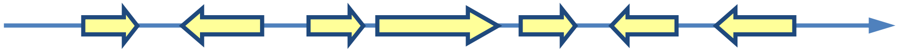


[Boyer et al., 2005]

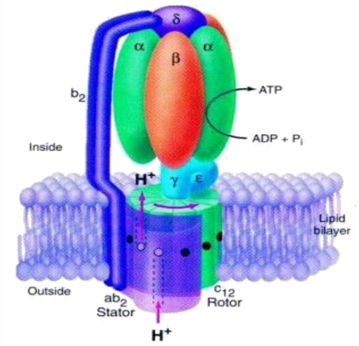
Recoupement de voisinages : approche ensembliste


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

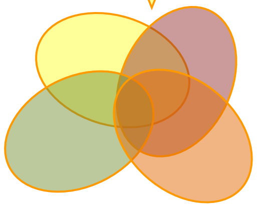
voies métaboliques



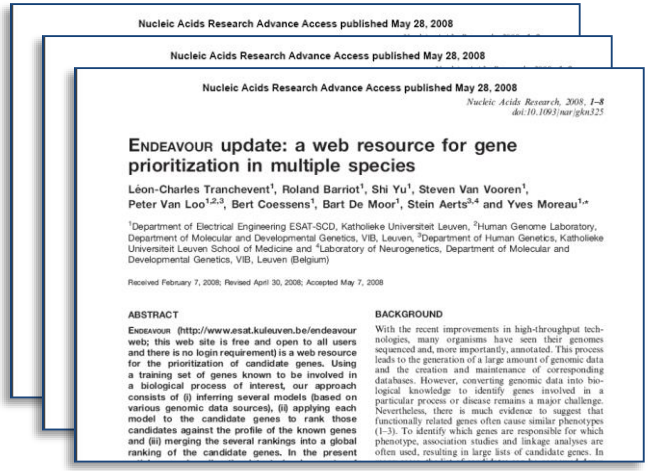
localisation chromosomique



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

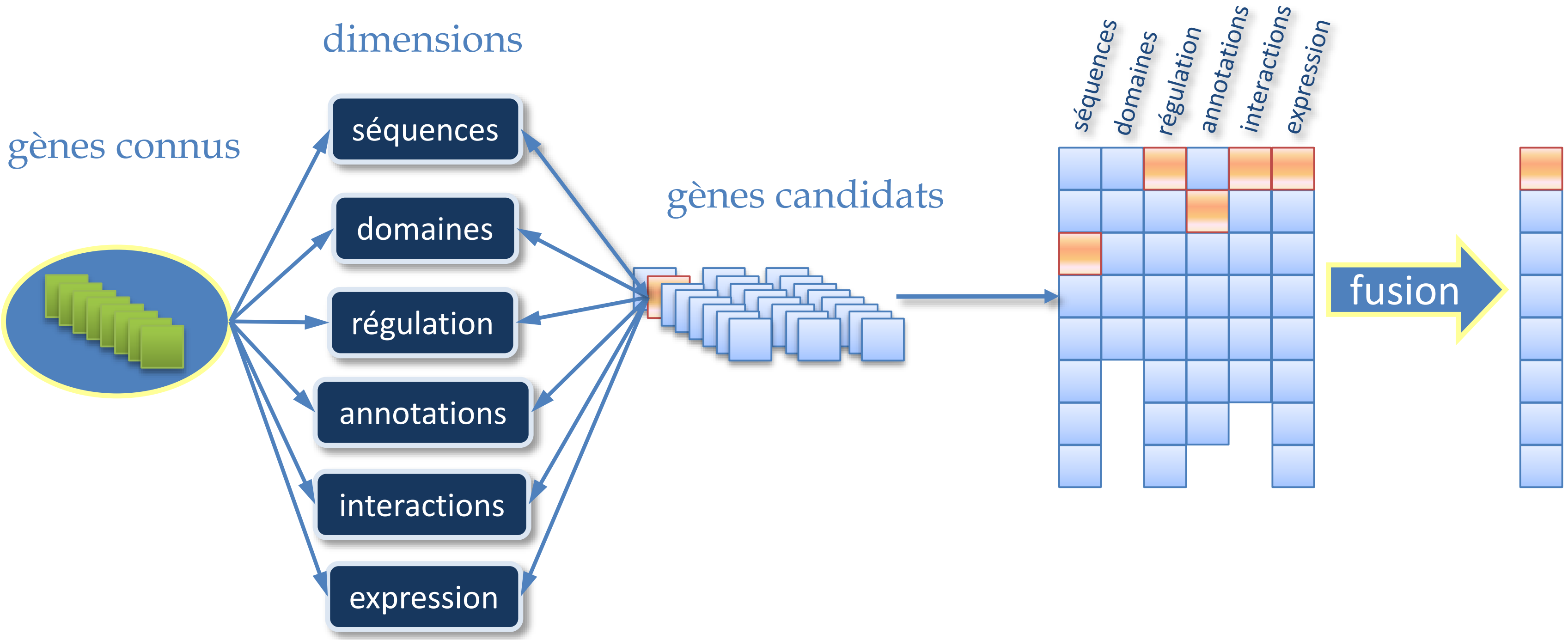
With the recent improvements in high-throughput technologies, many organisms have seen their genomes sequenced and, more importantly, annotated. This process leads to the generation of a large amount of genomic data and the creation and maintenance of corresponding databases. However, converting genomic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

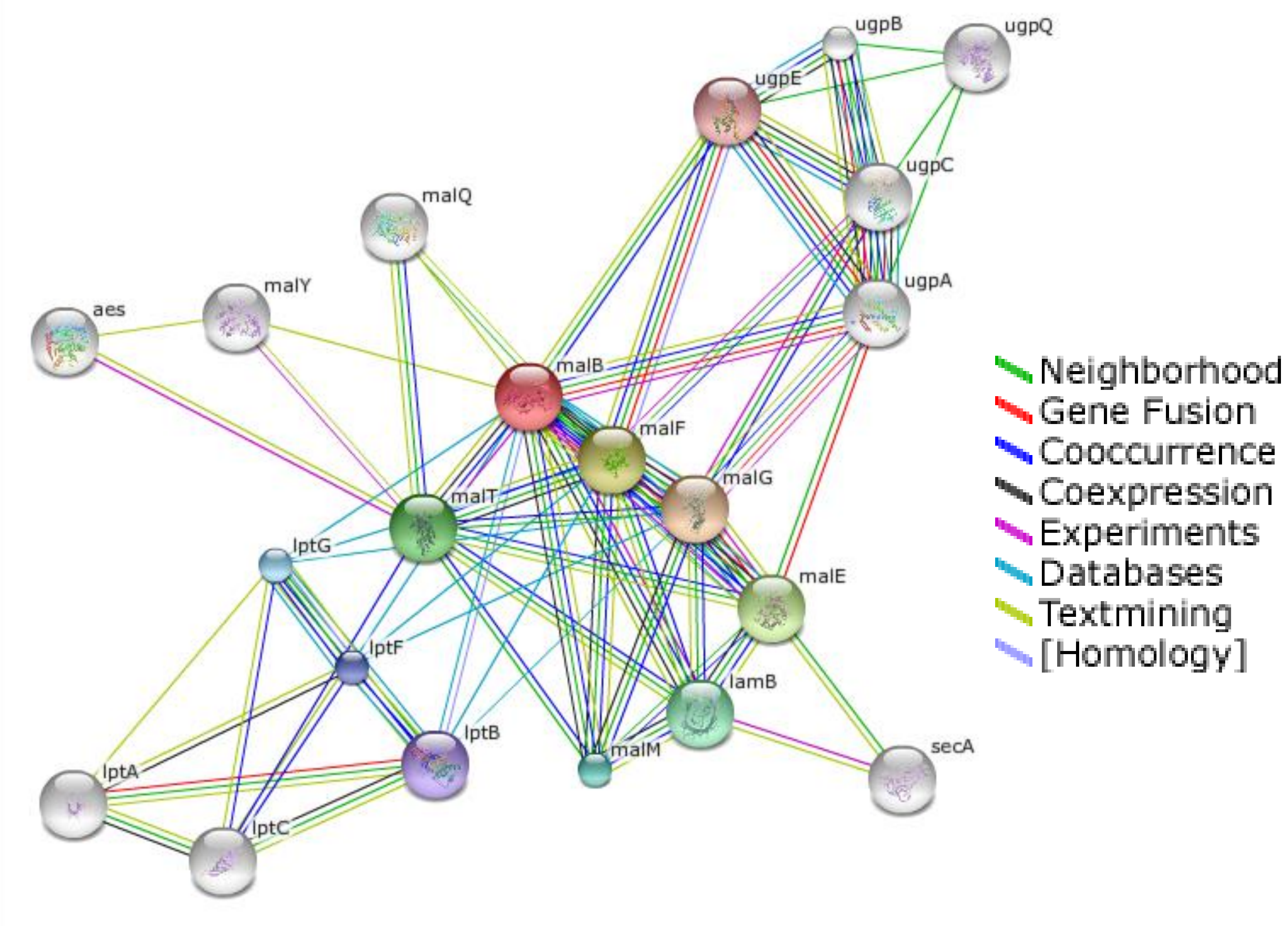
Gene Ontology

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

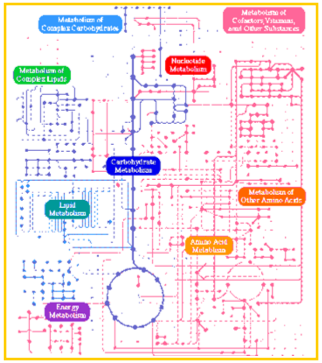


[Aerts *et al.*, 2006]

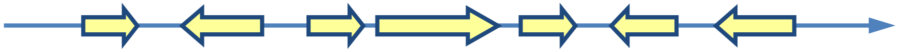
Fusion : approche graphique



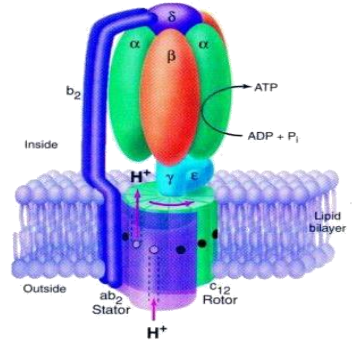
Recoupement de voisinages : approche ensembliste


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

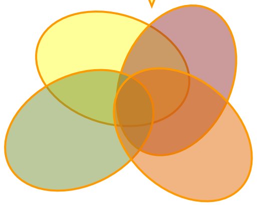
voies métaboliques



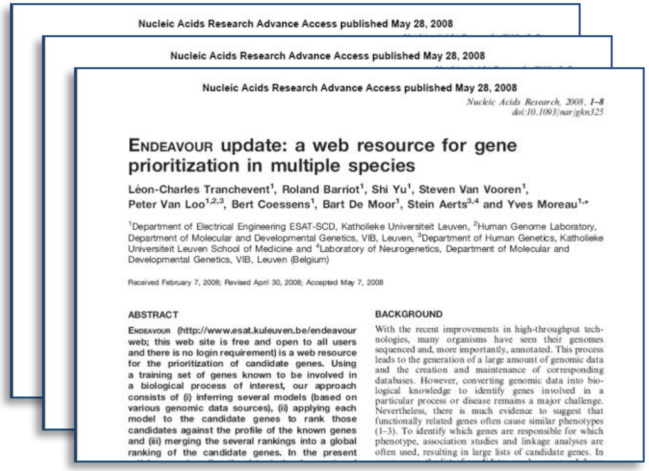
localisation chromosomique



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

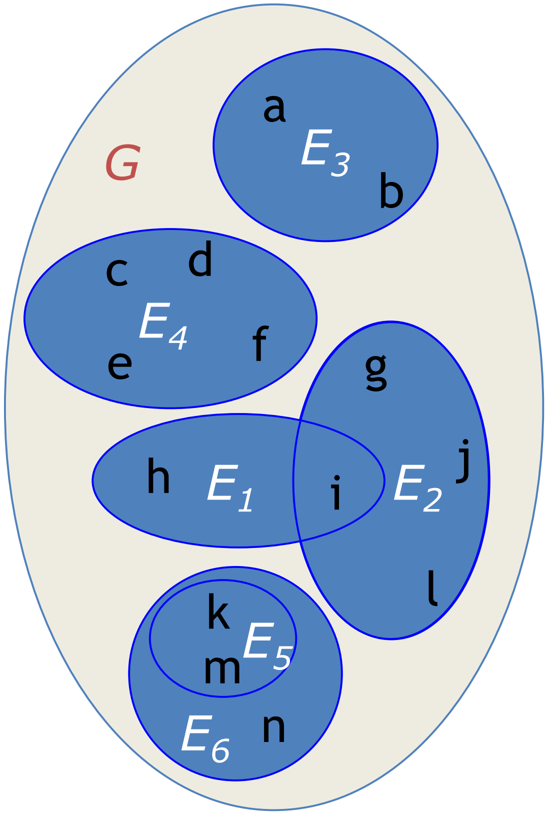
co-citation

domaines protéiques

Gene Ontology

Définitions

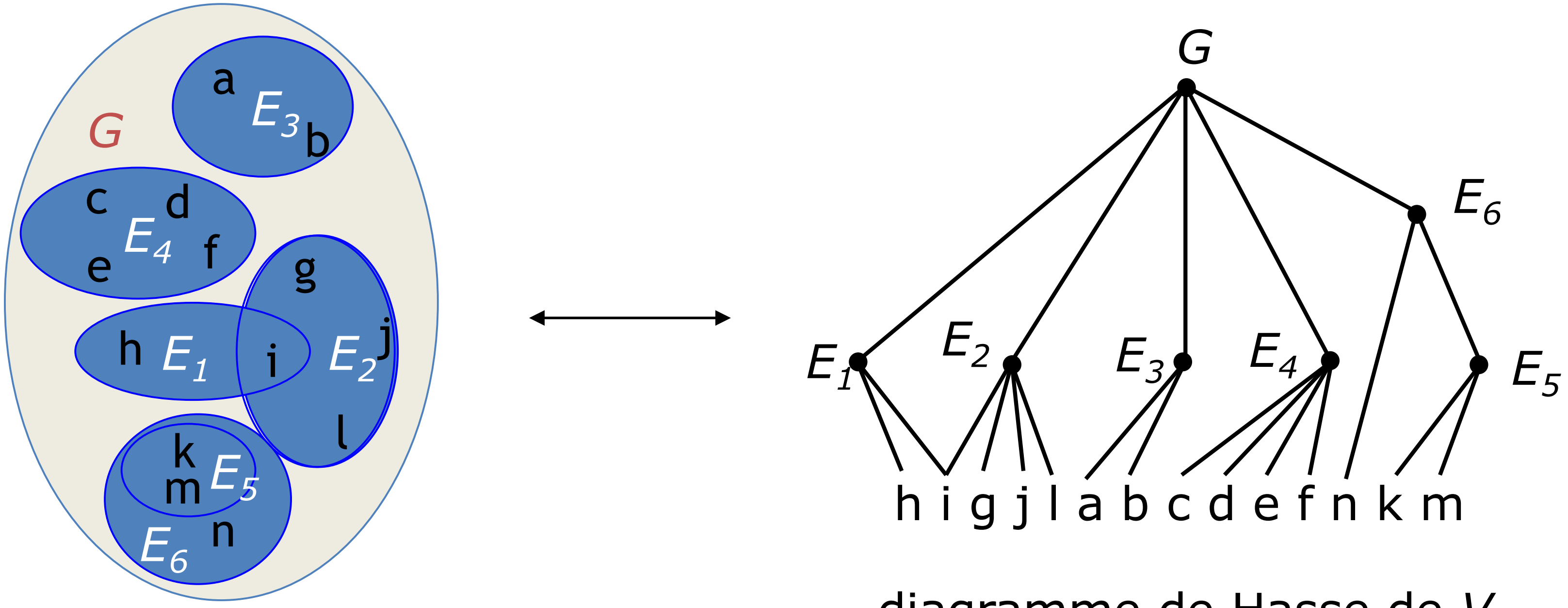
- (Identifiants de) gène → ARNm → protéine
- *G* : ensemble des gènes d'un organisme
- *Fonction de regroupement* : relation entre gènes basée sur un indice de similarité.
- *Ensemble de (gènes) voisins* : ensemble de gènes $E \subseteq G$ regroupés par une fonction de regroupement.
- *Voisinage* : sous-ensemble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

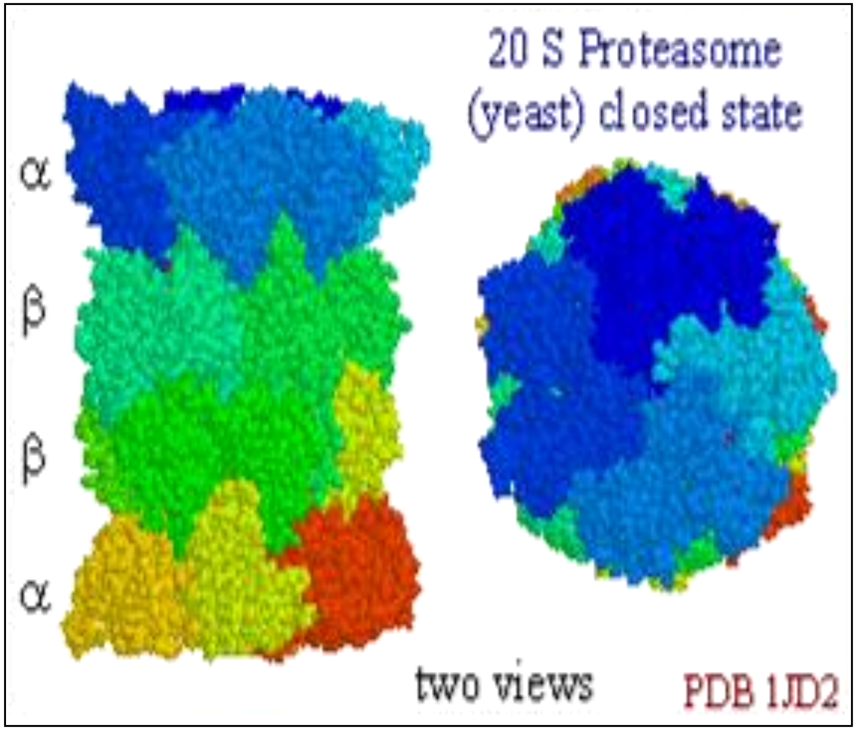
- Un voisinage est un ensemble (d'ensembles de voisins) ordonné par la relation d'inclusion ⊆

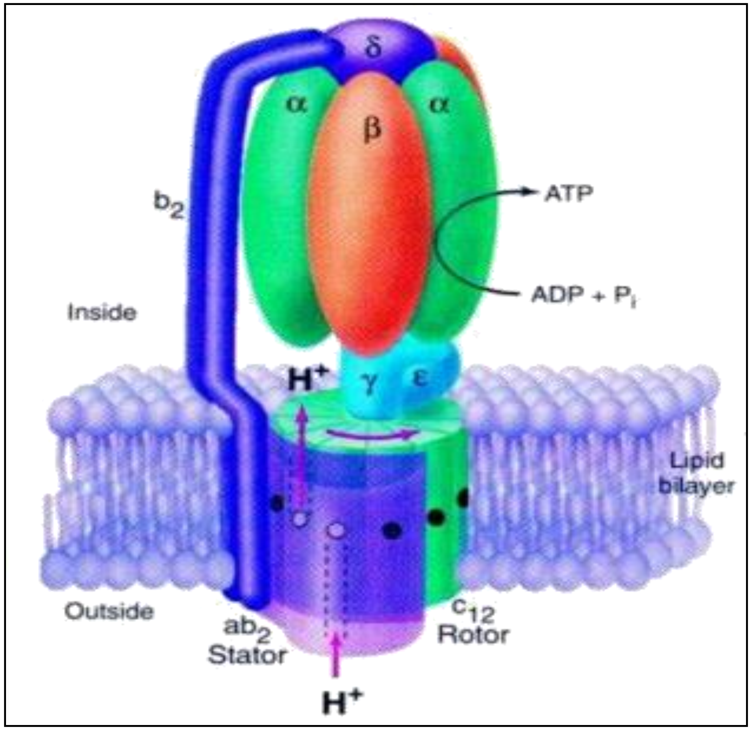


$V = \{E_1,E_2,E_3,E_4,E_5,E_6\}$

diagramme de Hasse de V

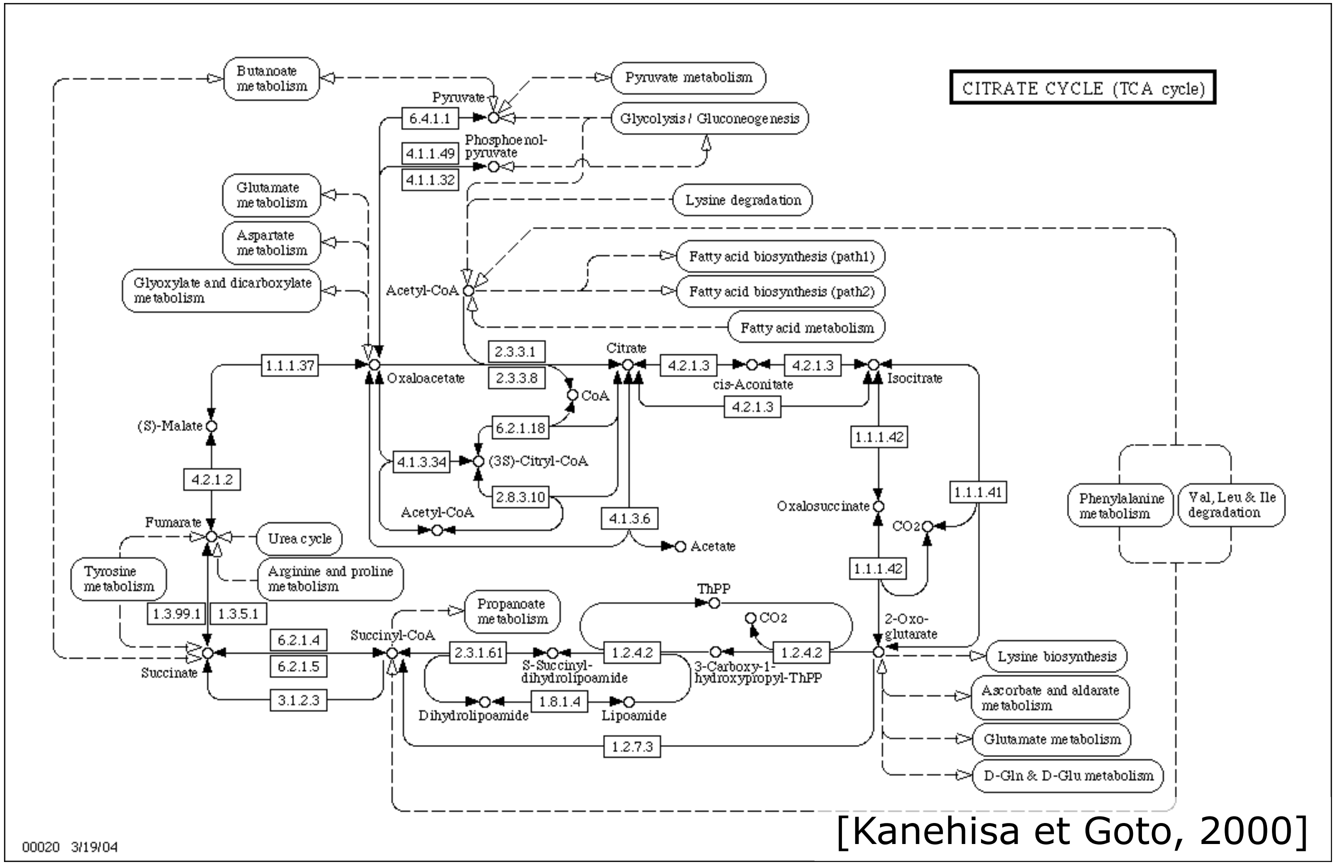
Exemple de fonction de regroupement : complexes protéiques





un complexe → un ensemble de protéines

Exemple de critère de regroupement : voies métaboliques

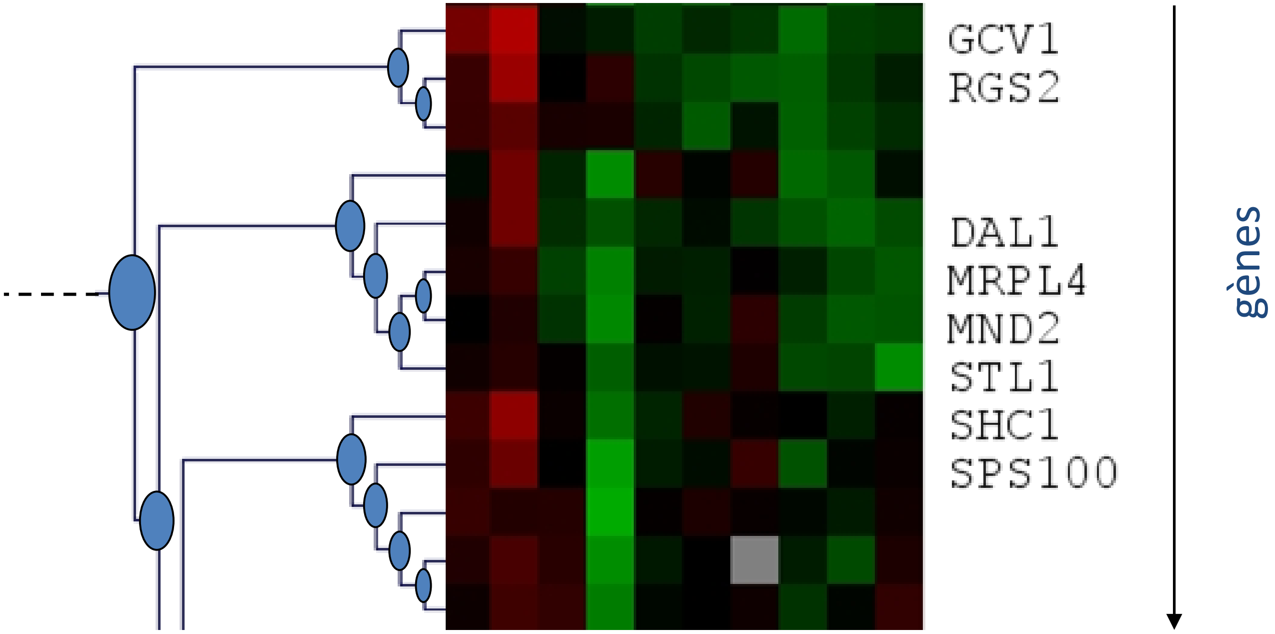


[Kanehisa et Goto, 2000]

une voie métabolique → un ensemble de protéines

Exemple de critère de regroupement : données d'expression

clustering hiérarchique
des profils



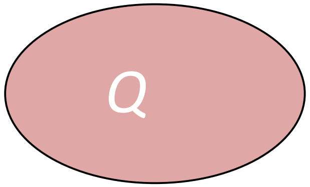
Conditions expérimentales

un cluster → un ensemble de gènes

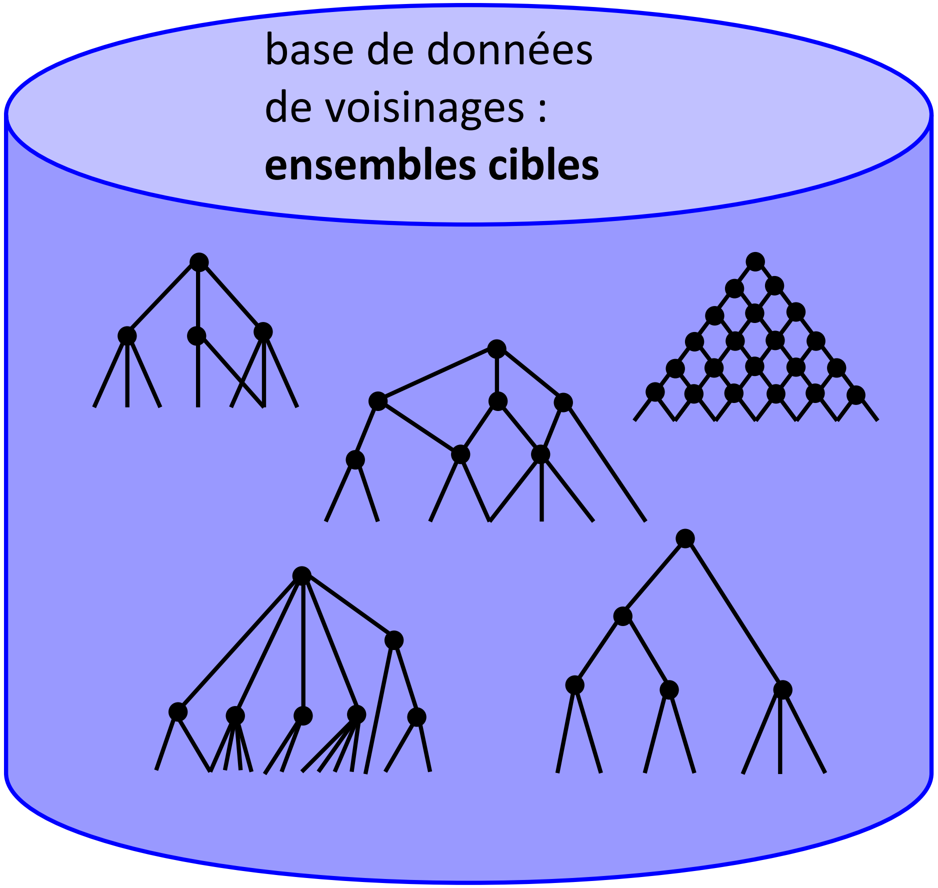
Mise en relation des données

- Recherche d’ensembles similaires

ensemble requête
$Q \subseteq G$

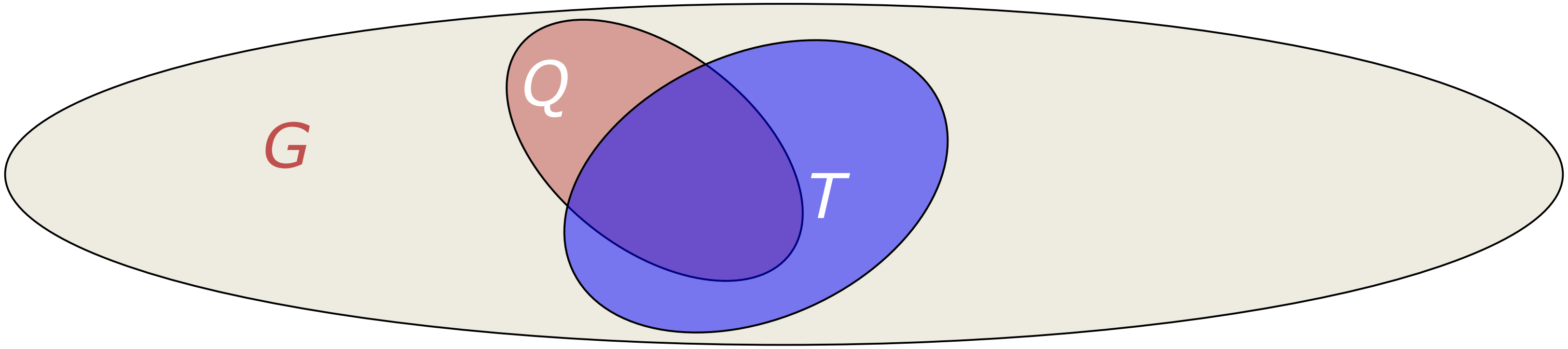


Quels sont les ensembles cibles qui lui sont similaires ?



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

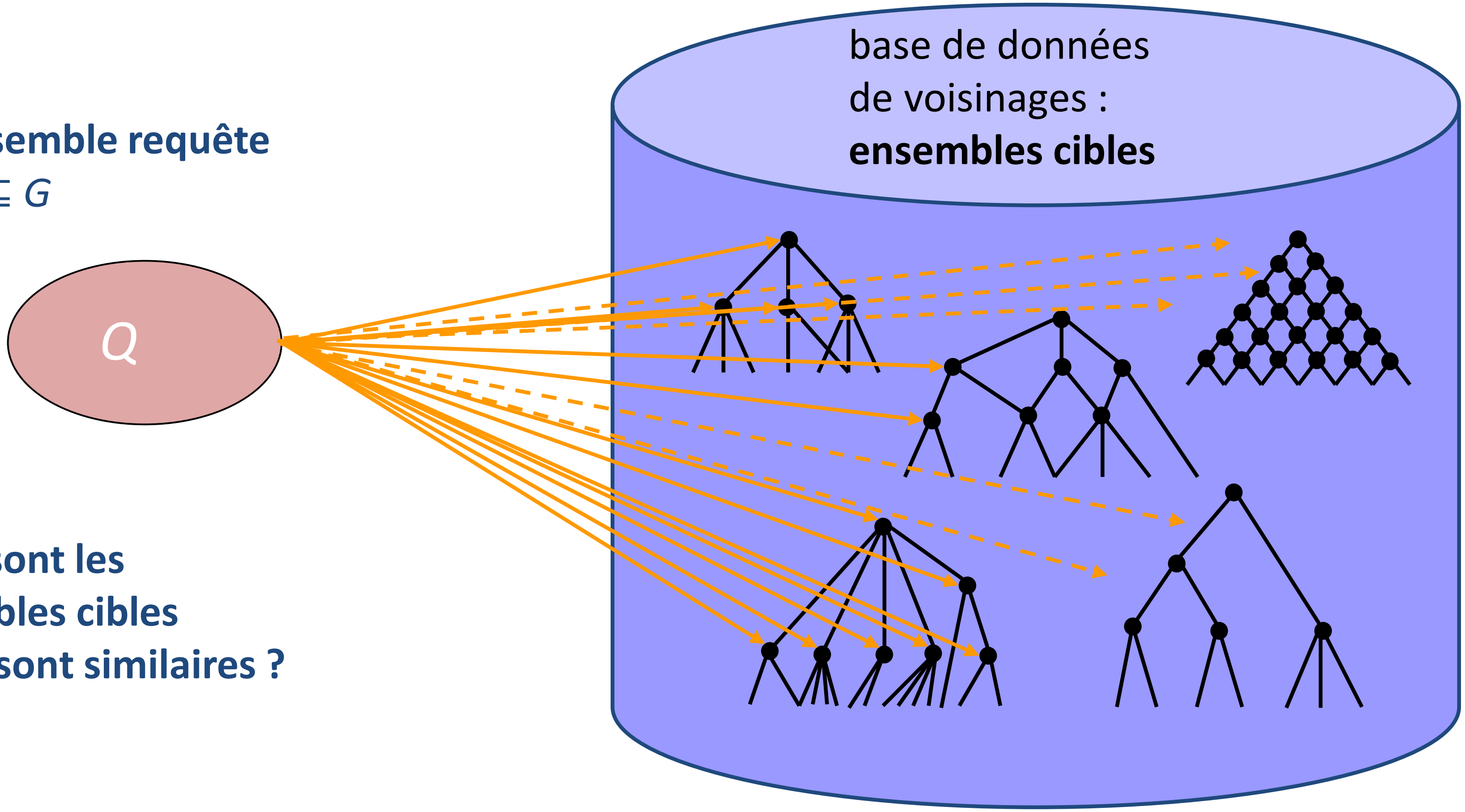
$$p-valeur(c,t,q,g)=\sum_{k=c}^{\min(q,t)}\frac{\binom{t}{k}\binom{g-t}{q-k}}{\binom{g}{q}}$$

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
$Q \subseteq G$

base de données
de voisinages :
ensembles cibles

Quels sont les
ensembles cibles
qui lui sont similaires ?

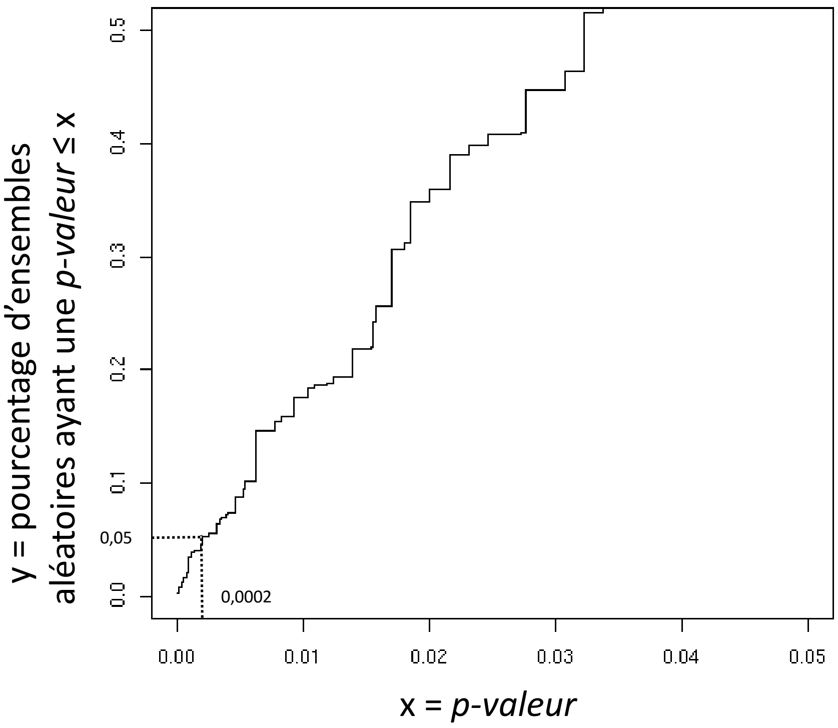
Tests multiples et significativité des p-valeurs

- 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

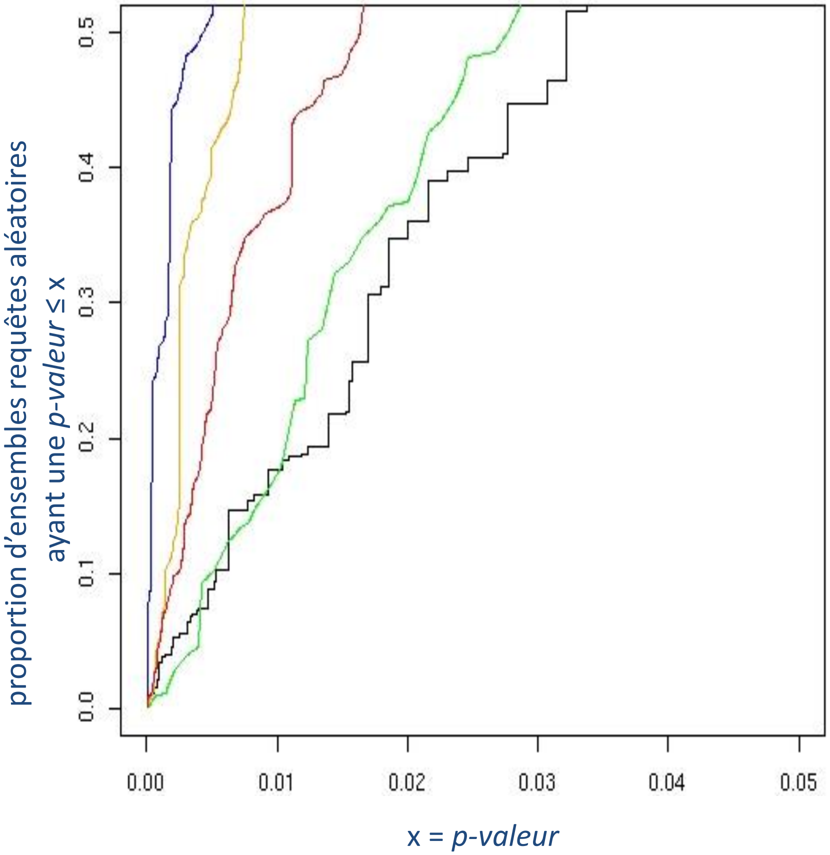
Étant donnée une p-valeur p
Combien ont un meilleur score ?



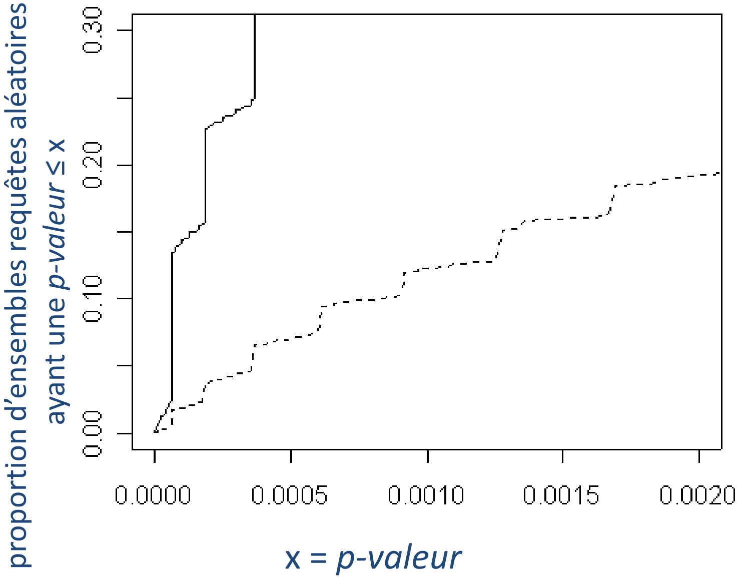
levure *Saccharomyces cerevisiae*
n=500, q=9, g=5786, KEGG Pathways

[Barriot *et al.* 2004]

Significativité des p-valeurs obtenues

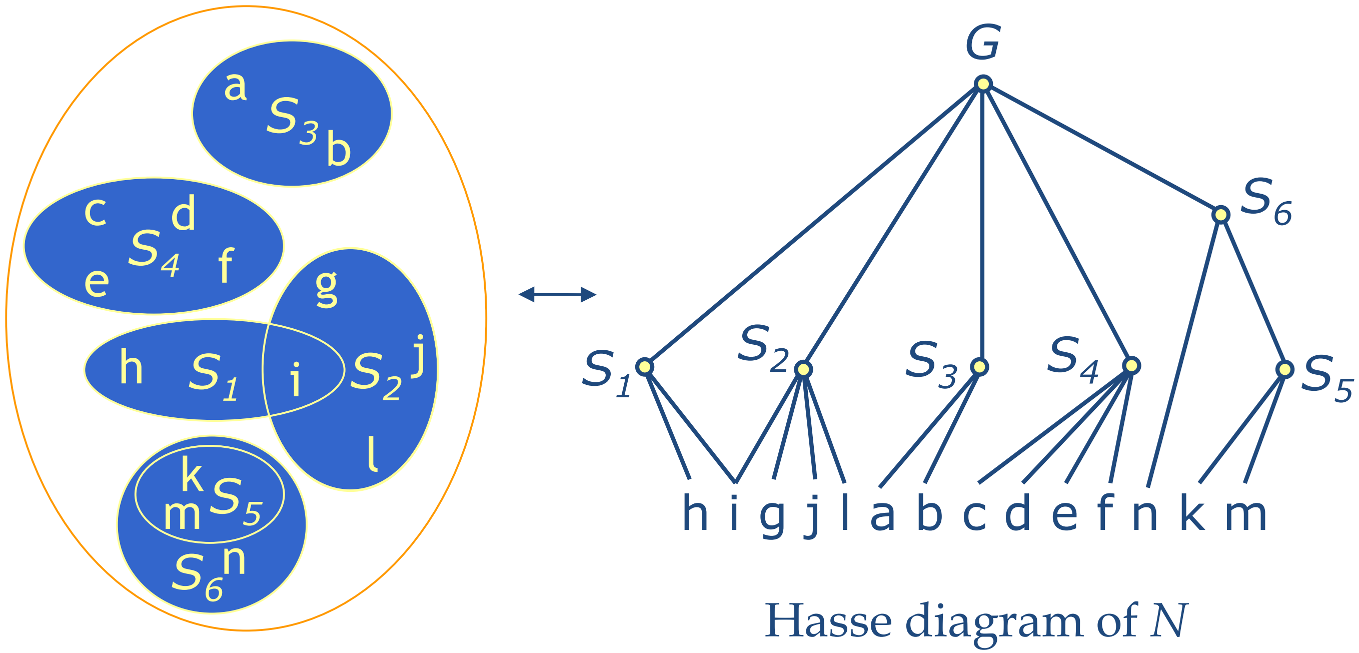


Saccharomyces cerevisiae
n=500, q=6-9-200-500-1000,
g=5786, KEGG Pathways



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

Optimisations



Hasse diagram of N

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

a target set T is **pertinent** if

$Q \cap T \neq \emptyset$

and

$\nexists\ T' \in N$ 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

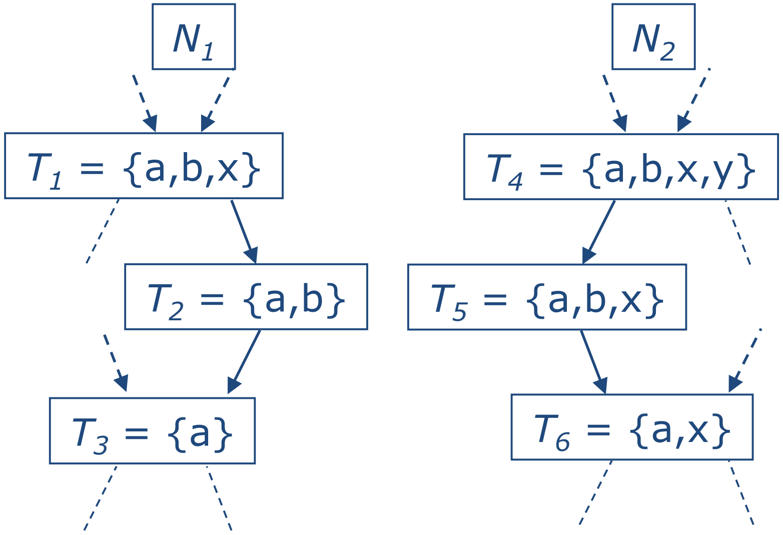
$$Q \cap T \neq \varnothing$$

and

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

and

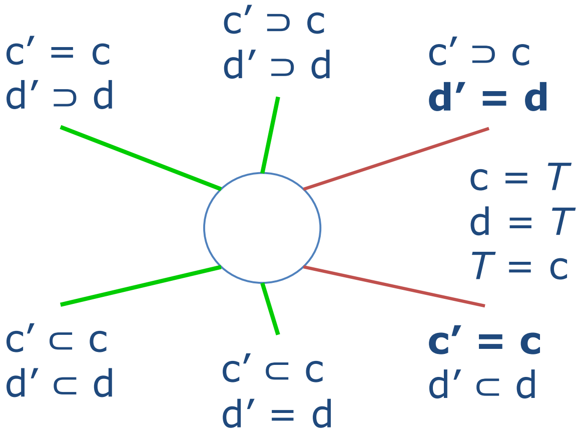
$$\nexists\, T' \in N \text{ such that } T \subset T' \text{ and } T' - Q = T - Q$$



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

Local decision

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



c = $T \cap Q$
d = $T - Q$
T = c ∪ d

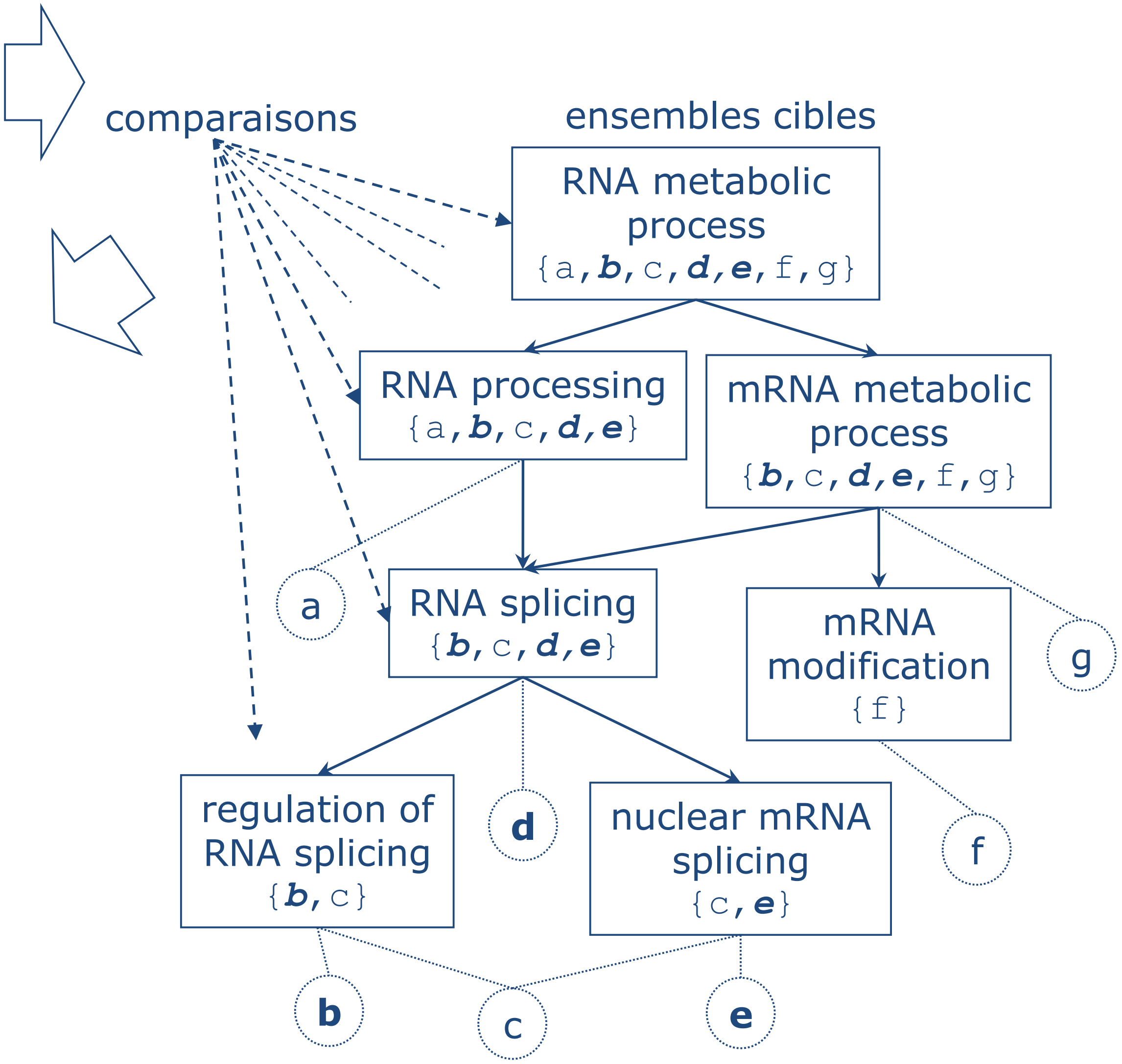
[Barriot, Dutour, Sherman, 2007, *BMC Bioinformatics*]

Illustration

- Pertinence des comparaisons & redondance des résultats

ensemble requête :
gènes d’intérêt

{*b*,*d*,*e*}

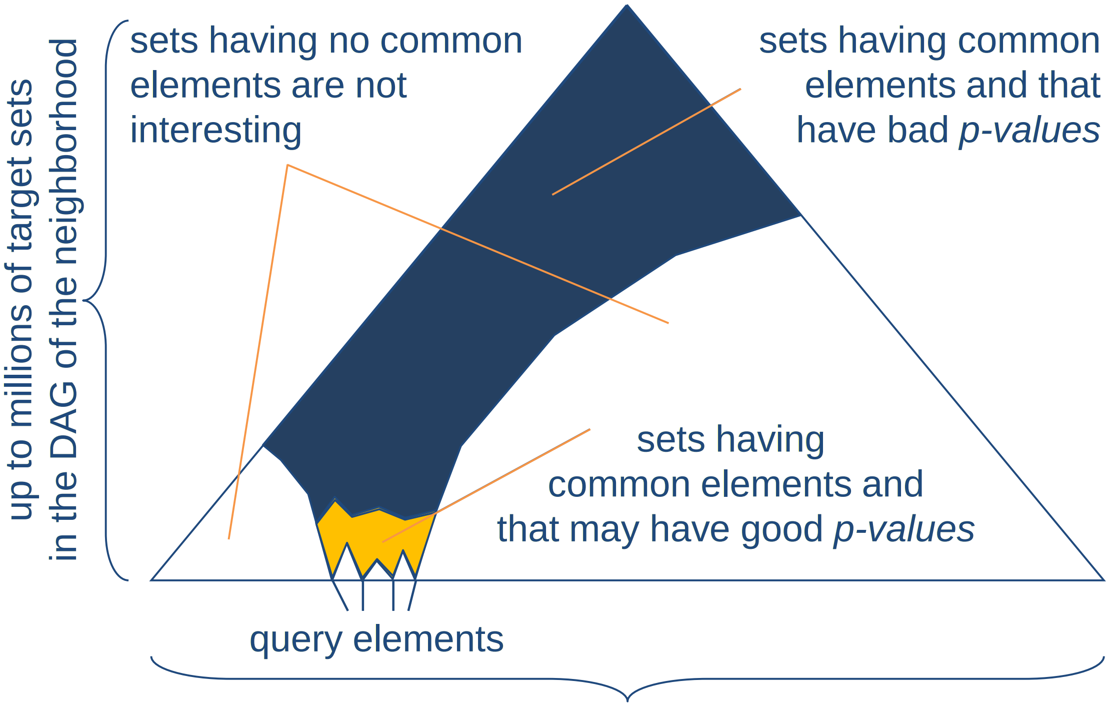


résultats

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*

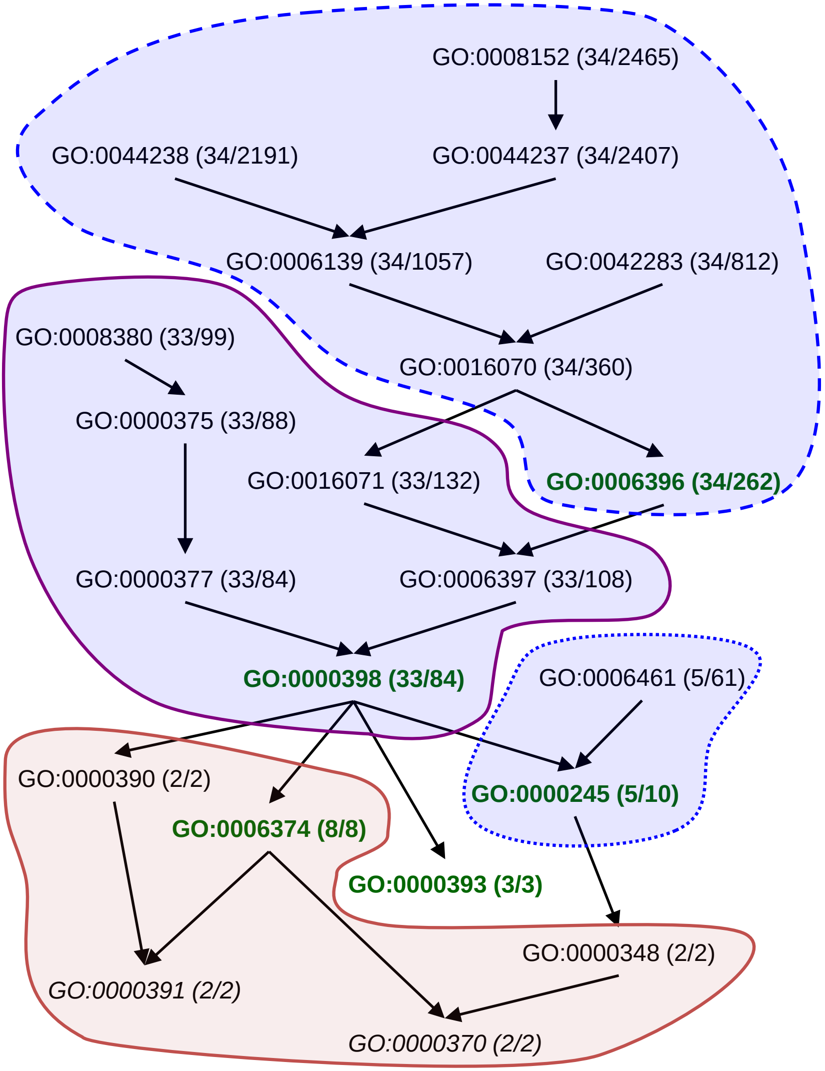
sets having common elements and that may have good *p-values*

query elements

thousands of elements (genes or proteins)

Complex 440.30.10 mRNA splicing

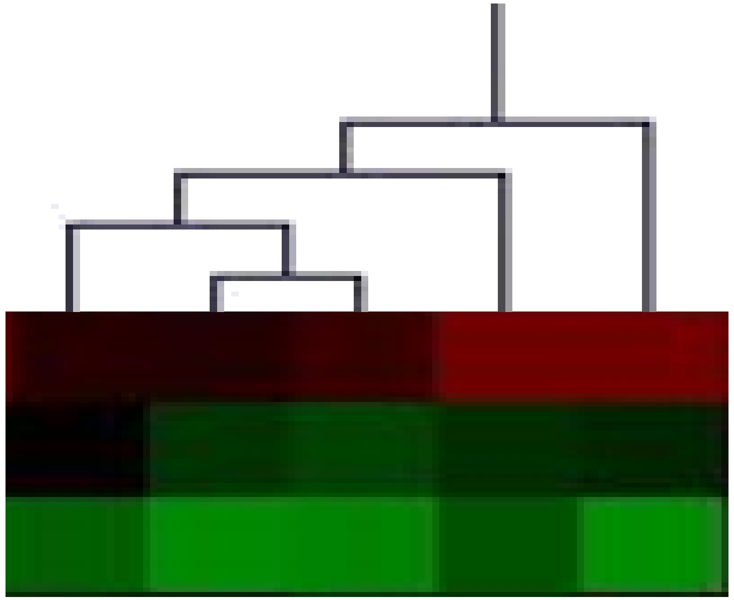
GO Term	Description	Target size	Common elements
GO:0000398	**nuclear mRNA splicing, via spliceosome**	**84**	**33**
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	84	33
GO:0000375	RNA splicing, via transesterification reactions	88	33
GO:0008380	RNA splicing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

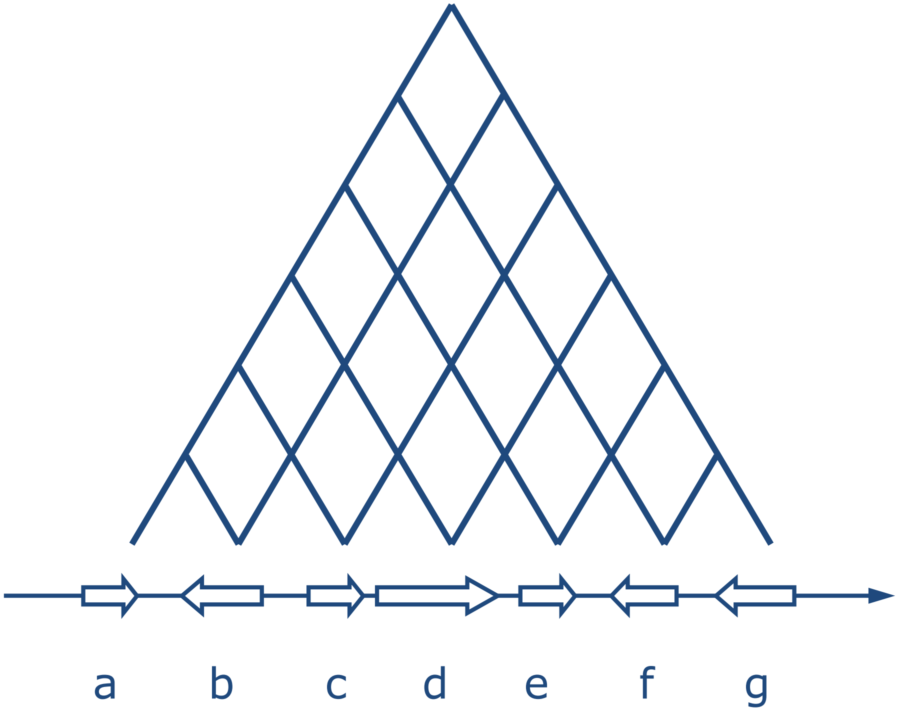


(((a (b c)) d) e)

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



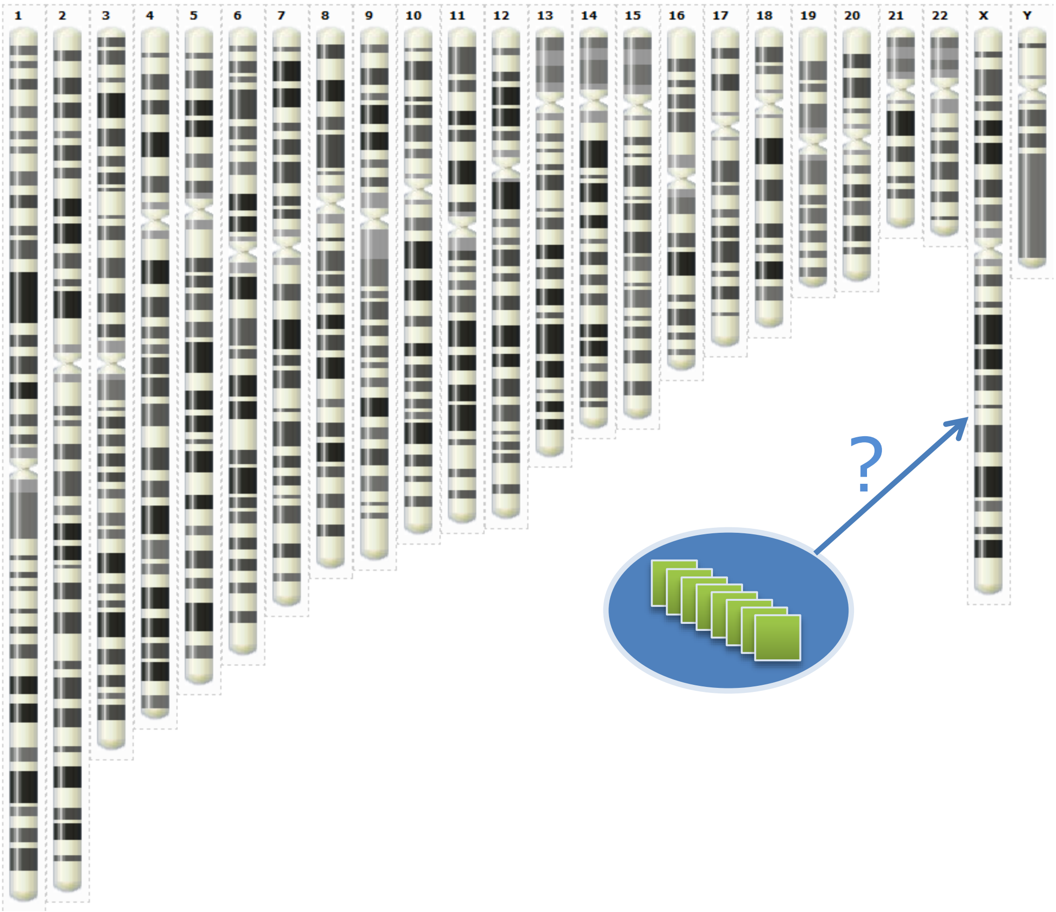
implicit

Context

Set of genes of interest

Examples

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



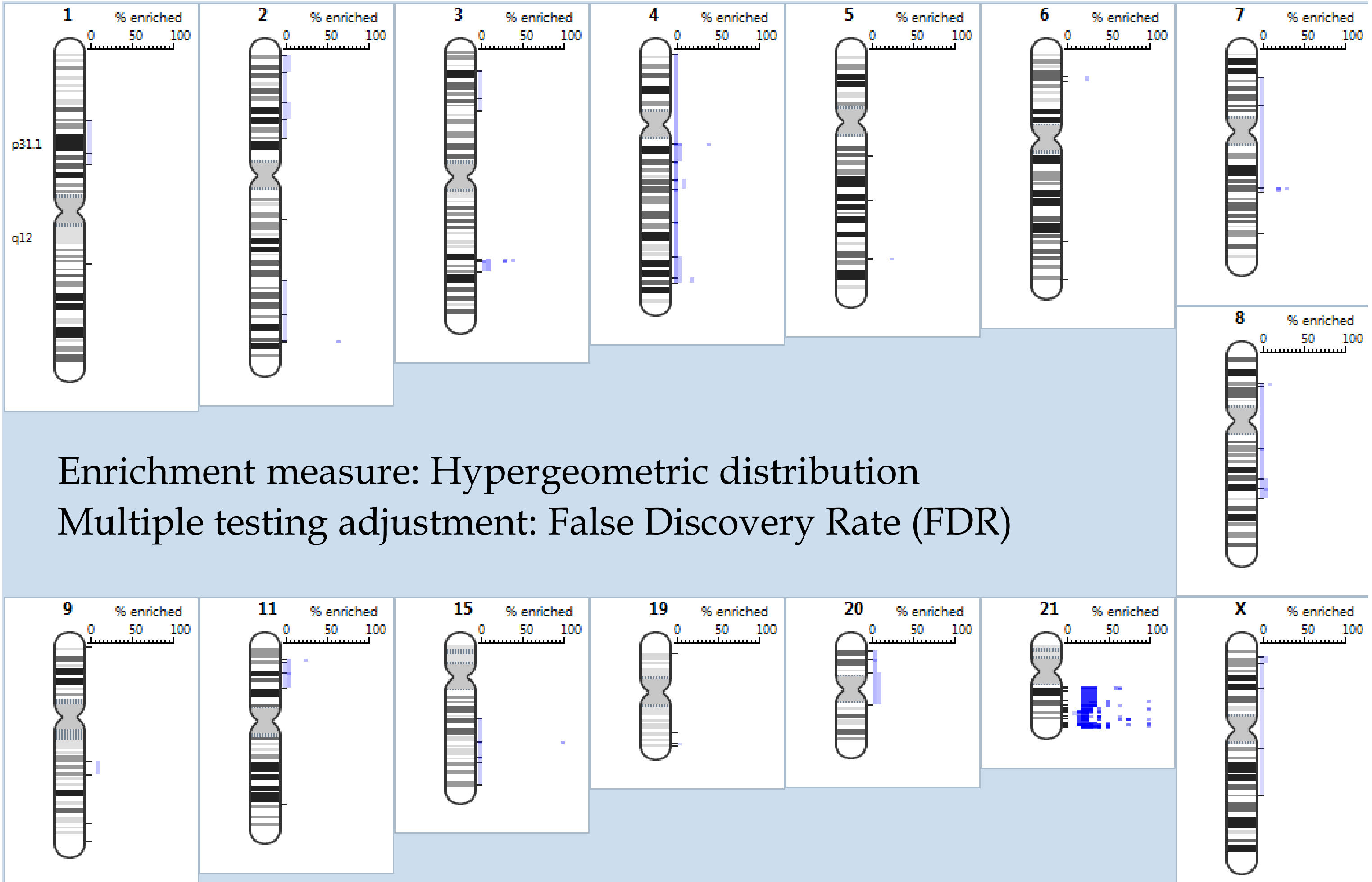
→ Question: Do those genes surprisingly cluster in the genome?

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)

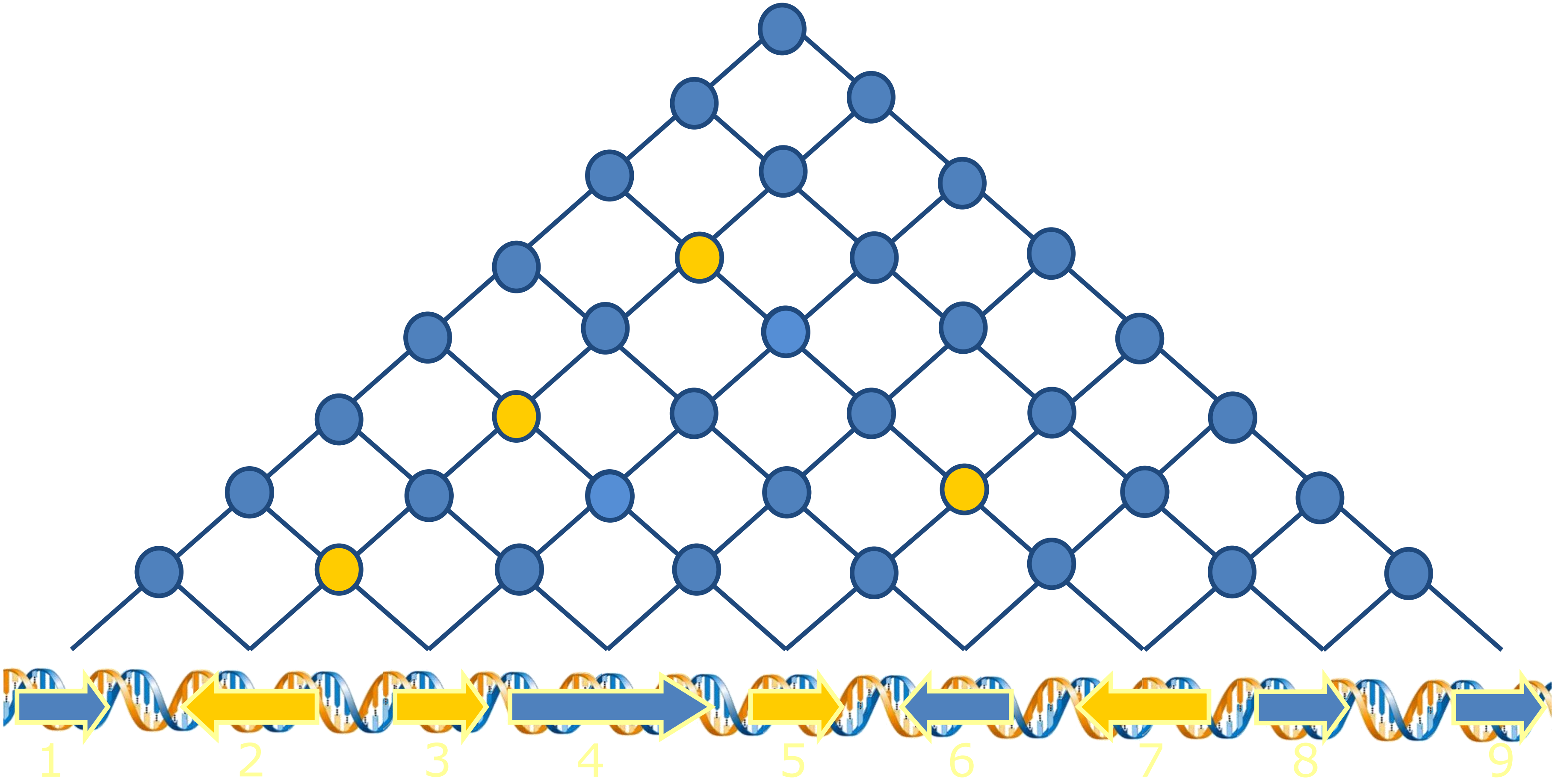
Issues:

- Number of regions to test
- False positives
- Redundancy

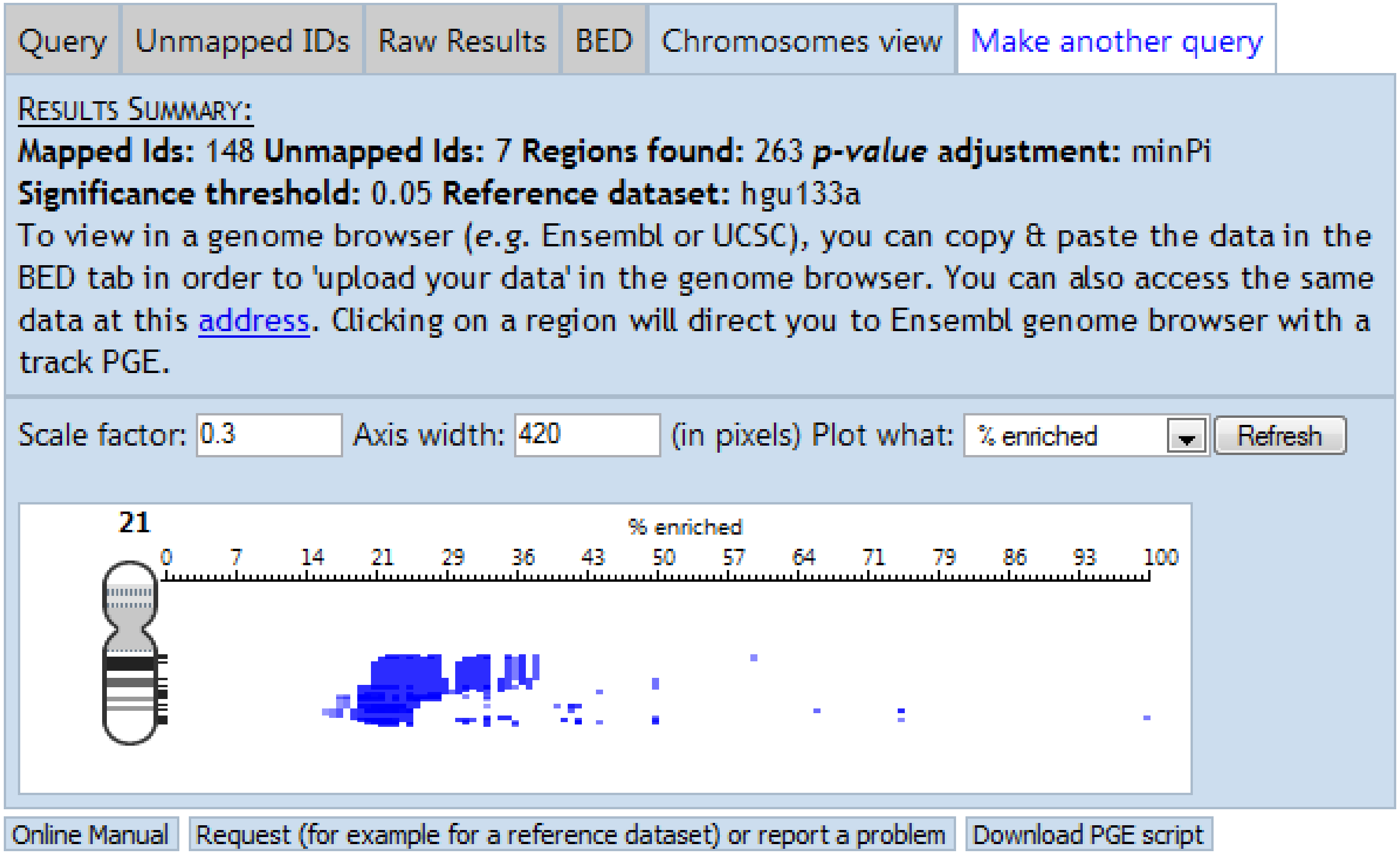
Pertinent Regions

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

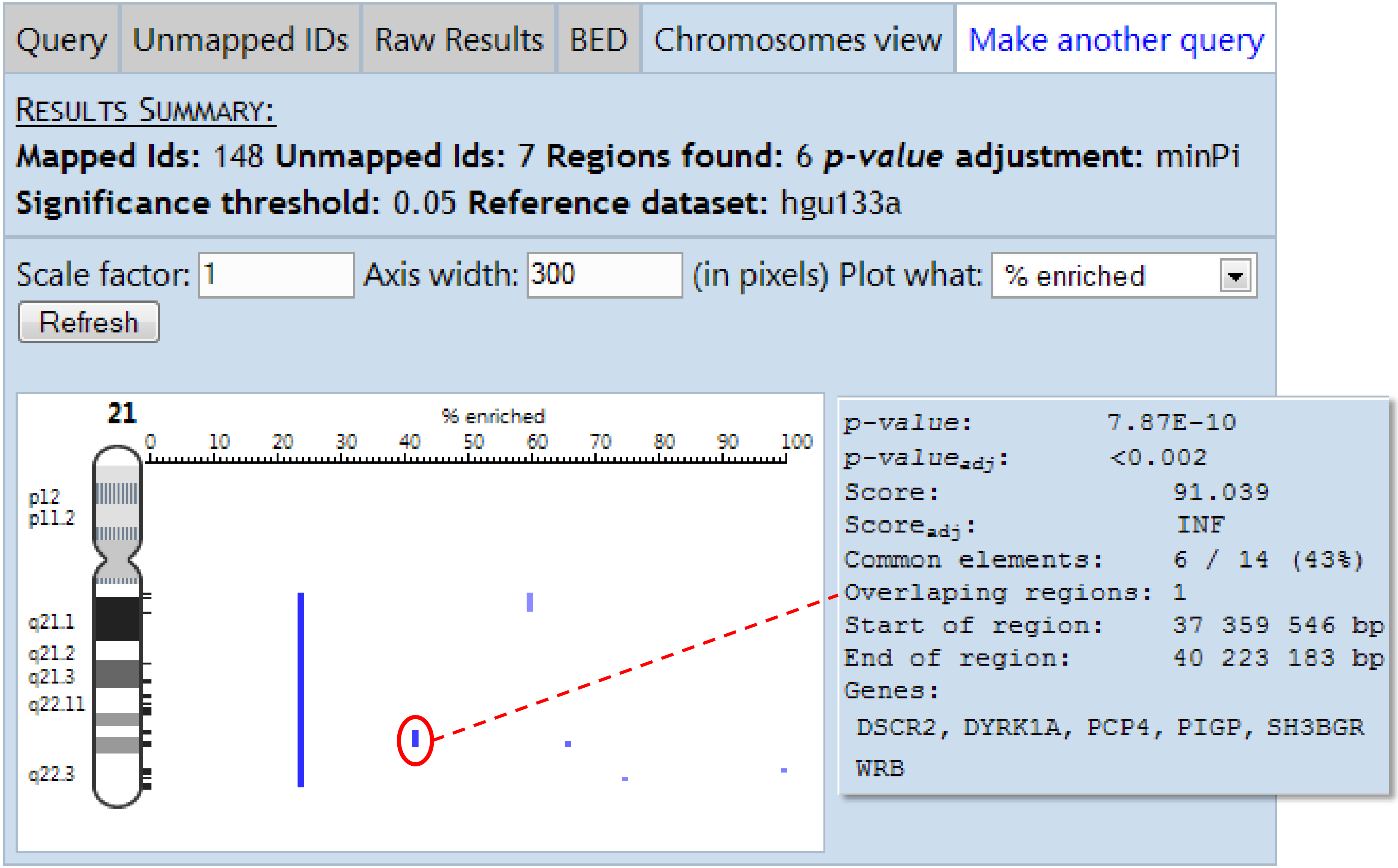
To view in a genome browser (*e.g.* Ensembl or UCSC), you can copy & paste the data in the BED tab in order to 'upload your data' in the genome browser. You can also access the same data at this address. Clicking on a region will direct you to Ensembl genome browser with a track PGE.

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



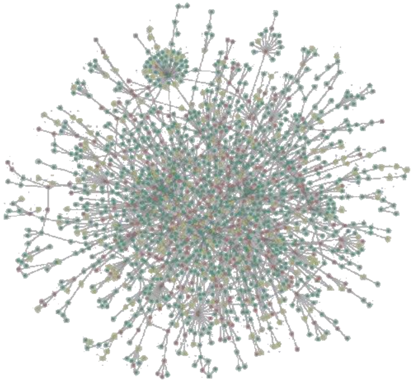
RESULTS SUMMARY:

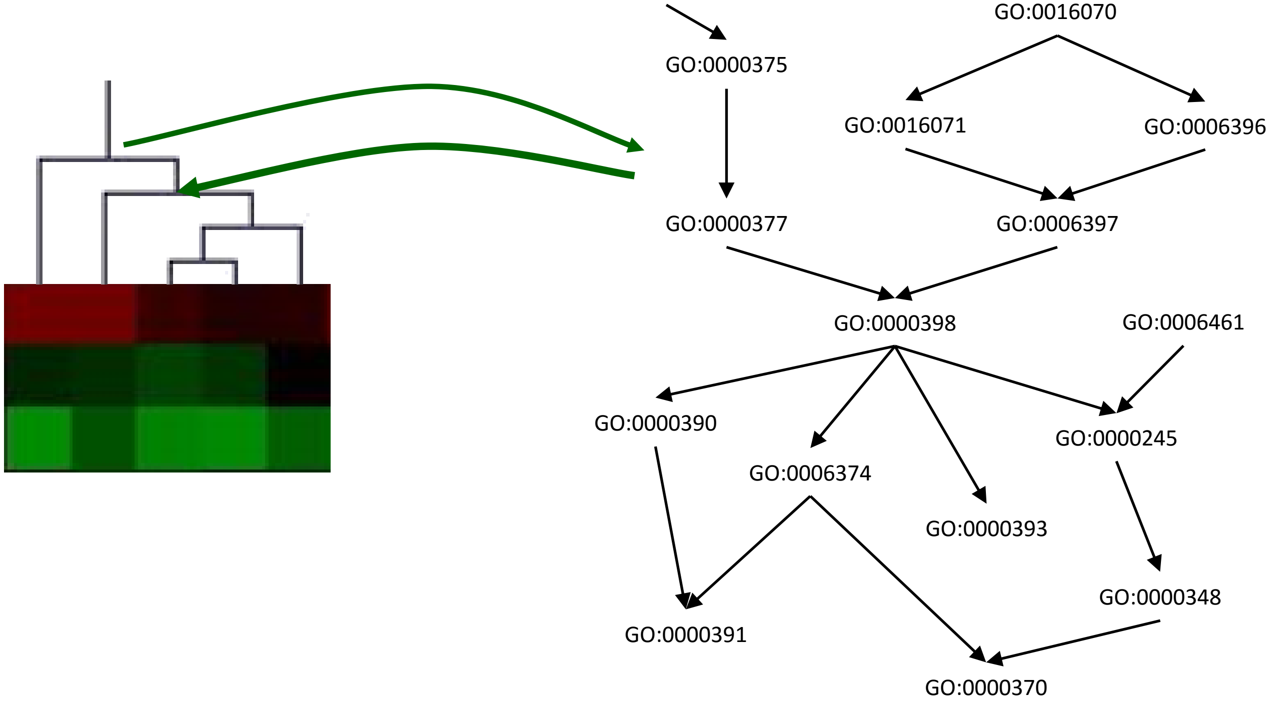
Mapped Ids: 148 **Unmapped Ids:** 7 **Regions found:** 6 ***p-value* adjustment:** minPi
Significance threshold: 0.05 **Reference dataset:** hgu133a

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
```

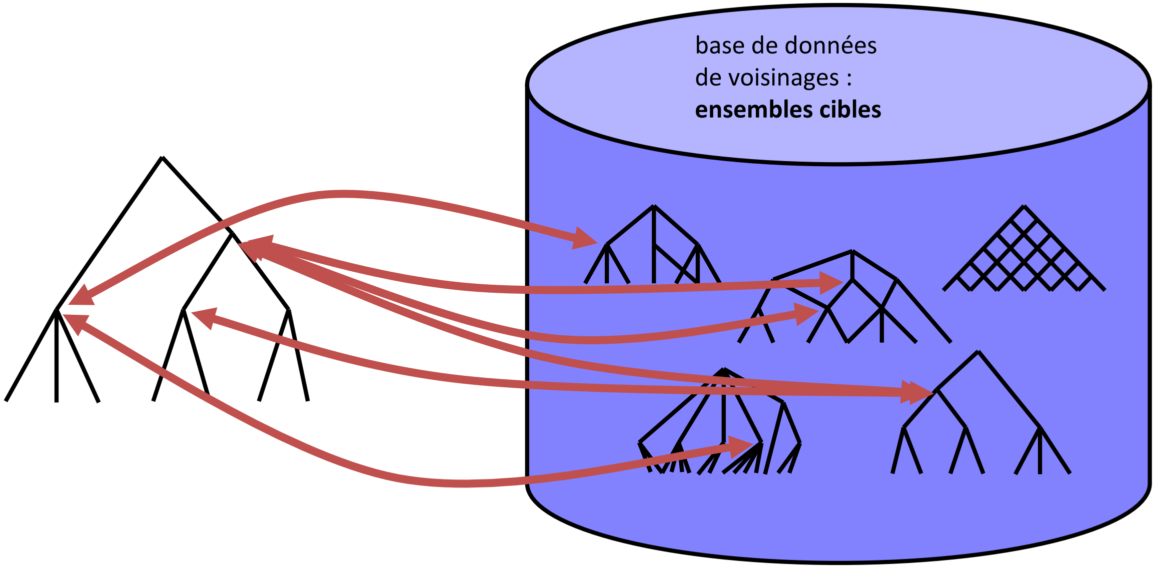
Challenges actuels





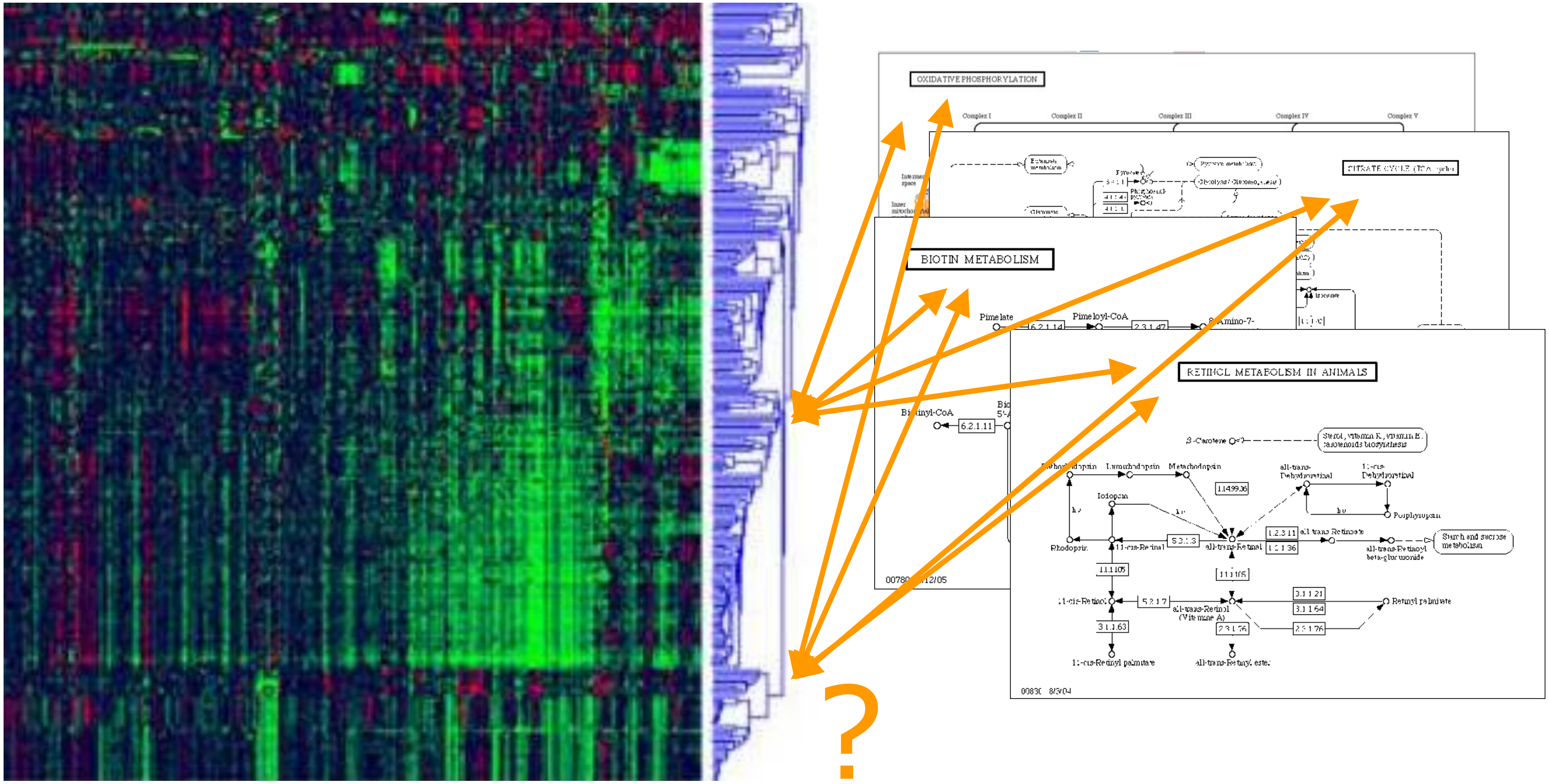
GO:0000375
GO:0016070
GO:0016071
GO:0006396
GO:0000377
GO:0006397
GO:0000398
GO:0006461
GO:0000390
GO:0006374
GO:0000245
GO:0000393
GO:0000348
GO:0000391
GO:0000370

Ensembles requêtes



base de données
de voisinages :
ensembles cibles

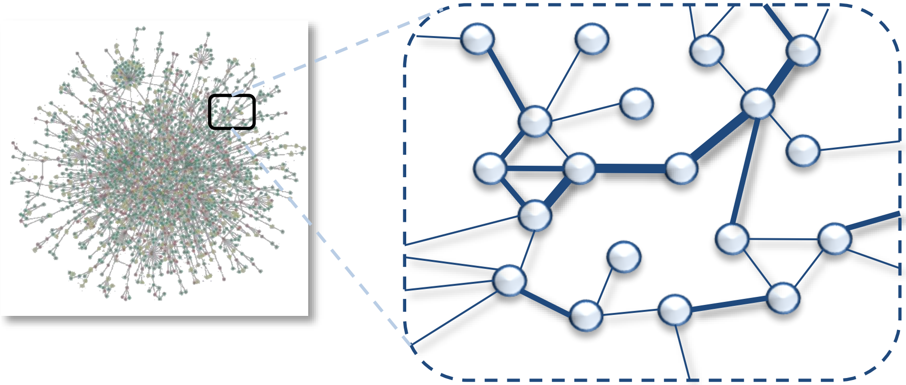
Analyse de données d’expression



[Ferea et al., 1999]

[Kanehisa et Goto, 2000]

Extraction de sous-graphe pertinent & visualisation



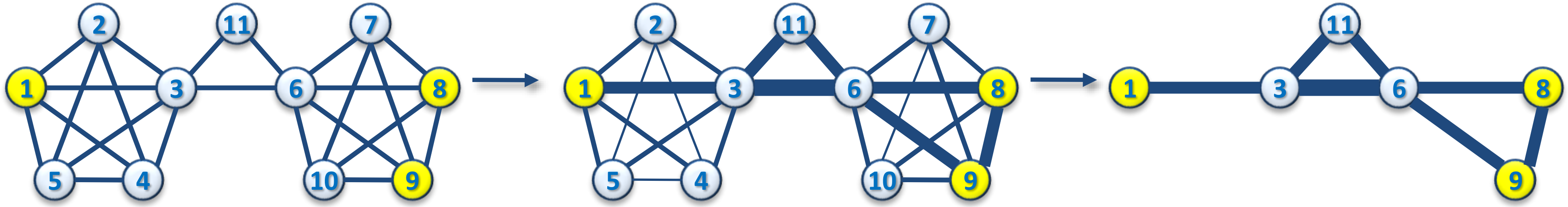
Idée :

- 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

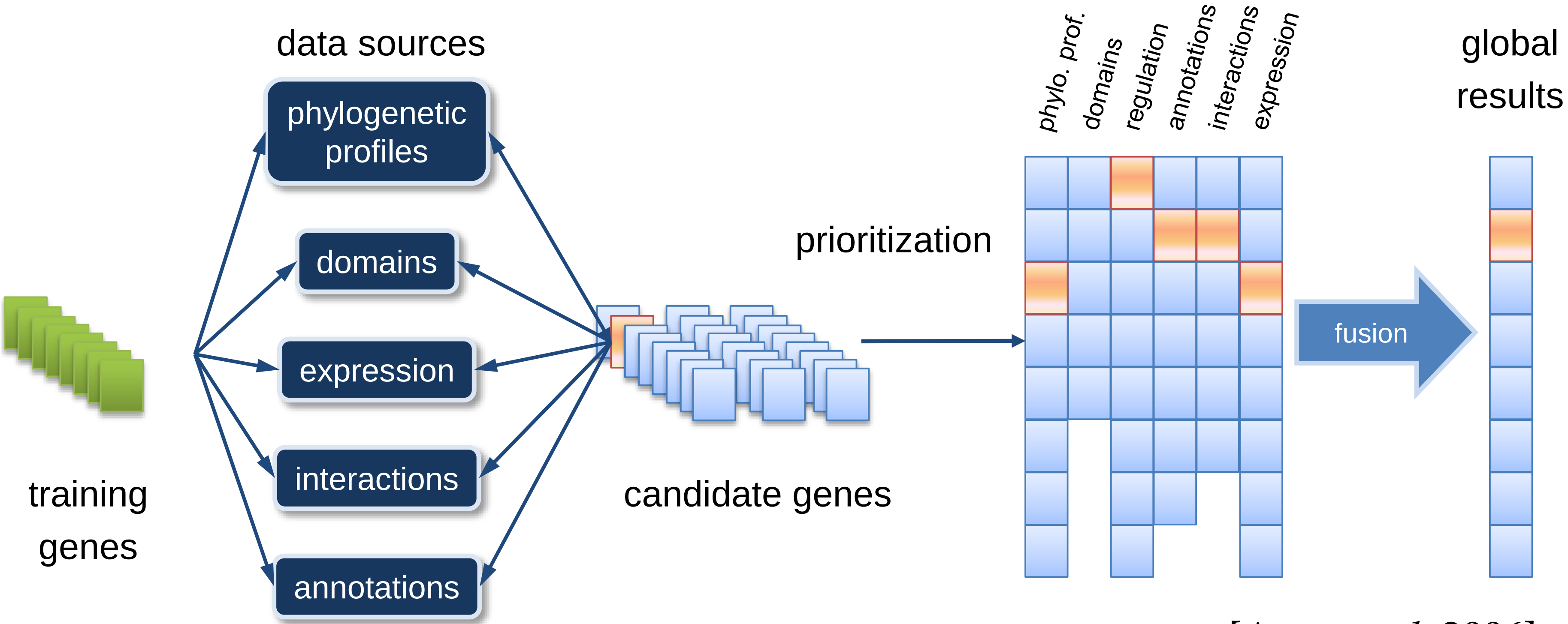
Surreprésentation :
sous-graphe pertinent



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

Candidate gene prioritization by genomic data fusion

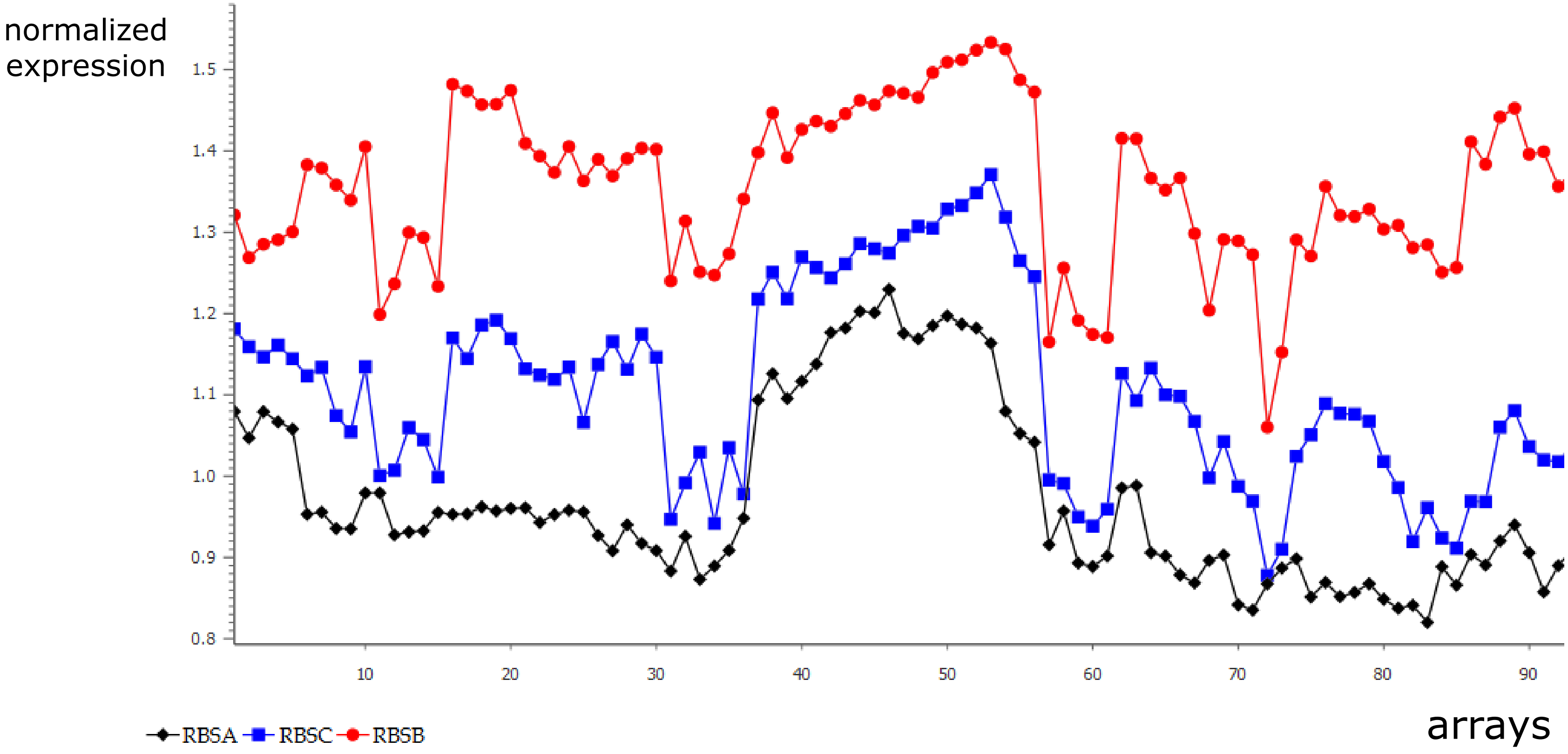
- Observation: more and more post-genomic data available
- Paradox: more difficult to select the best candidates
- Goal: objective and comprehensive evaluation of candidates



[Aerts *et al*. 2006]

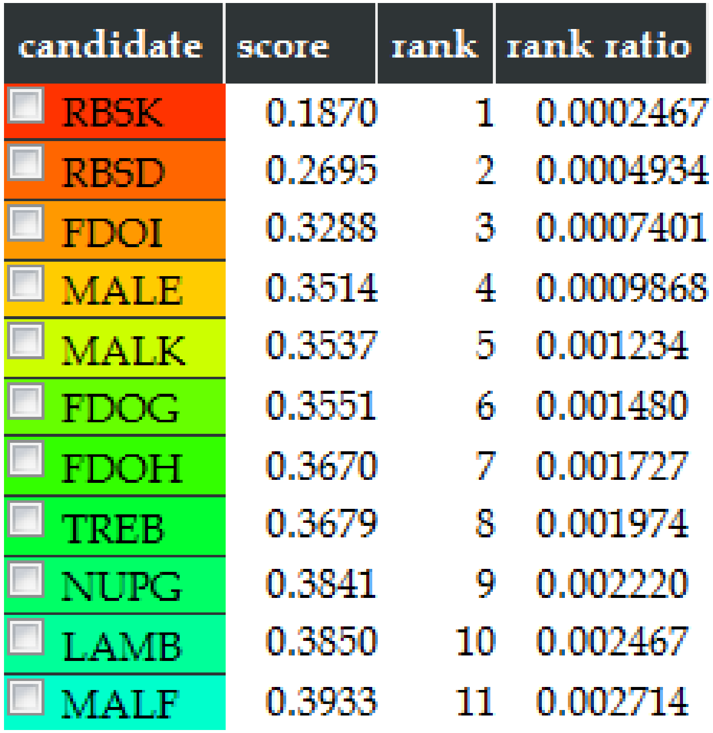
Gene expression

- 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

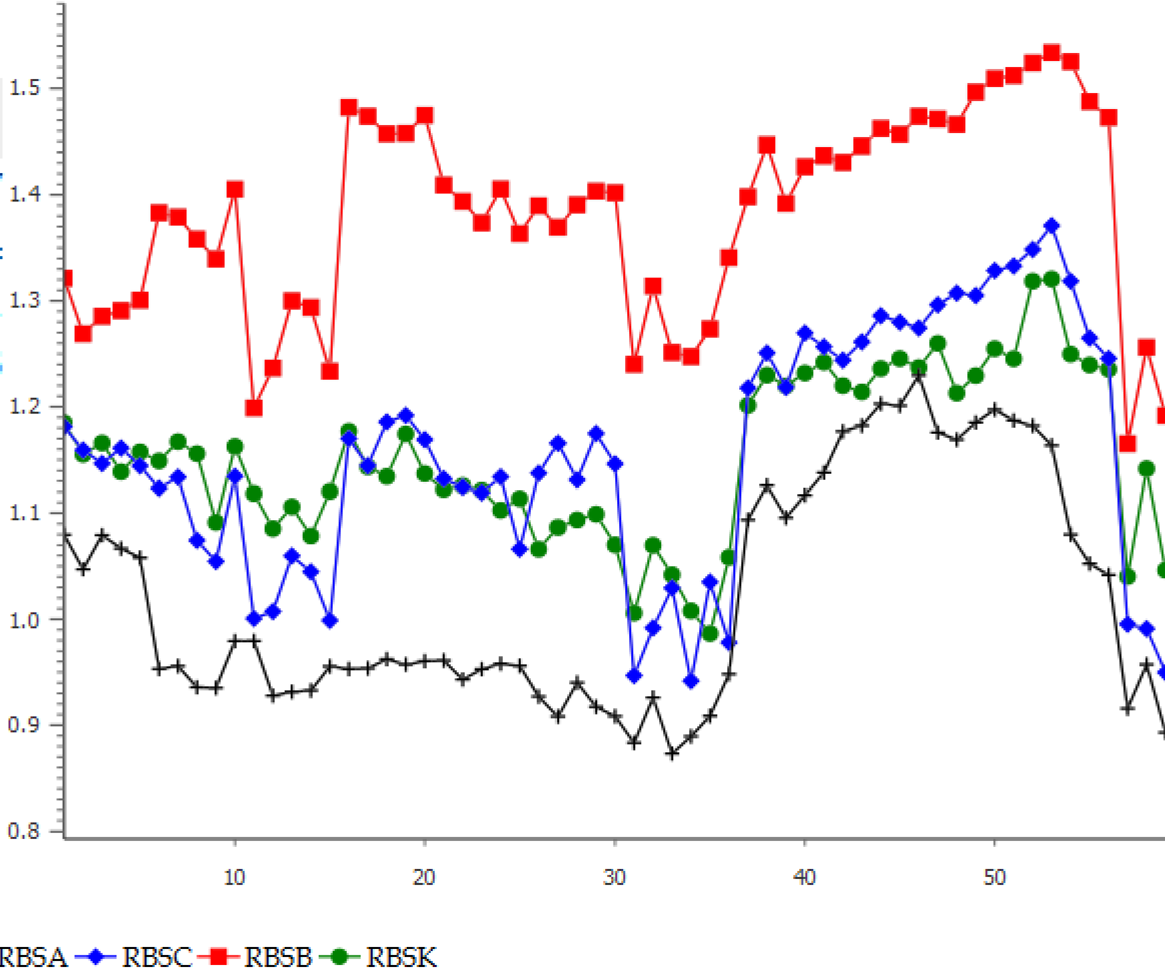


Gene expression example

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

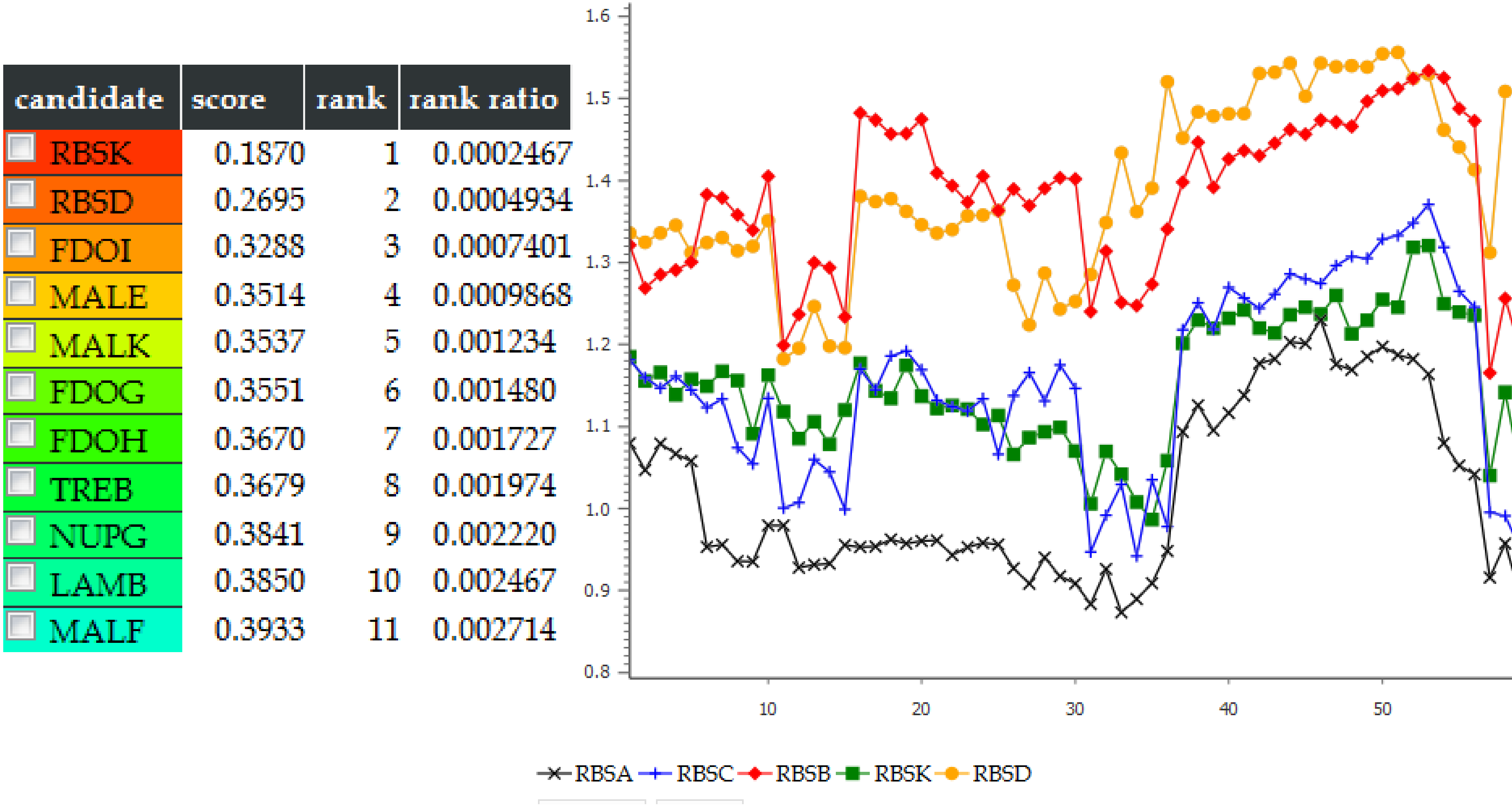


candidate	score	rank	rank ratio
□ RBSK	0.1870	1	0.0002467
□ RBSD	0.2695	2	0.0004934
□ FDOI	0.3288	3	0.0007401
□ MALE	0.3514	4	0.0009868
□ MALK	0.3537	5	0.001234
□ FDOG	0.3551	6	0.001480
□ FDOH	0.3670	7	0.001727
□ TREB	0.3679	8	0.001974
□ NUPG	0.3841	9	0.002220
□ LAMB	0.3850	10	0.002467
□ MALF	0.3933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



Gene expression example

- training: rbsA, rbsB, rbsC

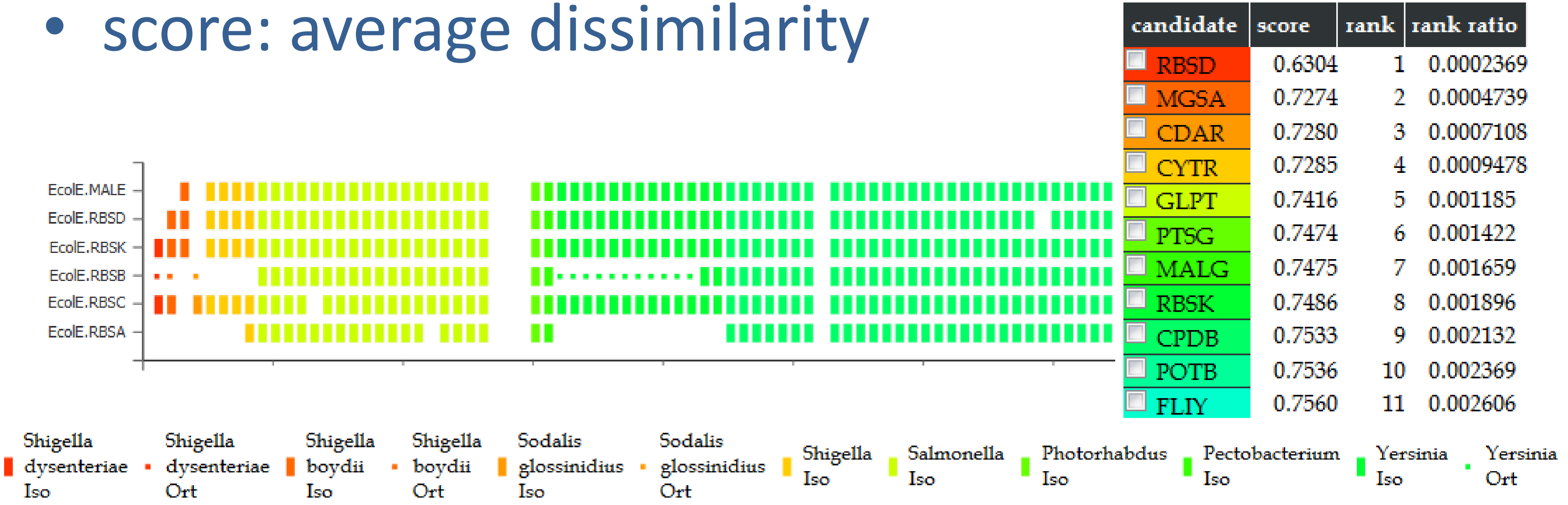


candidate	score	rank	rank ratio
□ RBSK	0.1870	1	0.0002467
□ RBSD	0.2695	2	0.0004934
□ FDOI	0.3288	3	0.0007401
□ MALE	0.3514	4	0.0009868
□ MALK	0.3537	5	0.001234
□ FDOG	0.3551	6	0.001480
□ FDOH	0.3670	7	0.001727
□ TREB	0.3679	8	0.001974
□ NUPG	0.3841	9	0.002220
□ LAMB	0.3850	10	0.002467
□ MALF	0.3933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

Phylogenetic profiles

- 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



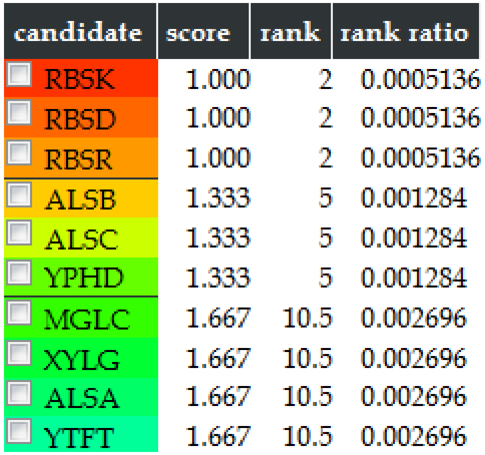
candidate	score	rank	rank ratio
□ RBSD	0.6304	1	0.0002369
□ MGSA	0.7274	2	0.0004739
□ CDAR	0.7280	3	0.0007108
□ CYTR	0.7285	4	0.0009478
□ GLPT	0.7416	5	0.001185
□ PTSG	0.7474	6	0.001422
□ MALG	0.7475	7	0.001659
□ RBSK	0.7486	8	0.001896
□ CPDB	0.7533	9	0.002132
□ POTB	0.7536	10	0.002369
□ FLIY	0.7560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

Interactions

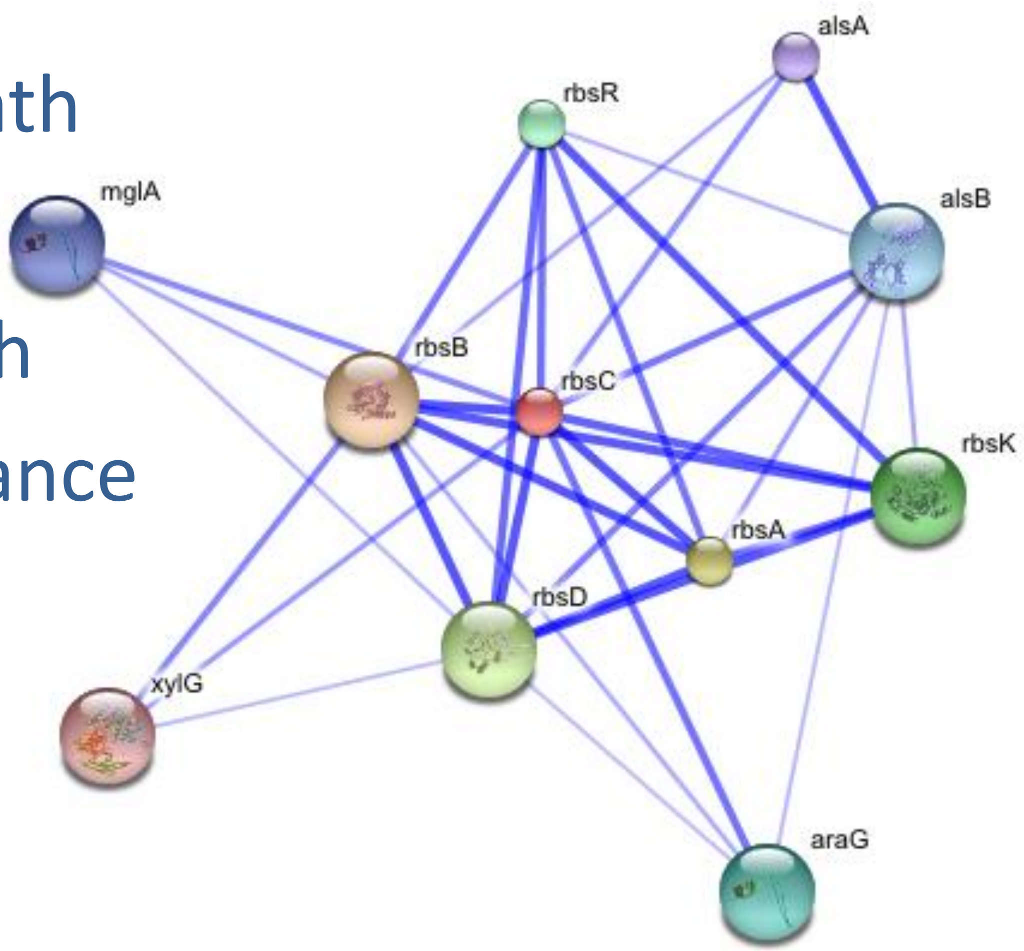
- all pairs shortest path
- a pair of gene:

  shortest path length
- score: average distance

training: rbsA, rbsB, rbsC



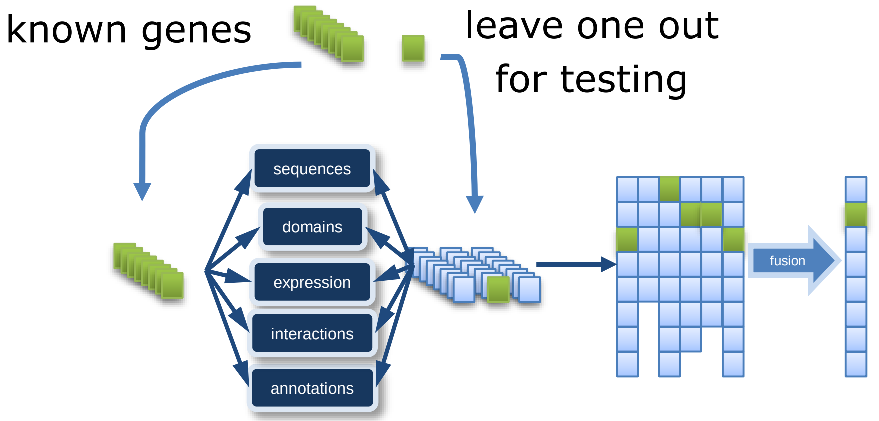
candidate	score	rank	rank ratio
□ RBSK	1.000	2	0.0005136
□ RBSD	1.000	2	0.0005136
□ RBSR	1.000	2	0.0005136
□ ALSB	1.333	5	0.001284
□ ALSC	1.333	5	0.001284
□ YPHD	1.333	5	0.001284
□ MGLC	1.667	10.5	0.002696
□ XYLG	1.667	10.5	0.002696
□ ALSA	1.667	10.5	0.002696
□ YTFT	1.667	10.5	0.002696



from STRING
http://string-db.org

Evaluation methodology

- Leave-one-out cross validation (LOOCV)



known genes

leave one out for testing

sequences

domains

expression

interactions

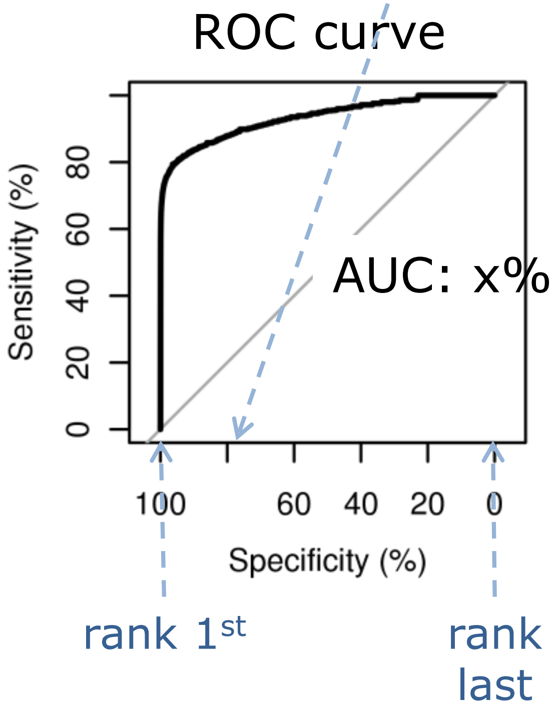
annotations

fusion

- 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



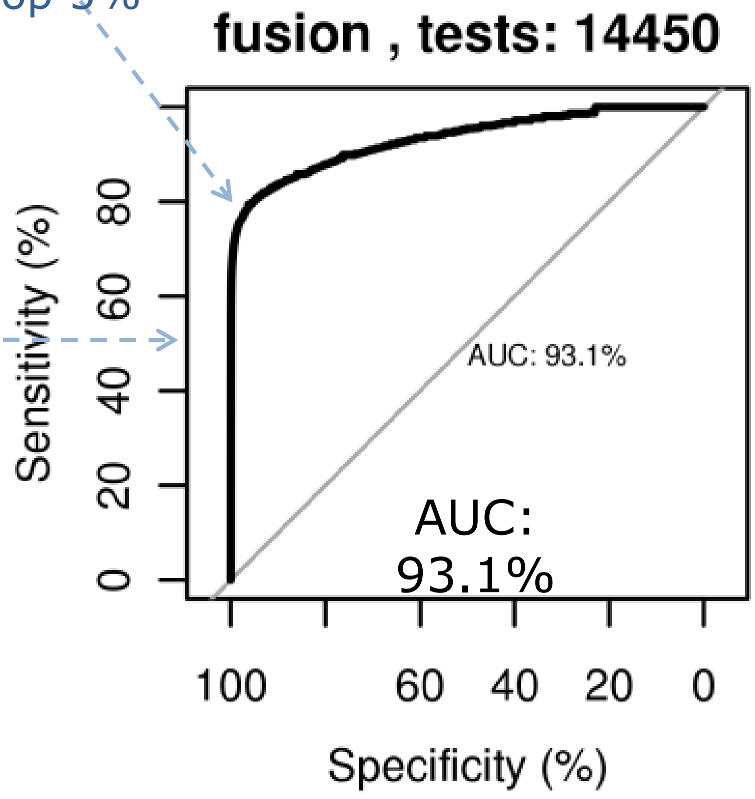
ROC curve

AUC: x%

Evaluation: global results

Gold standard

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

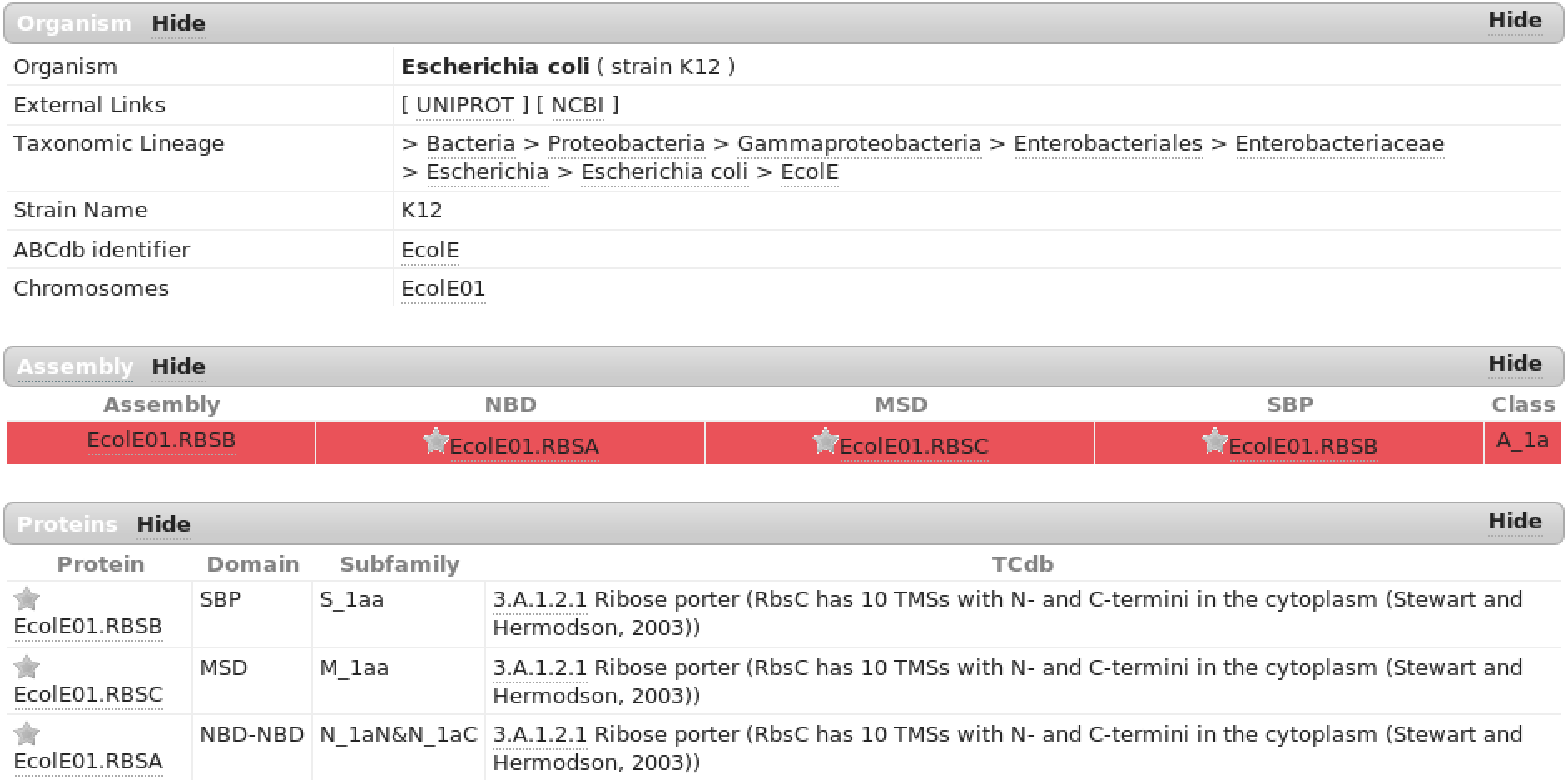


80% of the left out genes rank in the top 5%

53% of the left out genes rank 1st

AUC: 93.1%

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

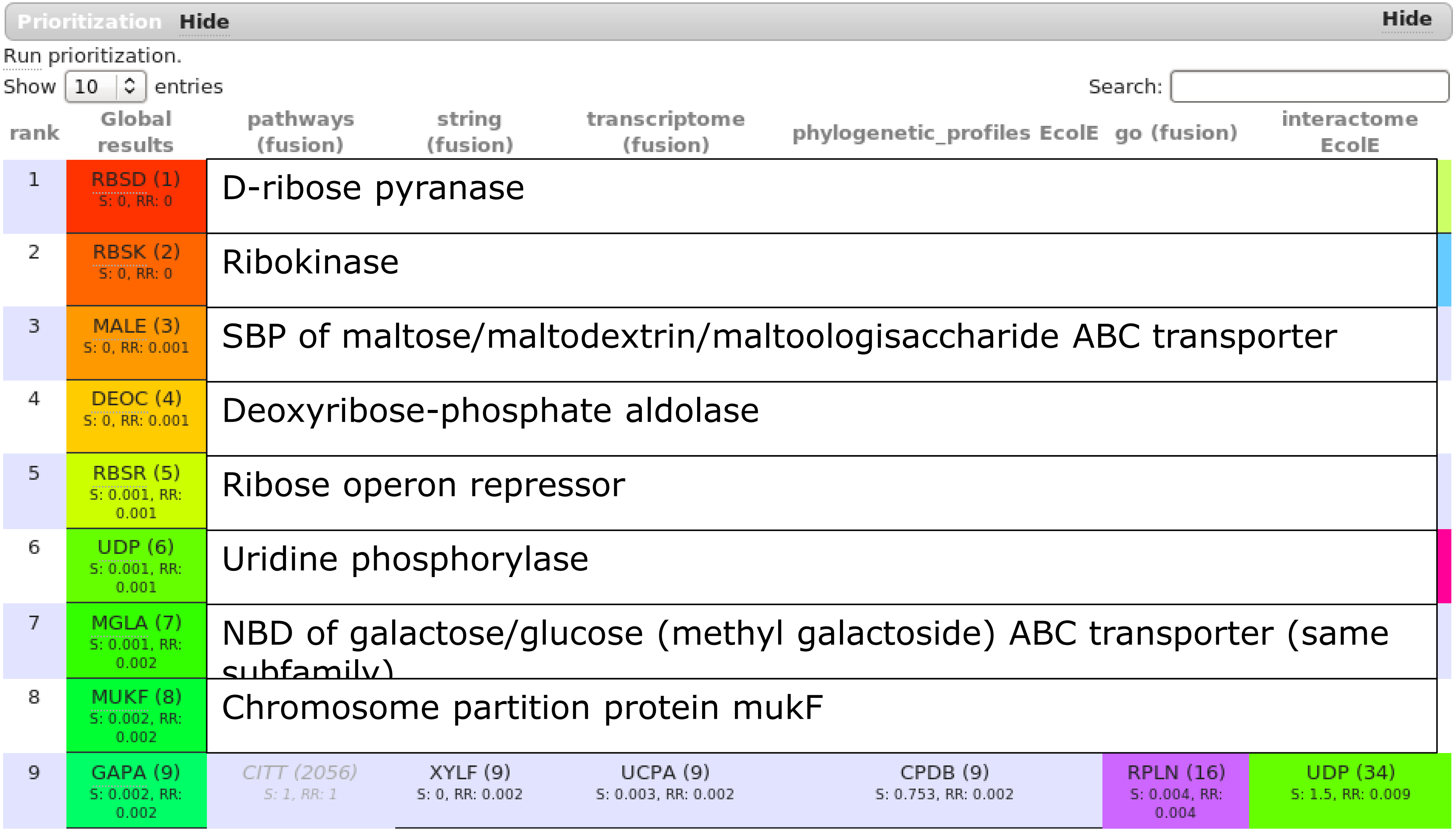
Proteins Hide Hide

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

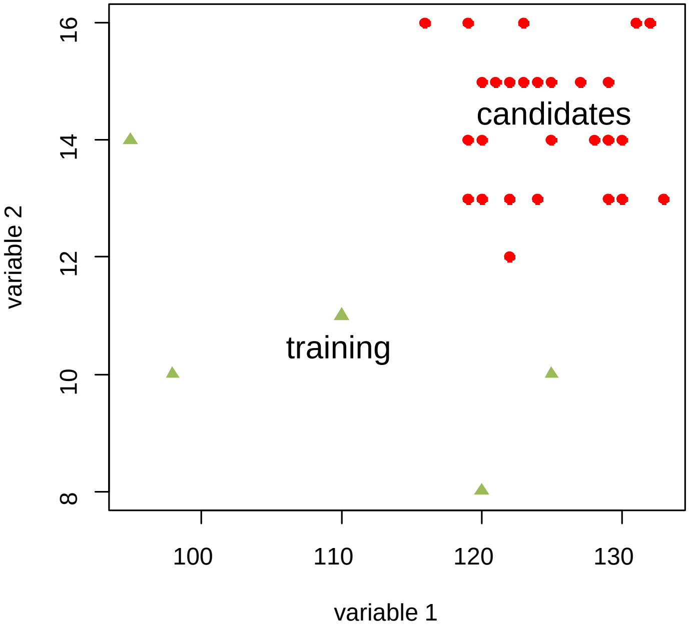
Run prioritization.

Show 10 ⇕ entries Search:

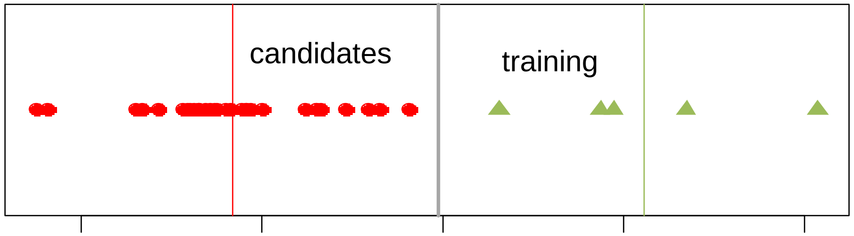
rank	Global results	pathways (fusion)	string (fusion)	transcriptome (fusion)	phylogenetic_profiles EcolE	go (fusion)	interactome EcolE
1	RBSD (1) S: 0, RR: 0	D-ribose pyranase					
2	RBSK (2) S: 0, RR: 0	Ribokinase					
3	MALE (3) S: 0, RR: 0.001	SBP of maltose/maltodextrin/maltoologisaccharide ABC transporter					
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

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)

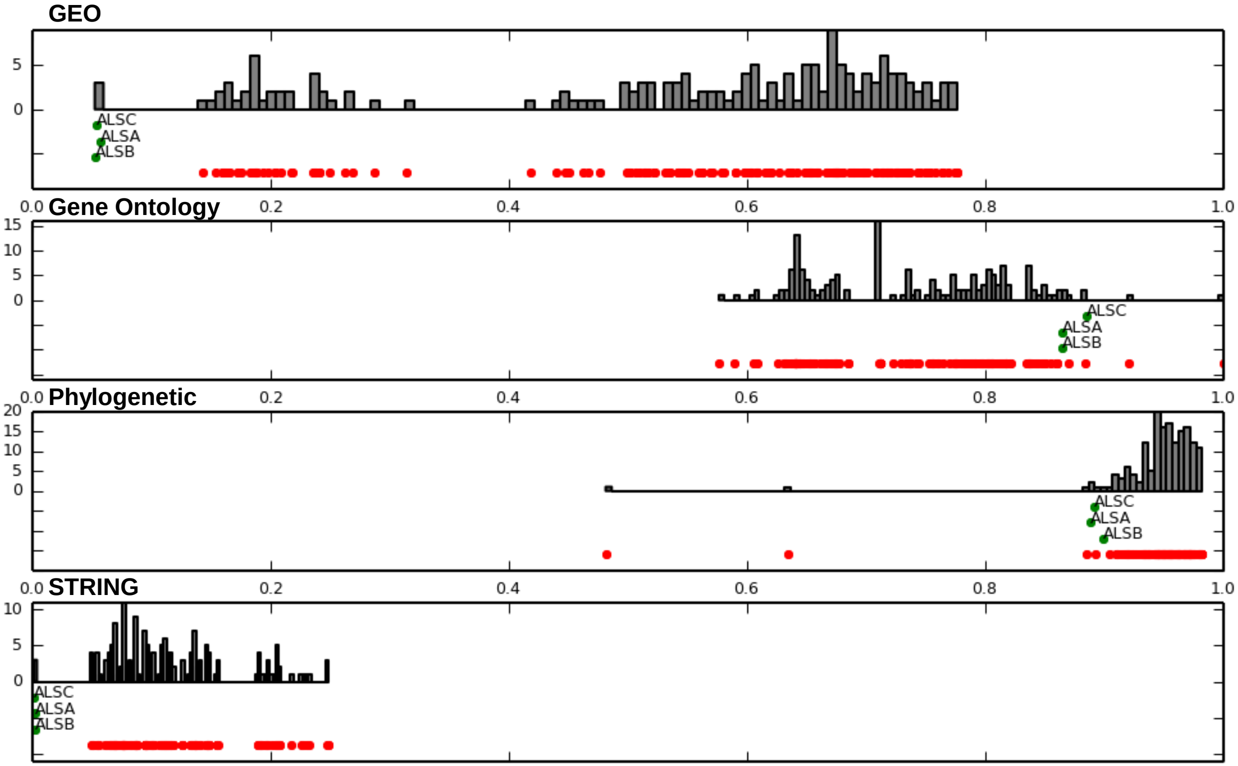


dimension	weight
variable 1	-0.1307346
variable 2	-0.7031850



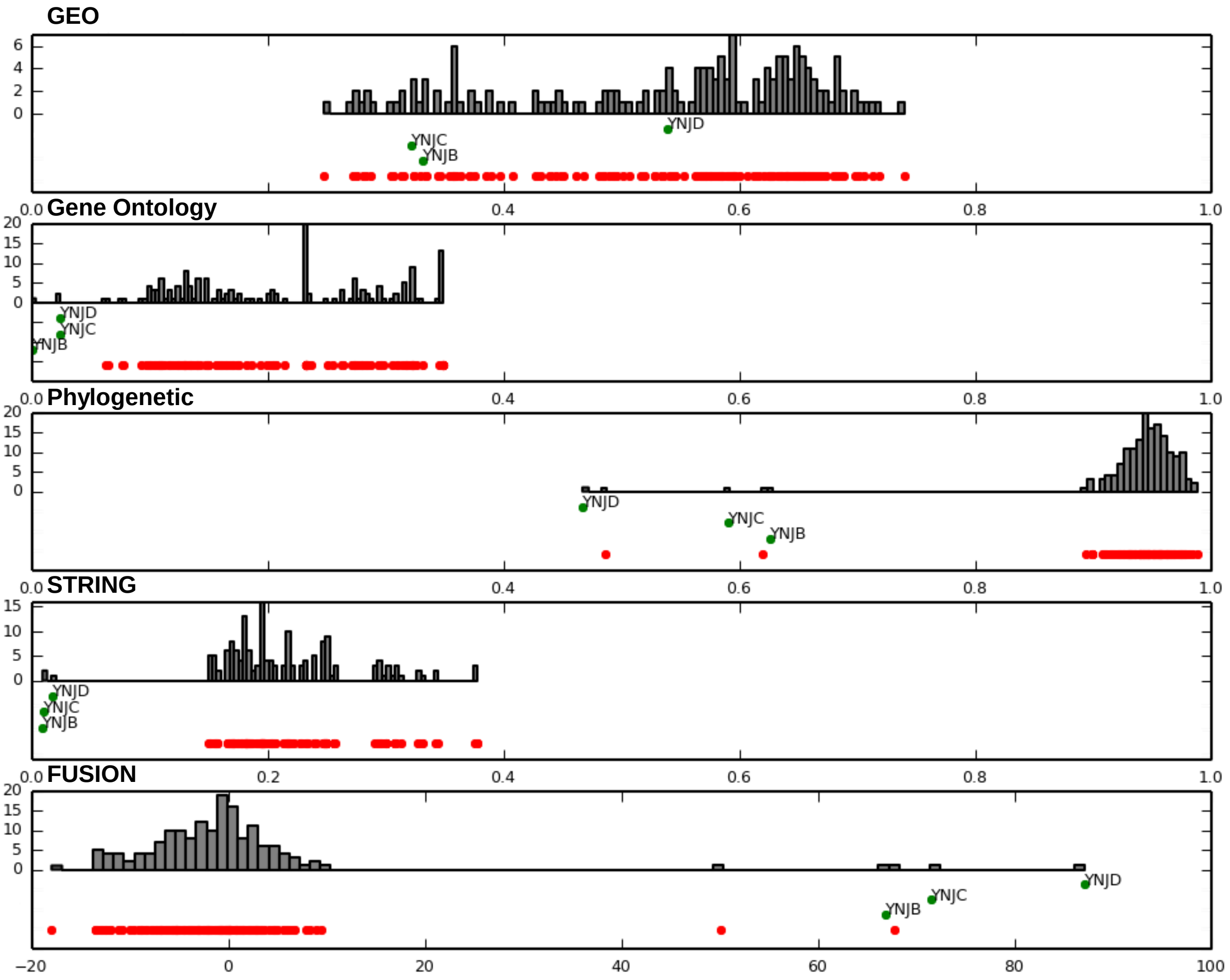
Application to *E. coli* alsA system: *alsA, alsB, alsC*

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

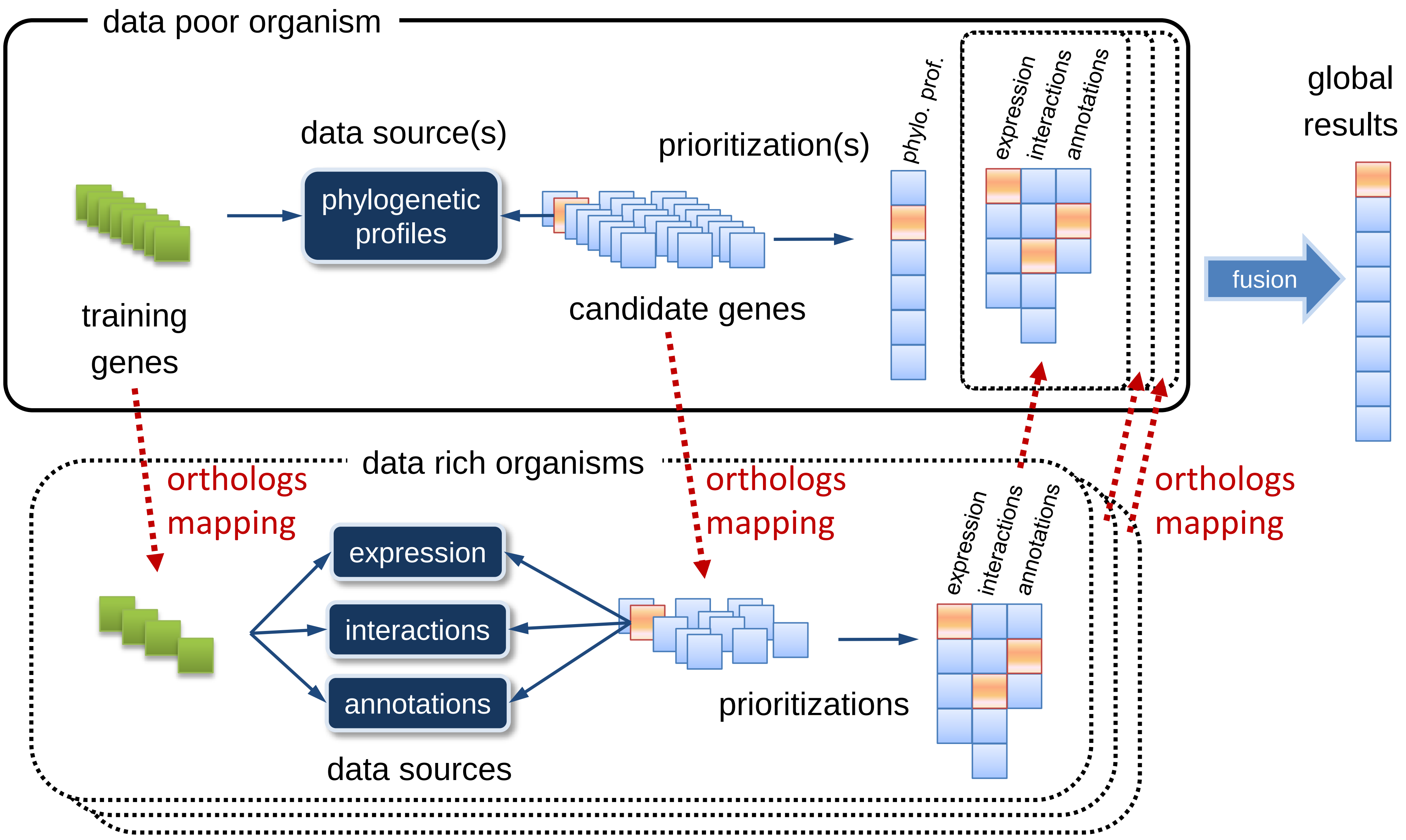


Application to *E. coli* ynjD system: *ynjB, ynjC, ynjD*

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

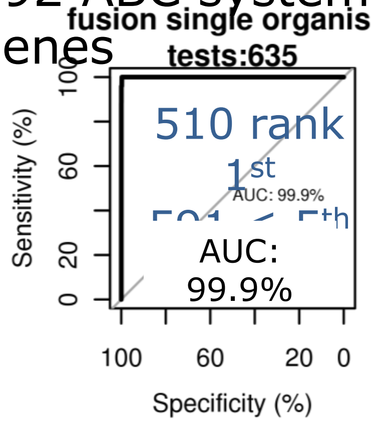


Extension of the method to data poor organisms



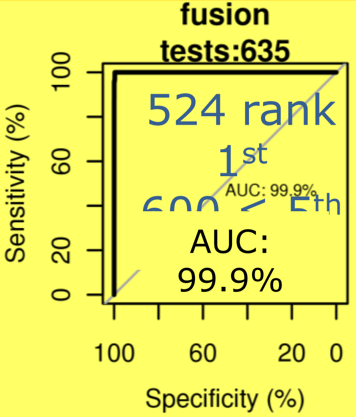
Performances: using other organisms data through orthology

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



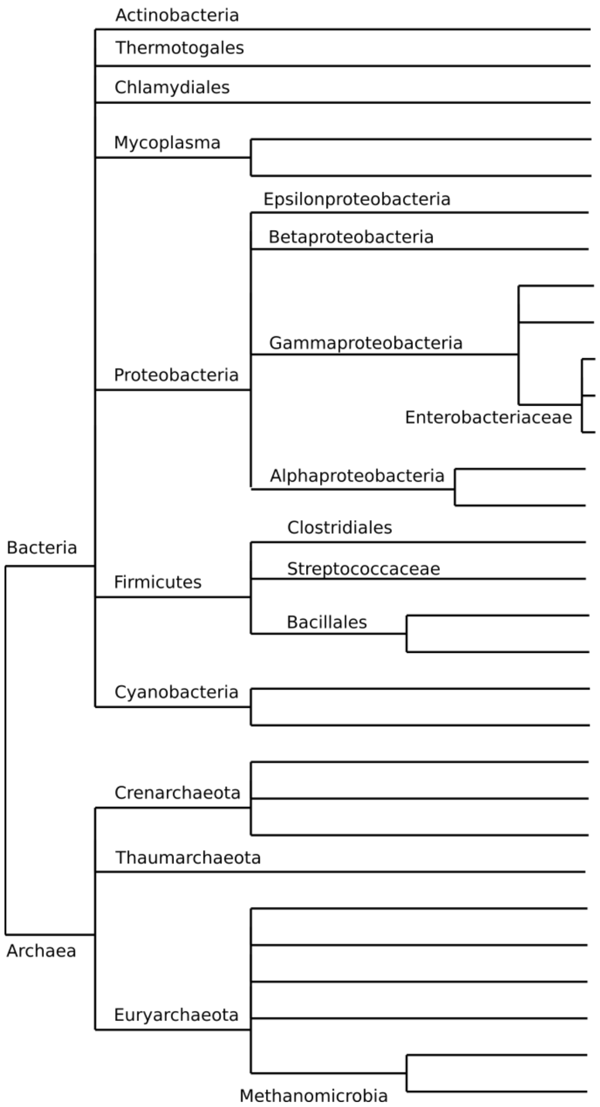
510 rank 1st
501 < 5th

AUC: 99.9%



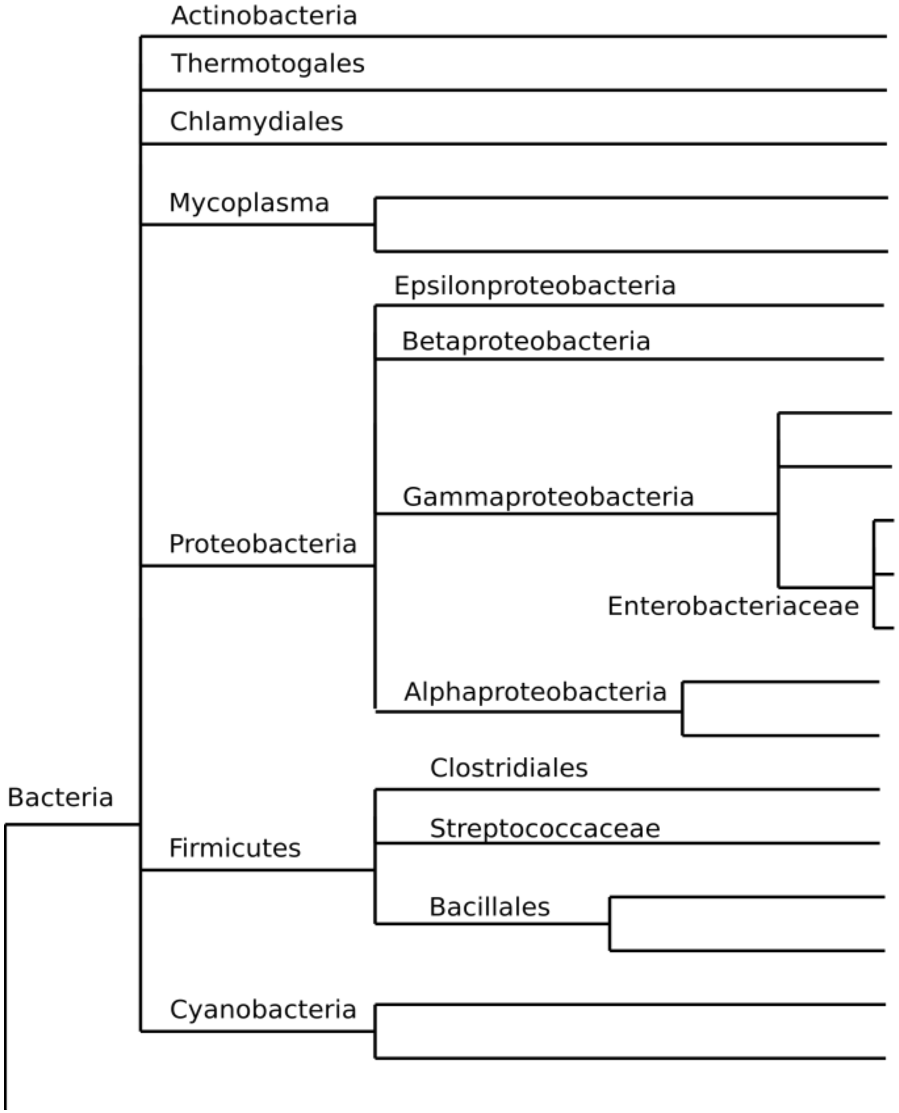
524 rank 1st
600 < 5th

AUC: 99.9%



	# ABC genes	AUC (%)	data sources
Streptomyces coelicolor	434	93	
Thermotoga maritima	170	95.3	
Chlamydia trachomatis	31	89.8	
Mycoplasma gallisepticum	47	82.4	
Mycoplasma genitalium	36	74.8	
Helicobacter pylori	42	90.7	
Nitrosomonas europaea	76	98	
Pseudomonas aeruginosa	285	99.9	★★★★
Coxiella burnetii	37	97.8	
Escherichia coli	216	99.9	★★★★
Salmonella enterica	198	98.4	★★
Shigella flexneri	192	99.3	★
Bradyrhizobium japonicum	619	99.9	★
Anaplasma marginale	18	96.2	
Clostridium perfringens	141	92.7	
Streptococcus pneumoniae	163	96.7	
Staphylococcus aureus	145	98.3	★★
Bacillus subtilis	208	99.9	★★★★
Nostoc sp.	234	96	★
Synechocystis sp.	132	96	
Thermofilum pendens	141	89.9	
Metallosphaera sedula	60	81.2	
Aeropyrum pernix	104	90.4	
Nitrosopumilus maritimus	33	90.2	
Methanocaldococcus jannaschii	40	95	
Thermococcus onnurineus	65	90.6	
Halobacterium sp.	83	93	★★
Methanosphaera stadtmanae	35	96.5	
Candidatus Methanoregula boonei	83	93.4	
Methanosarcina mazei	119	91.9	

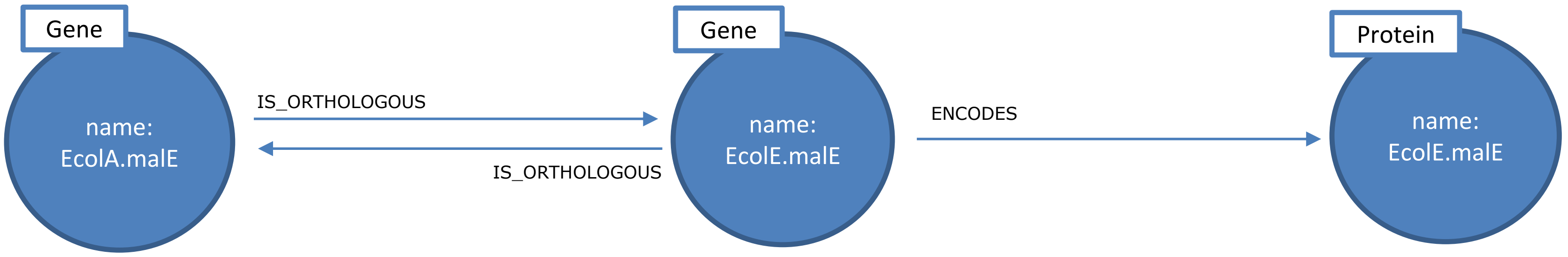
Performances: using other organisms data through orthology



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

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

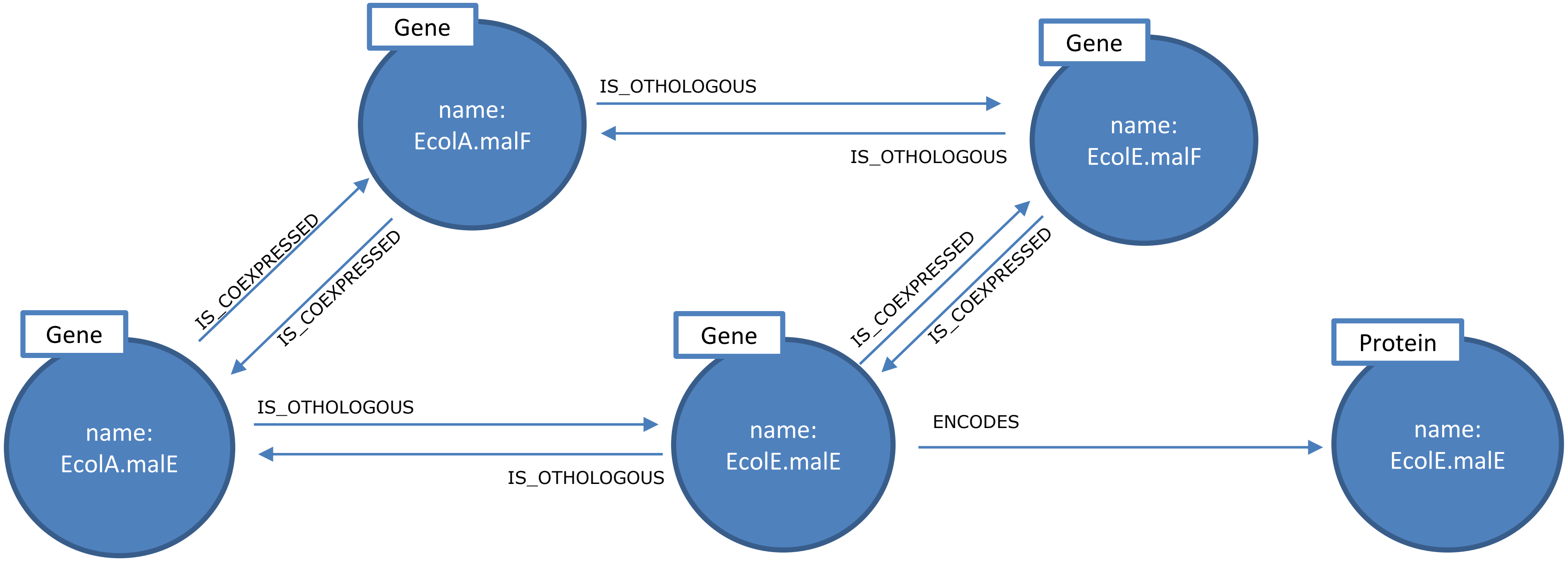


Labeled Property Graph

```
(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

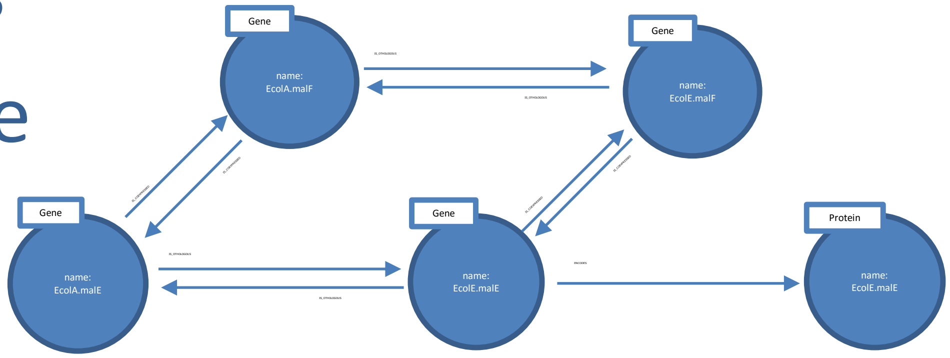


Labeled Property Graph

```
(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

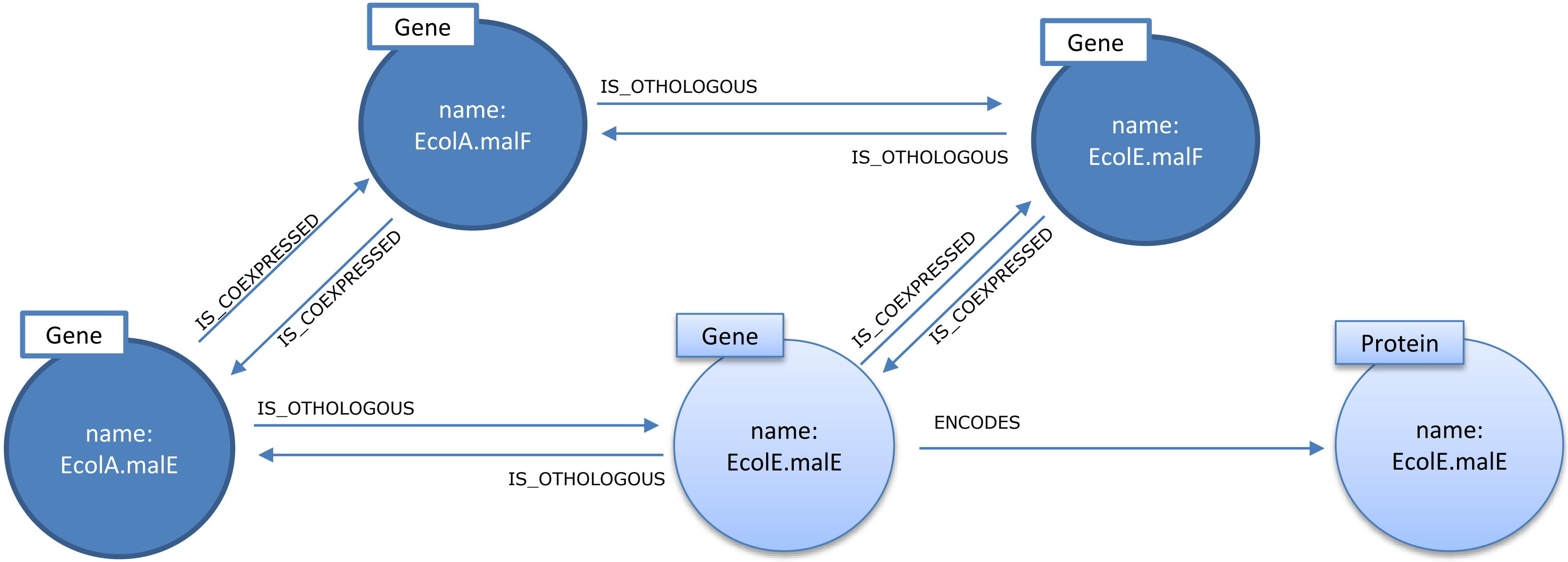
Un graphe avec propriétés étiquetées est constitué de sommets, relations, propriétés et étiquettes :



- 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
name:
EcolA.malF
IS_OTHOLOGOUS
IS_OTHOLOGOUS
Gene
name:
EcolE.malF
IS_COEXPRESSED
IS_COEXPRESSED
IS_COEXPRESSED
IS_COEXPRESSED
Gene
name:
EcolA.malE
IS_OTHOLOGOUS
IS_OTHOLOGOUS
Gene
name:
EcolE.malE
ENCODES
Protein
name:
EcolE.malE