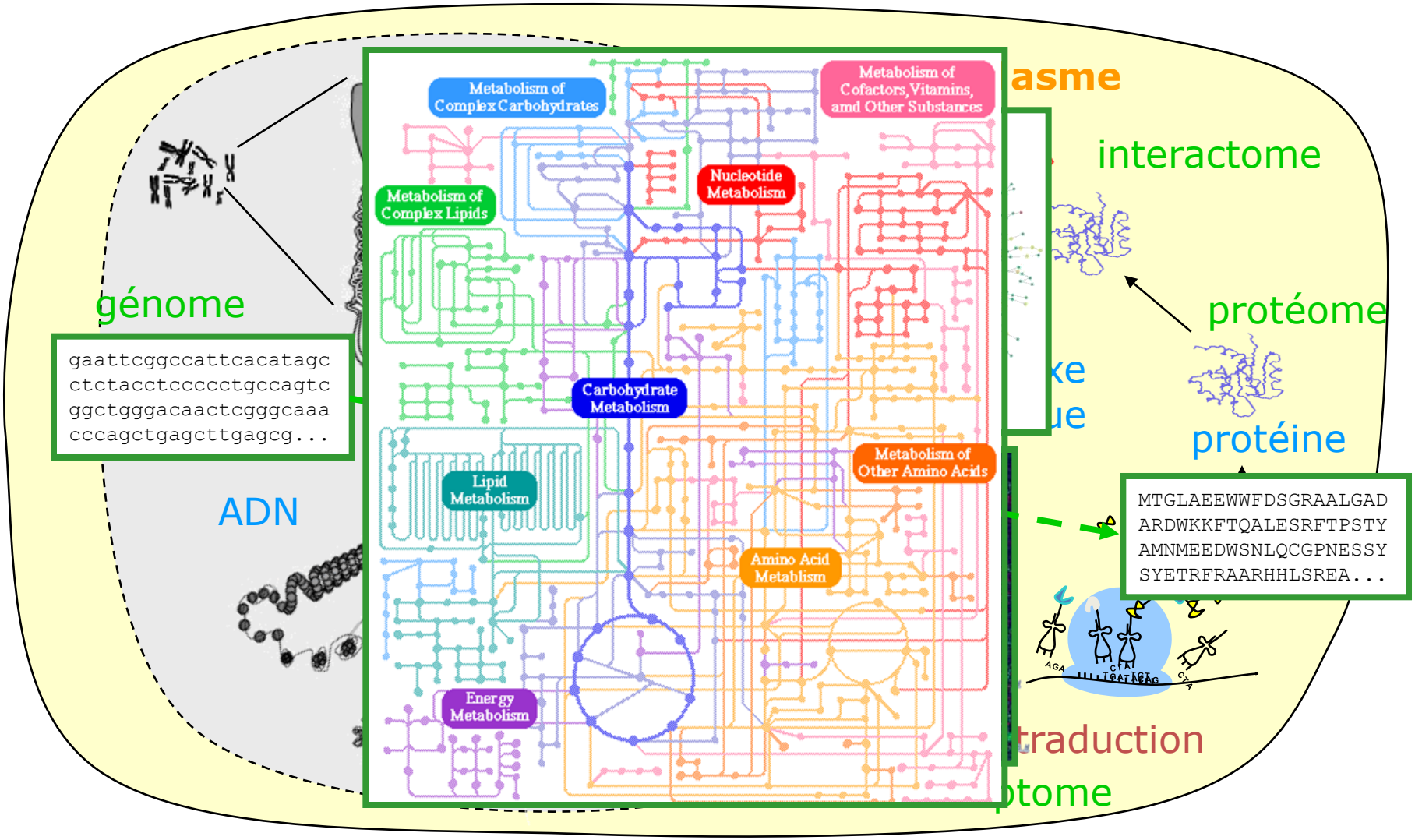
Intégration de données hétérogènes

Master 2 MABS

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

Biologie, mesures, données et connaissances

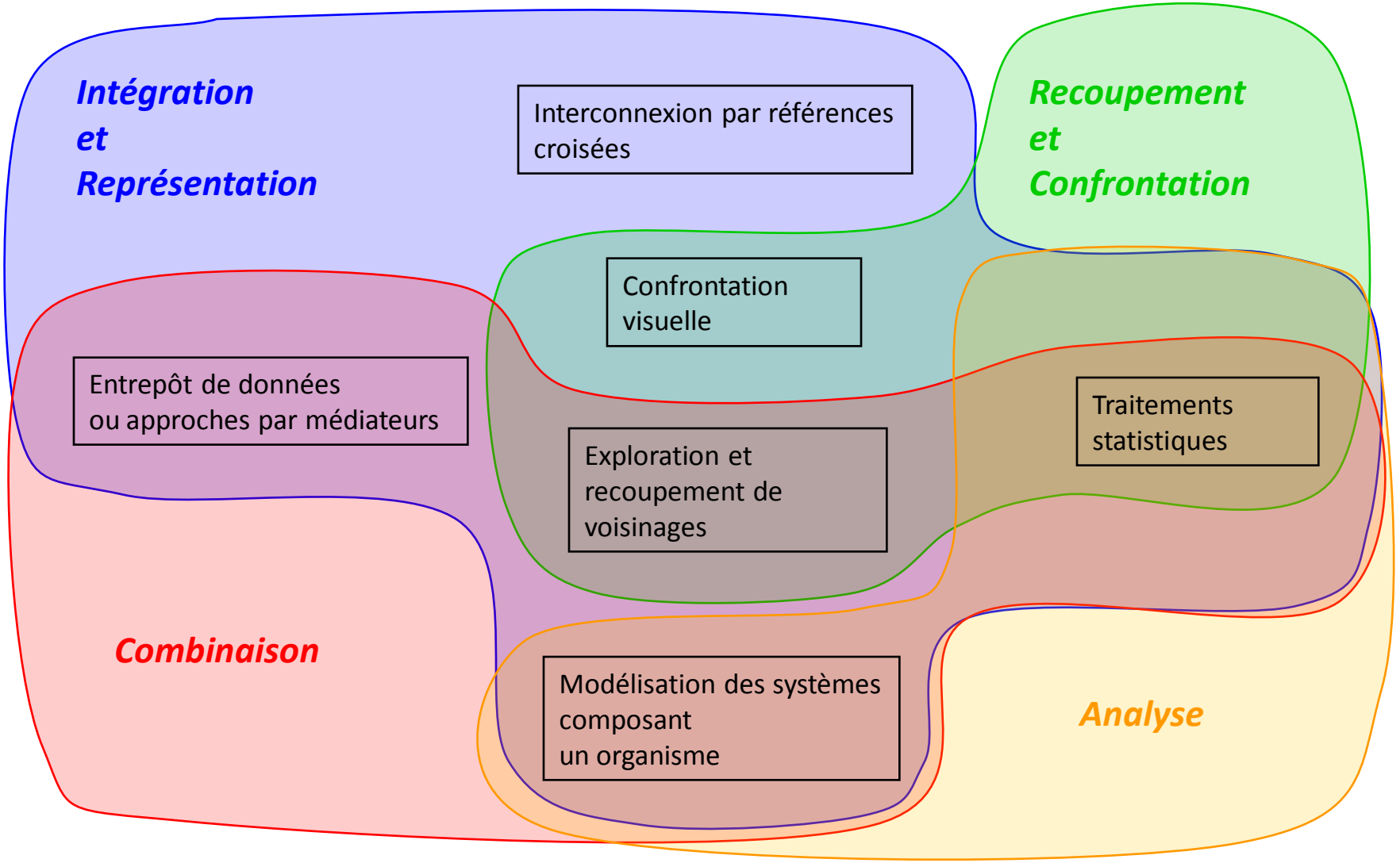


Cellule eucaryote

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

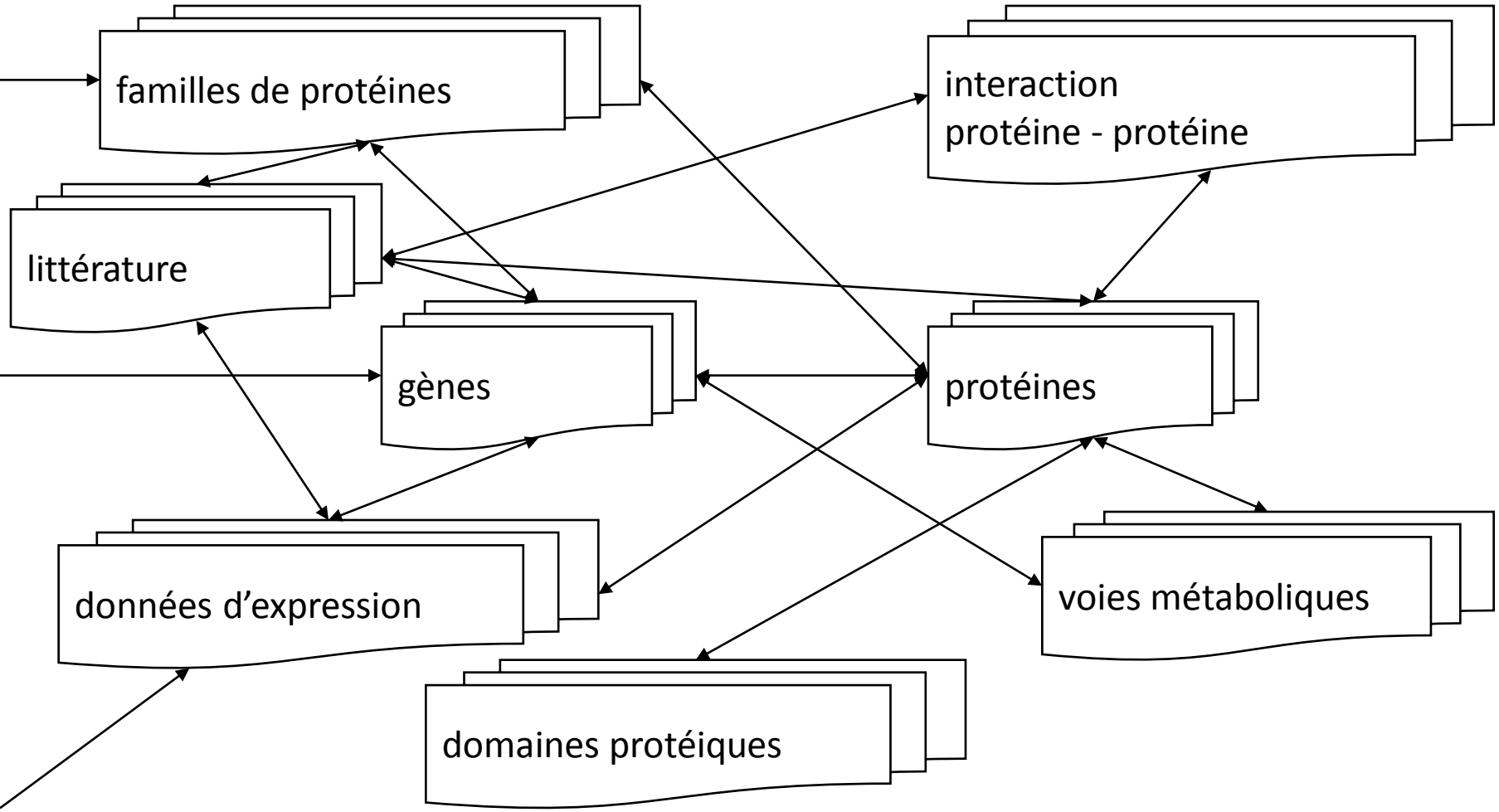
Vue générale



Intégration : vers une vue unifiée

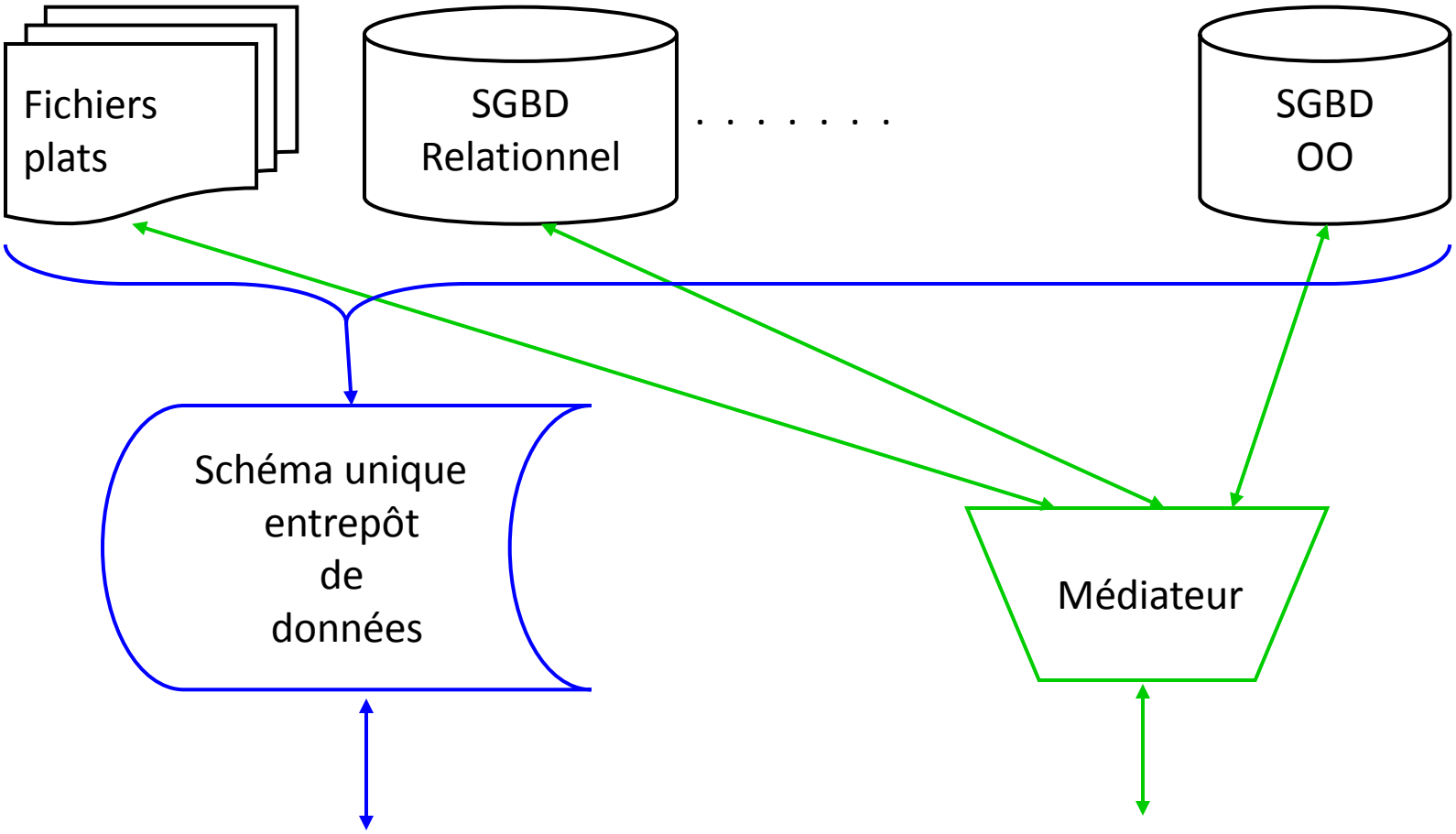
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

Intégration par interconnexion : principe



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

Intégration par fusion ou par médiateurs



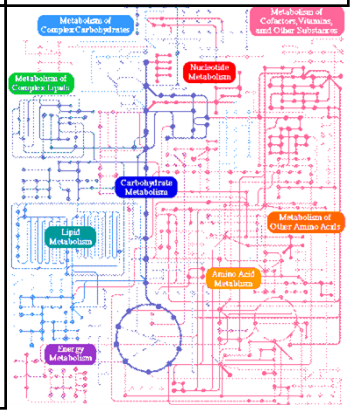
Integr8 [Kersey et al., 2005], BioMart [Kasprzyk et al., 2004],
WInGS [Abergel et al., 2004], BioKleisli [Davidson et al., 1997], …

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	

Modèle
- système de réaction
- interactions moléculaires
- pi calcul

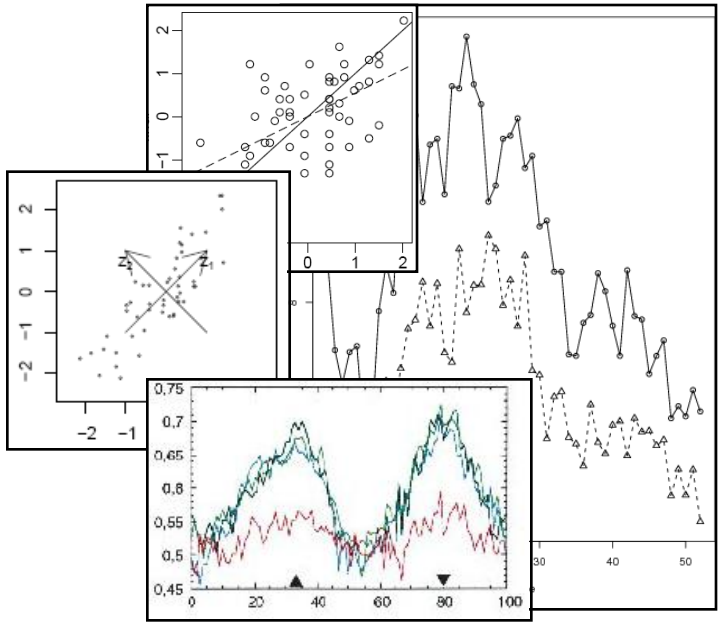
Caractérisation et simulations
- validation
- prédictions
- propriétés émergentes

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

Intégration : vers une vue unifiée

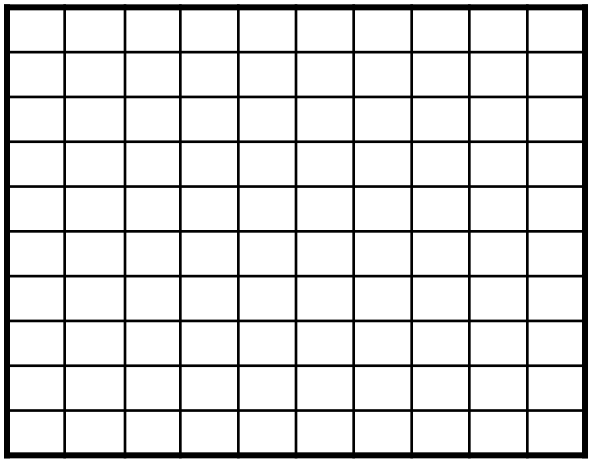
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

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



attributs

gènes

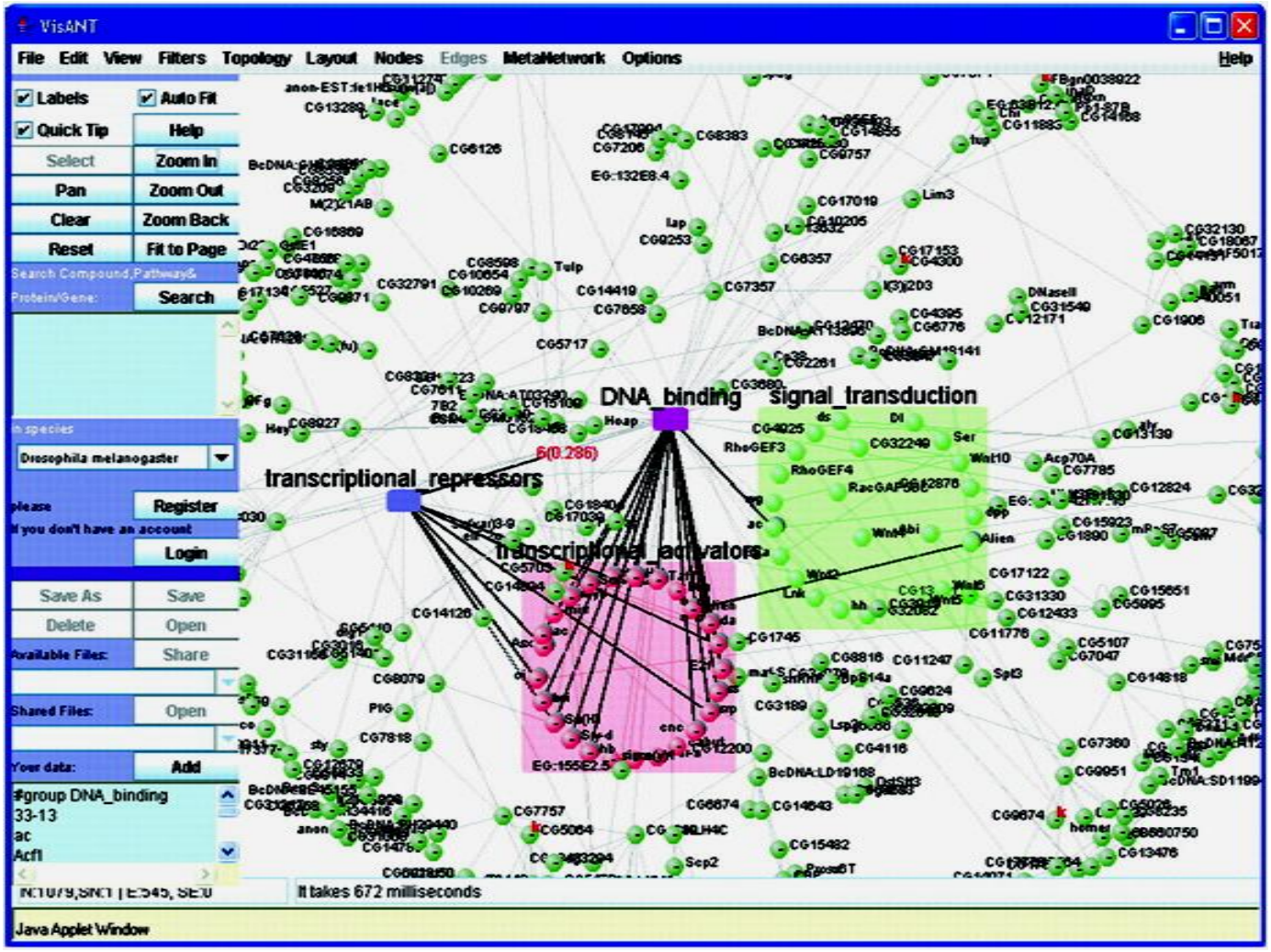


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	0	1	0	1	0	0	0	0	1	0	1	0	0	0	1	0	0	1	0	0	1
2	0	1	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1
3	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	1
4	1	0	0	1	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1
5	1	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	1	0
6	1	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	1	0	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

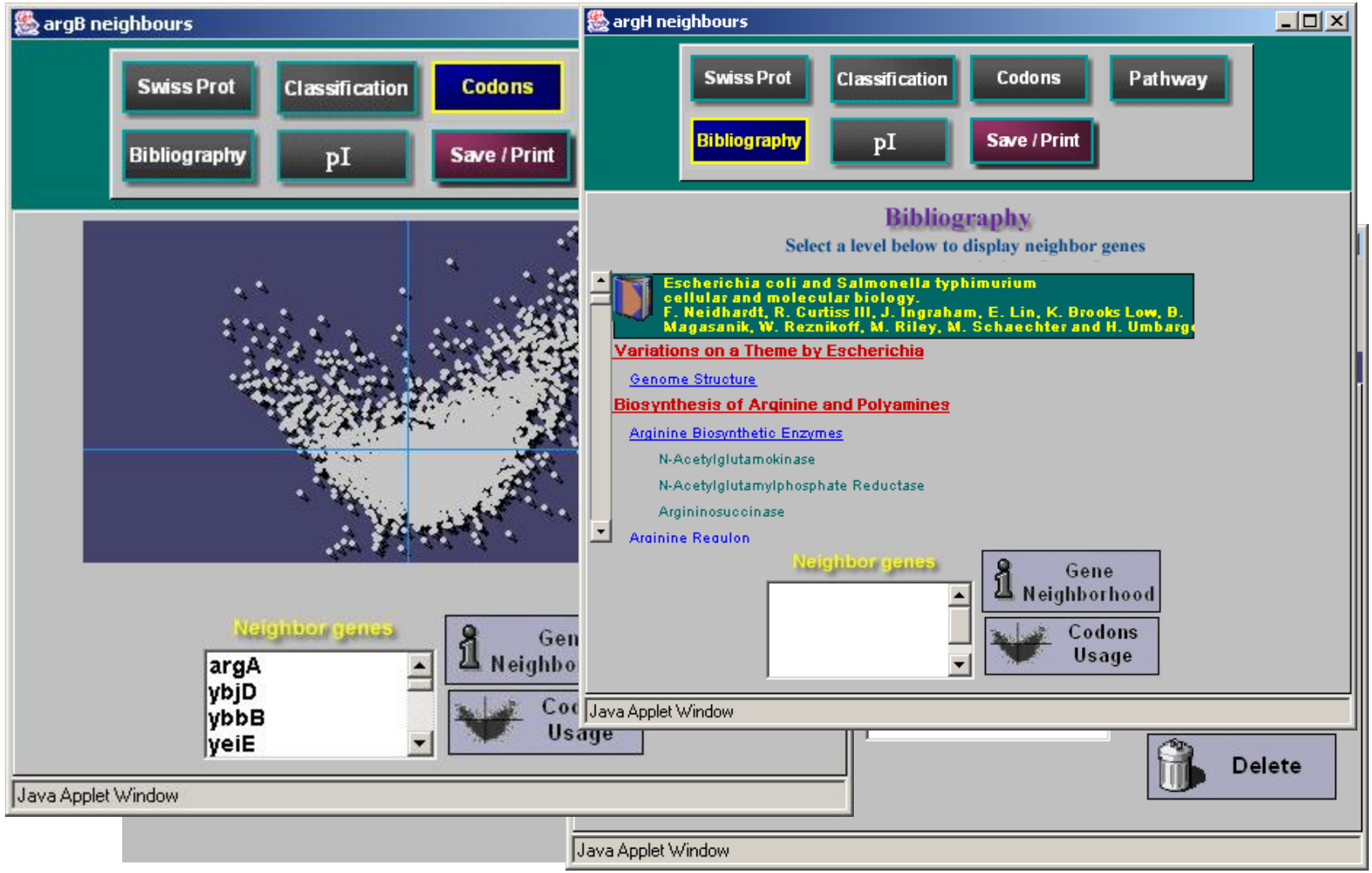


Visant [Hu et al., 2005]

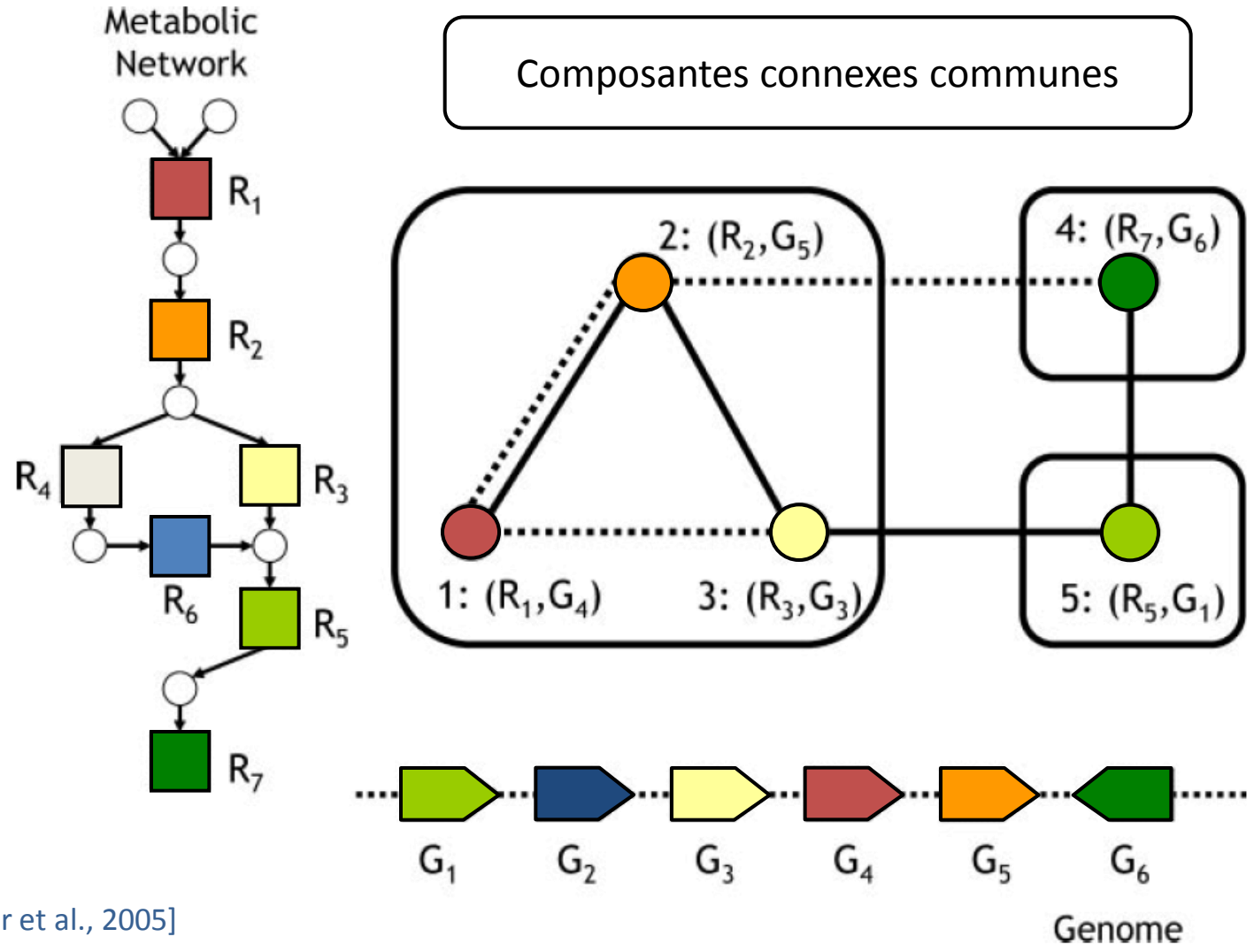
Intégration : vers une vue unifiée

- 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

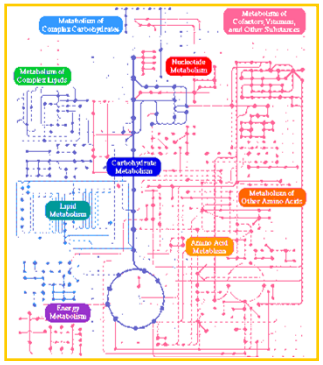


Recoupement de voisinages : approche graphique



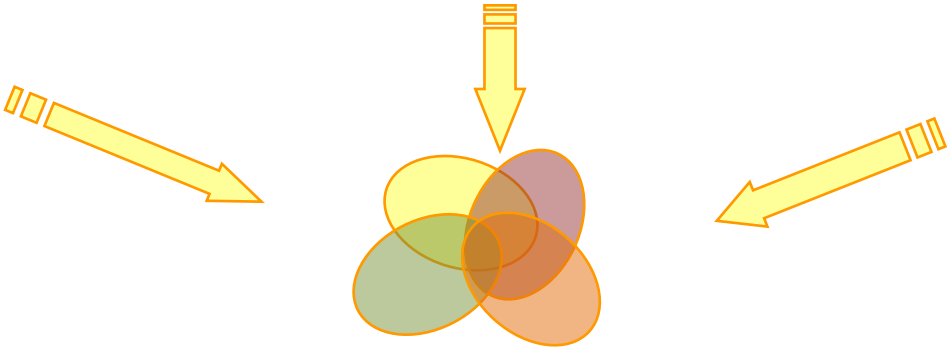
[Boyer et al., 2005]

Recoupement de voisinages : approche ensembliste

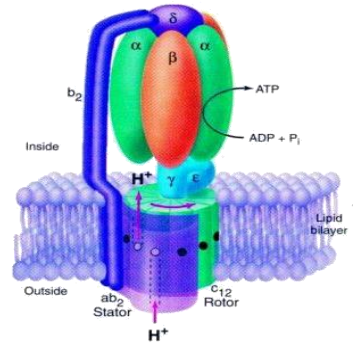


voies métaboliques

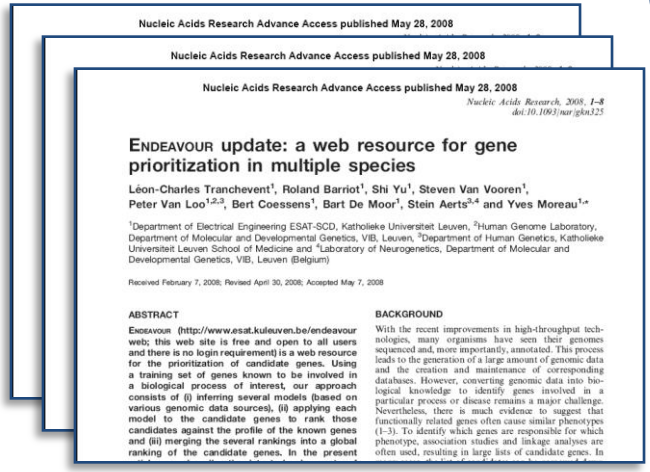
localisation chromosomique



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

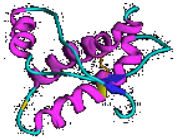
ABSTRACT

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

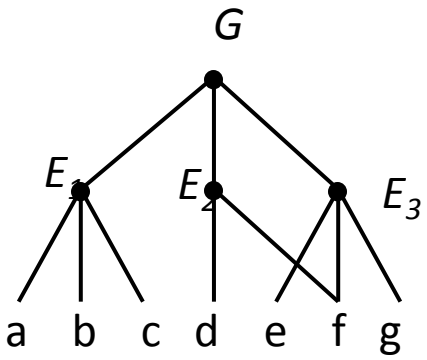
co-citation

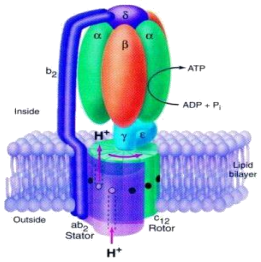


domaines protéiques

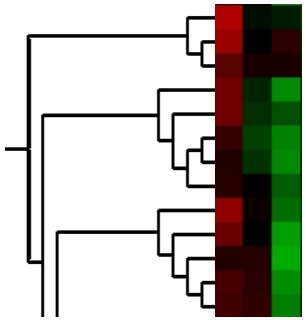
Gene Ontology

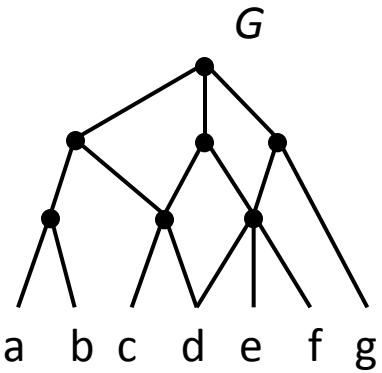
Ensembles



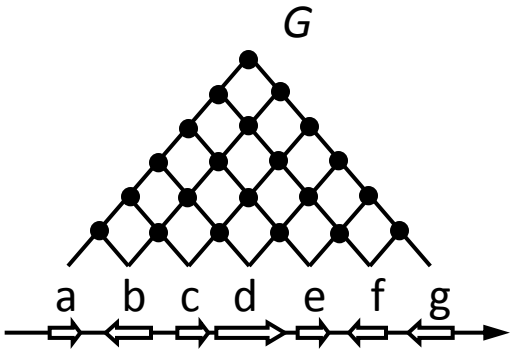


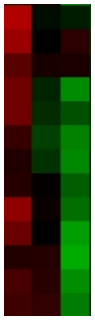
Hiérarchies



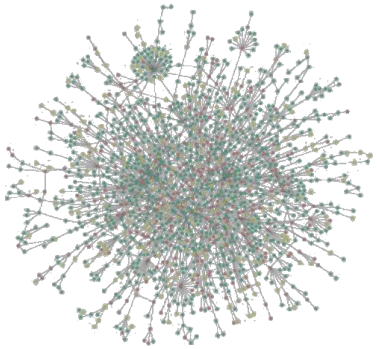


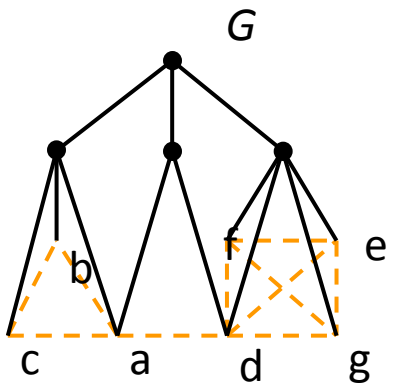
Vecteurs





Graphes

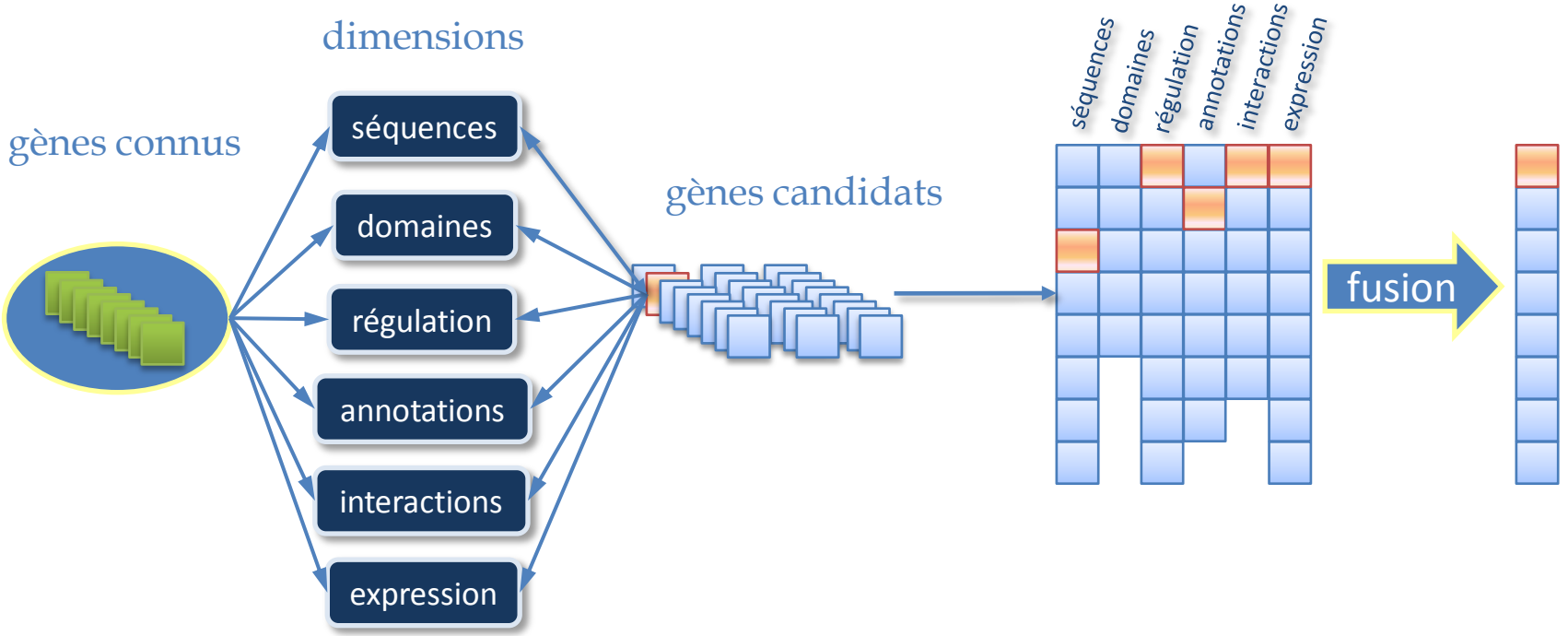




Intégration : vers une vue unifiée

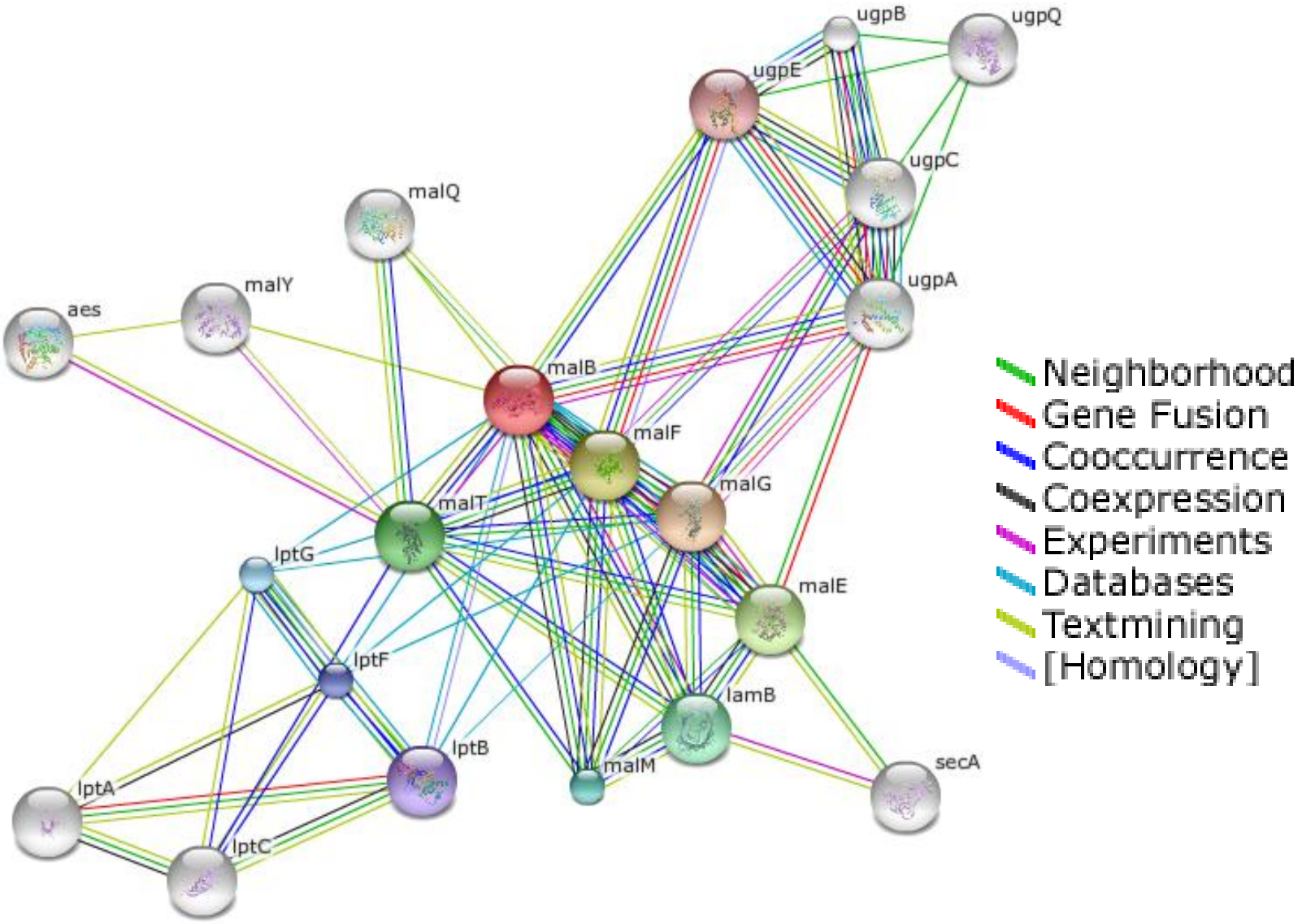
- Exploitation des références (croisées)
  - interconnexion
  - schéma unifié matérialisé : entrepôt
  - schéma unifié virutel : médiateur
- Modélisation
- Statistiques
- Confrontation visuelle, exploratoire
- Exploitation de la notion de voisinage
  - exploration
  - recoupement
  - confrontation
 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

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



[Aerts *et al.*, 2006]

Fusion : approche graphique



L'hétérogénéité sous toutes ses formes

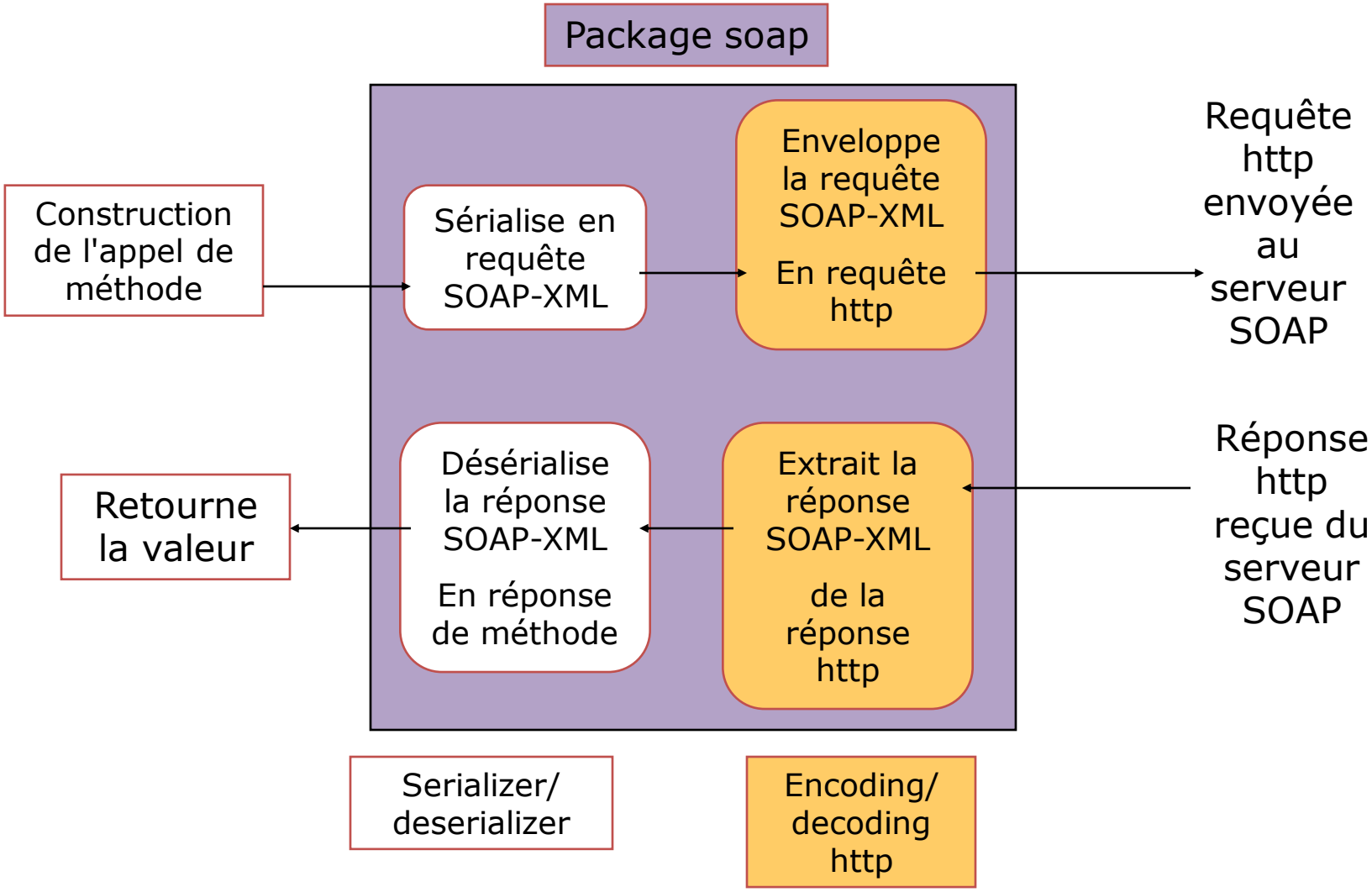
- Sémantique
  - Gene Ontology BP/KEGG pathways/BioCyc
  - structure d'un gène/peptide
- Modèle et format
  - SGBDr, SGBDoo, LDAP, fichier
    - schémas, attibuts et unités
  - XML, FASTA, EMBL/GenBank, SWISSPROT
- Architecture et accès
  - SGBD, SOAP, REST, pipeline (galaxy, ergatis)

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

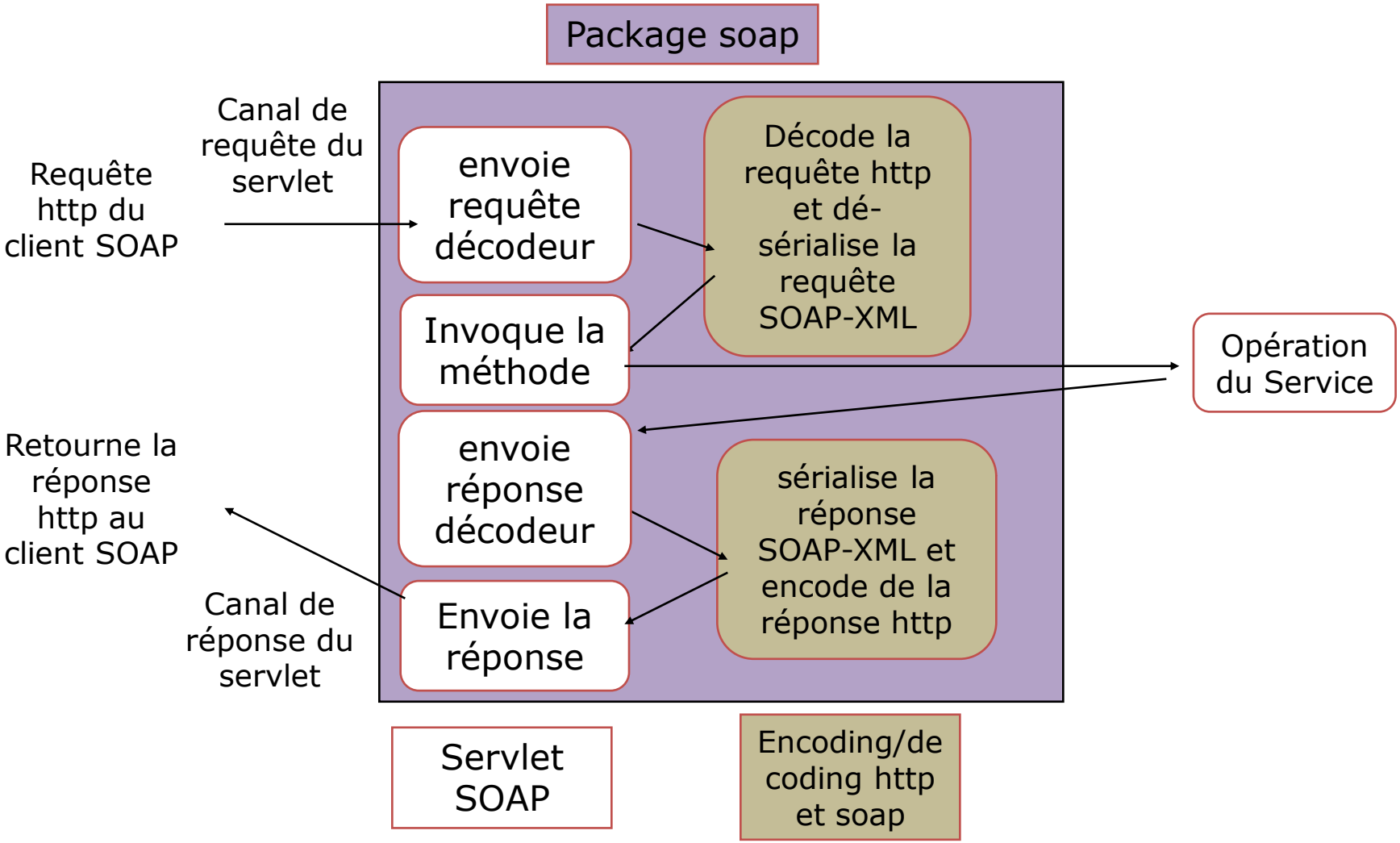
- Principe
  - appel de procédure à distance
  - permet de faire communiquer différents systèmes (plateforme, OS, langage de programmation, ...) à travers XML
  - un fournisseur propose certaines fonctionnalités
  - description du mode d'accès à ces fonctions : WSDL (Web Service Description Language)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

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

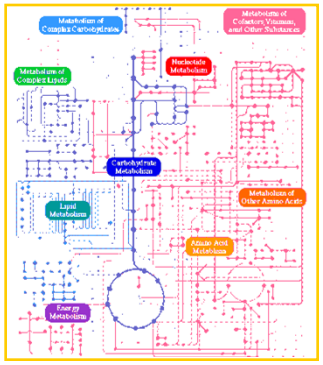
- Processus Client avec binding http



- Processus Serveur avec binding http

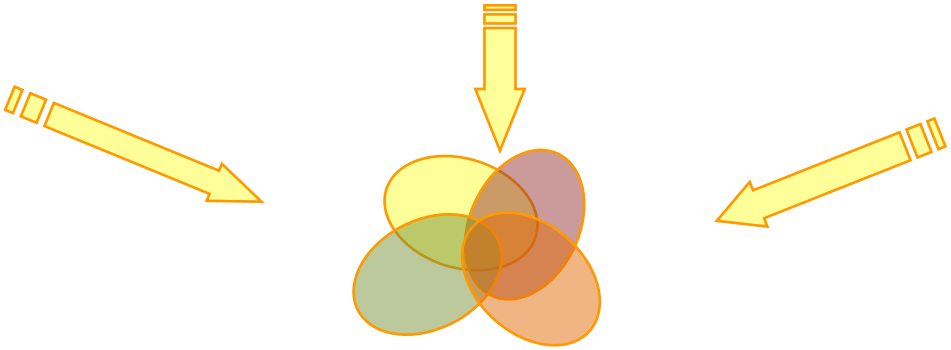


Recoupement de voisinages : approche ensembliste

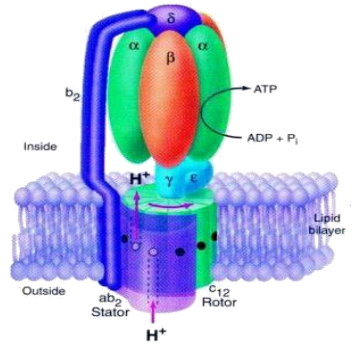


voies métaboliques

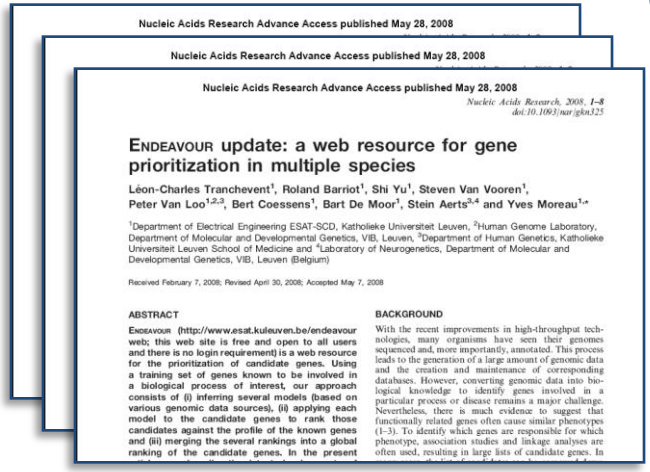
localisation chromosomique



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

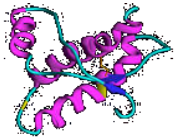
ABSTRACT

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

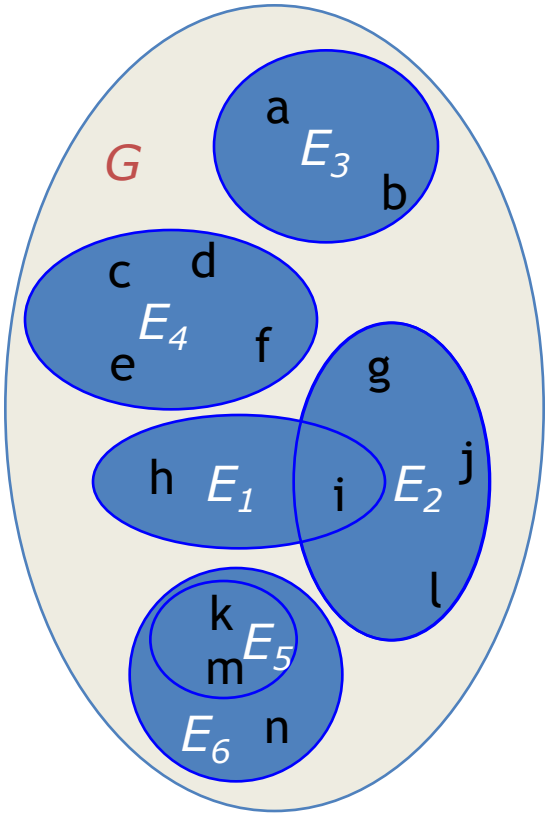


domaines protéiques

Gene Ontology

Définitions

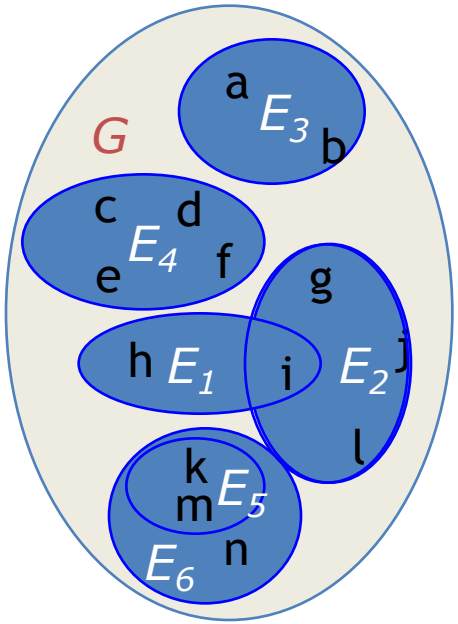
- (Identifiants de) gène → ARNm → protéine
- *G* : ensemble des gènes d'un organisme
- *Fonction de regroupement* : relation entre gènes basée sur un indice de similarité.
- *Ensemble de (gènes) voisins* : ensemble de gènes $E \subseteq G$ regroupés par une fonction de regroupement.
- *Voisinage* : sous-ensemble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



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

Représentation d'un voisinage : ordre partiel (poset)

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



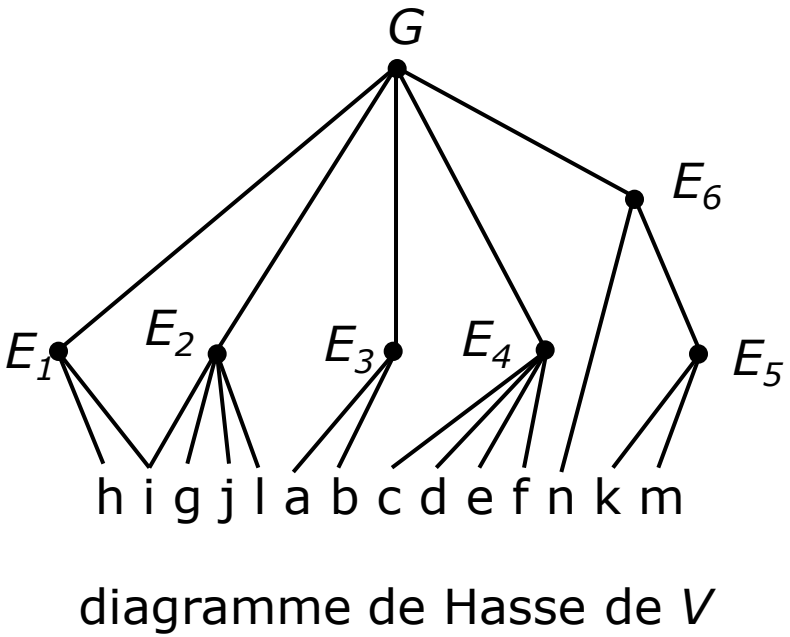
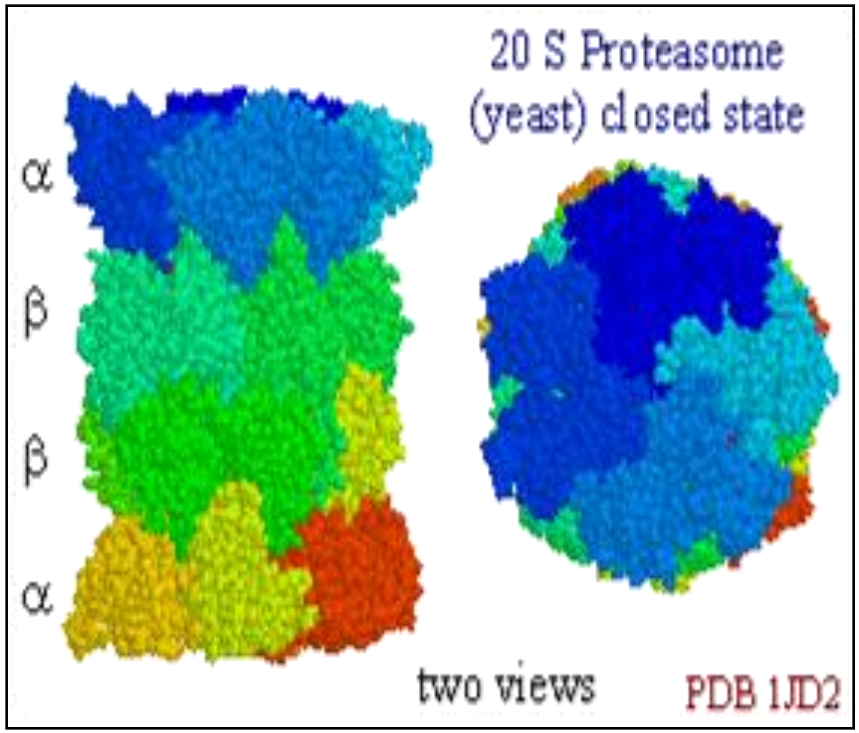
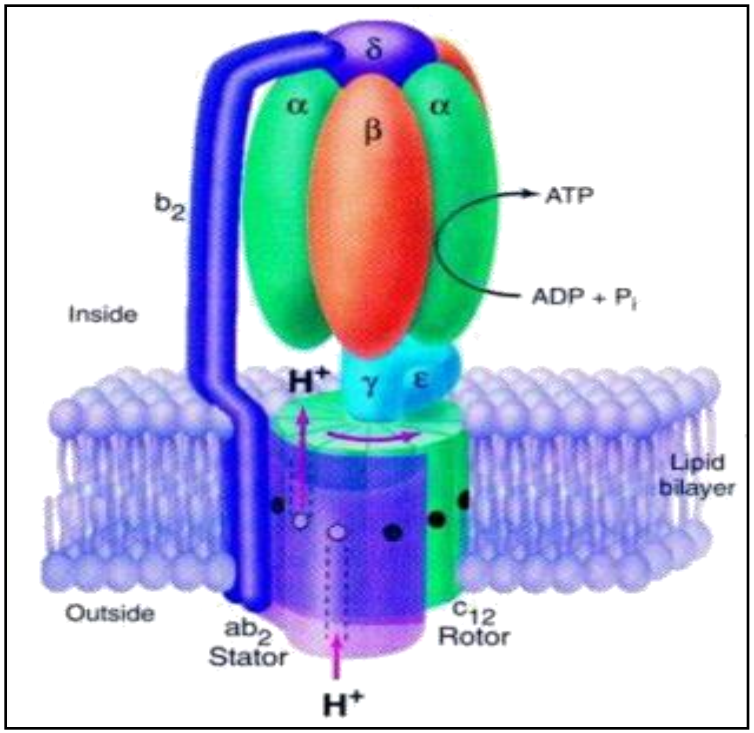


diagramme de Hasse de V

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

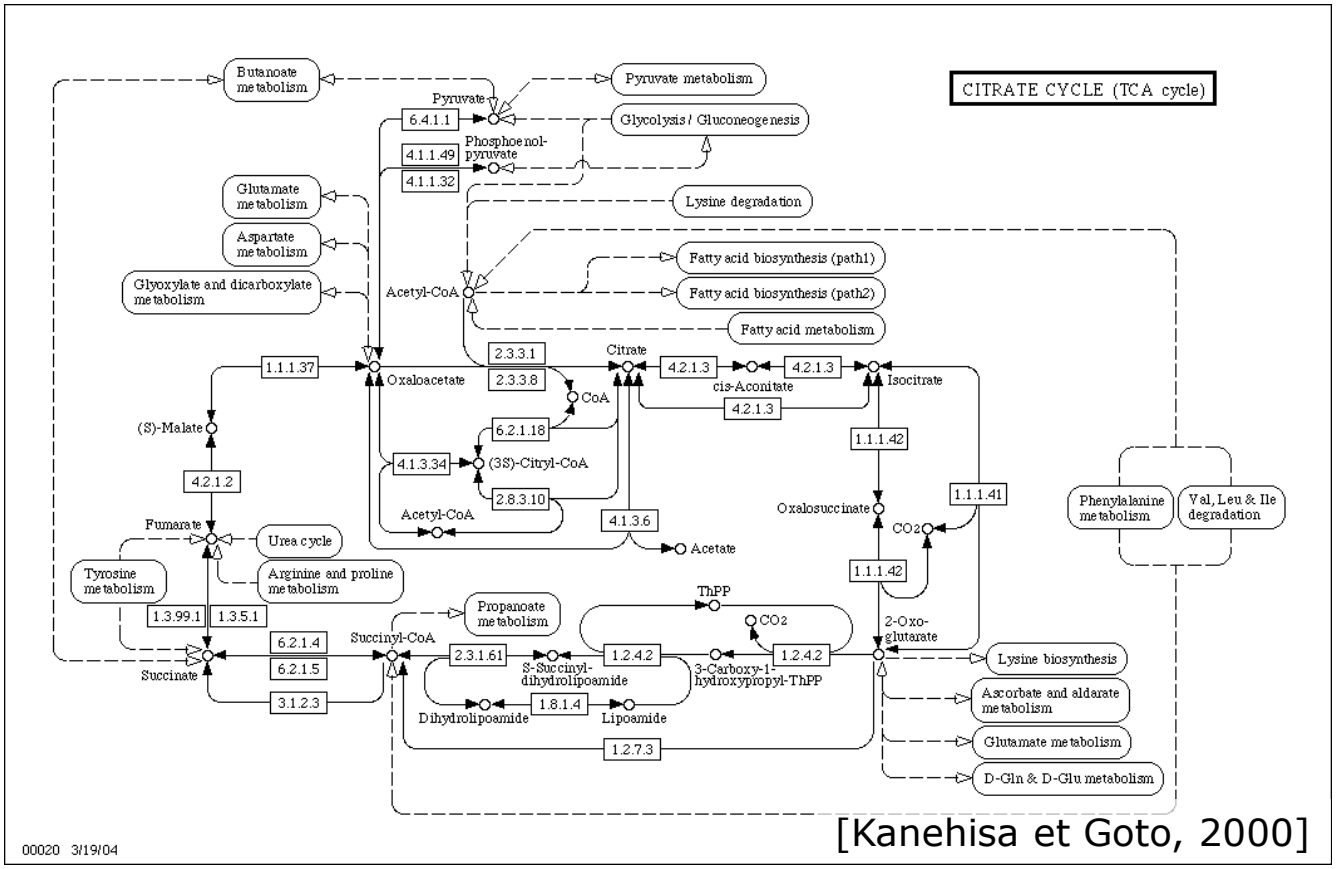
Exemple de fonction de regroupement : complexes protéiques





un complexe → un ensemble de protéines

Exemple de critère de regroupement : voies métaboliques

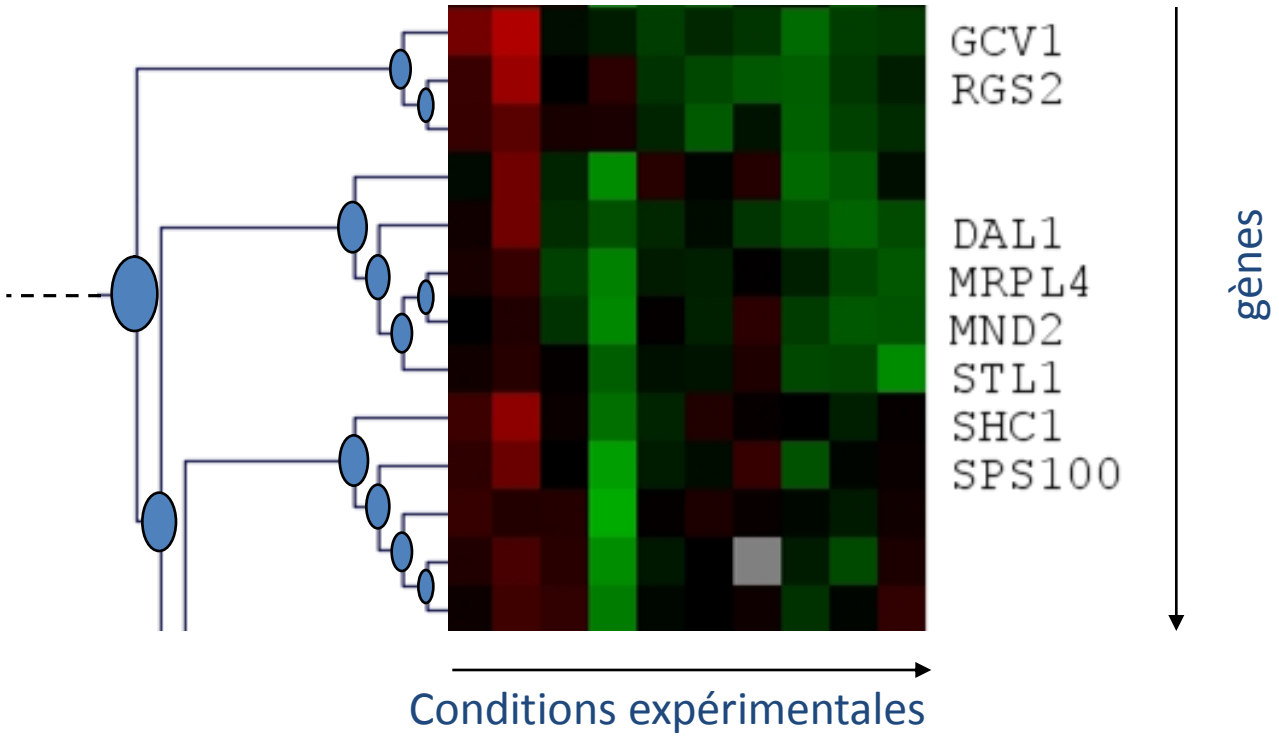


[Kanehisa et Goto, 2000]

une voie métabolique → un ensemble de protéines

Exemple de critère de regroupement : données d'expression

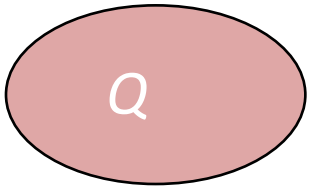
clustering hiérarchique
des profils



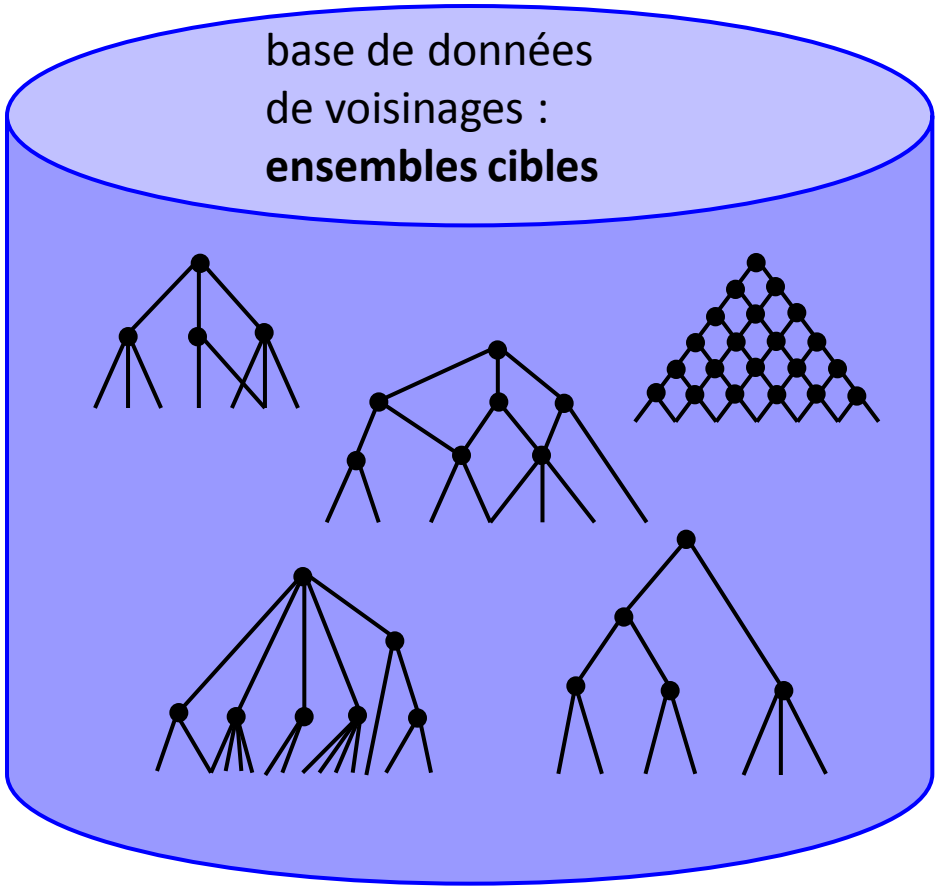
un cluster → un ensemble de gènes

- Recherche d'ensembles similaires

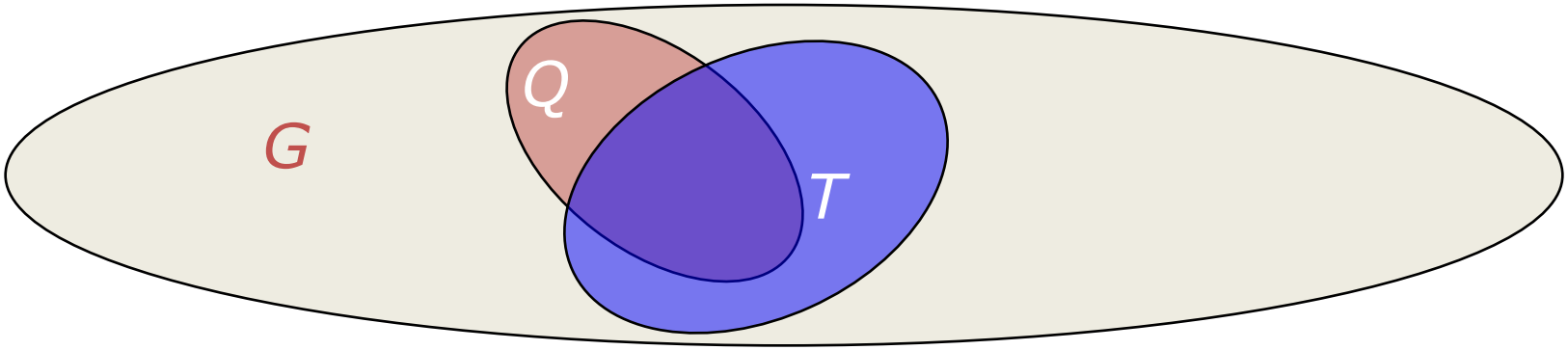
ensemble requête
$Q \subseteq G$



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

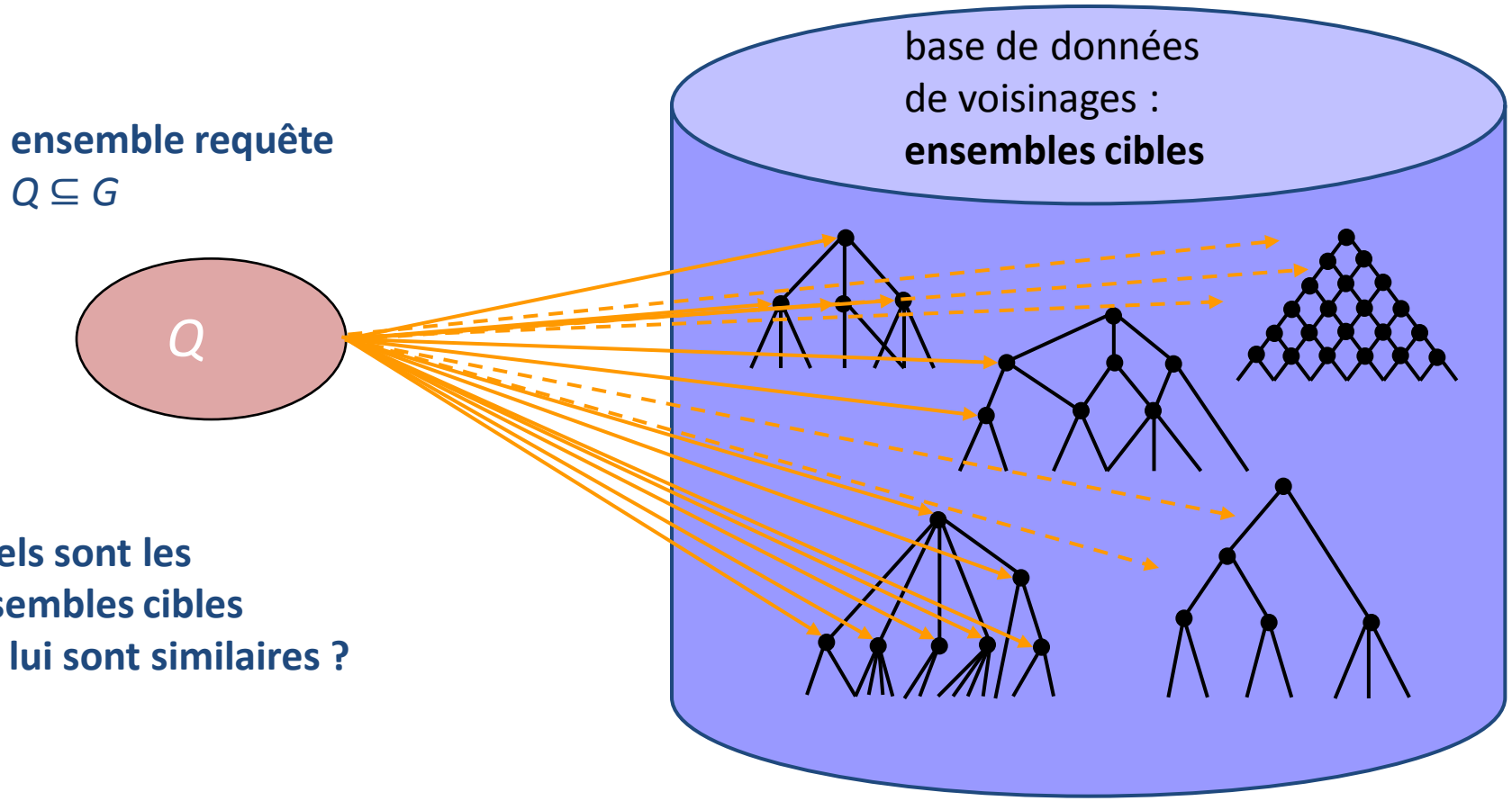
$$p - valeur\ (c,t,q,g) = \sum_{k=c}^{\min(q,t)} \frac{\binom{t}{k}\binom{g-t}{q-k}}{\binom{g}{q}}$$

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

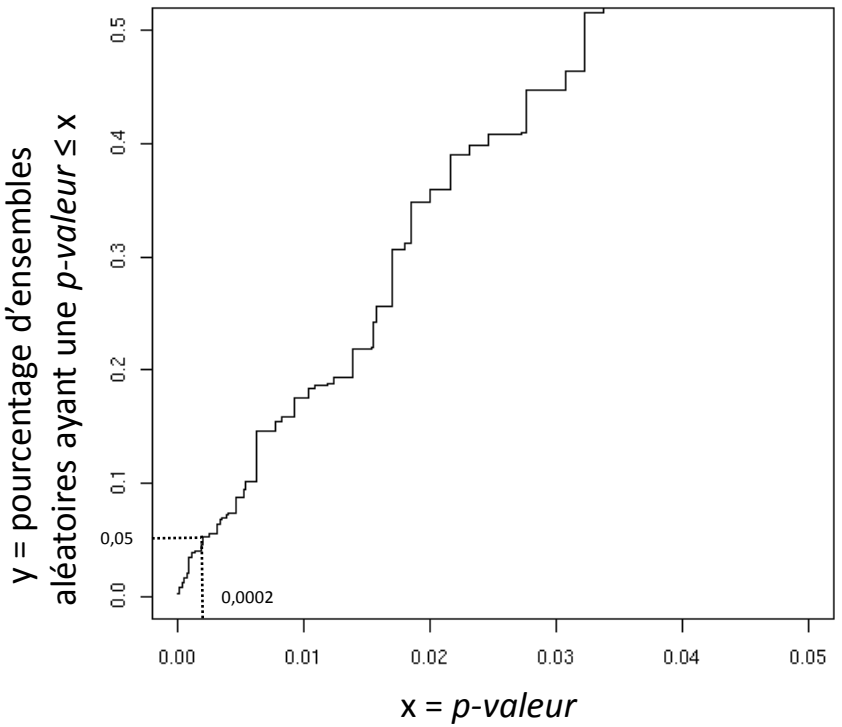


- Probabilité d'obtenir une p-valeur aussi faible par hasard : fonction de répartition des p-valeurs minimales

- Simulations

RandomSet_1, minPi = M1
RandomSet_2, minPi = M2
.
.
RandomSet_n, minPi = Mn

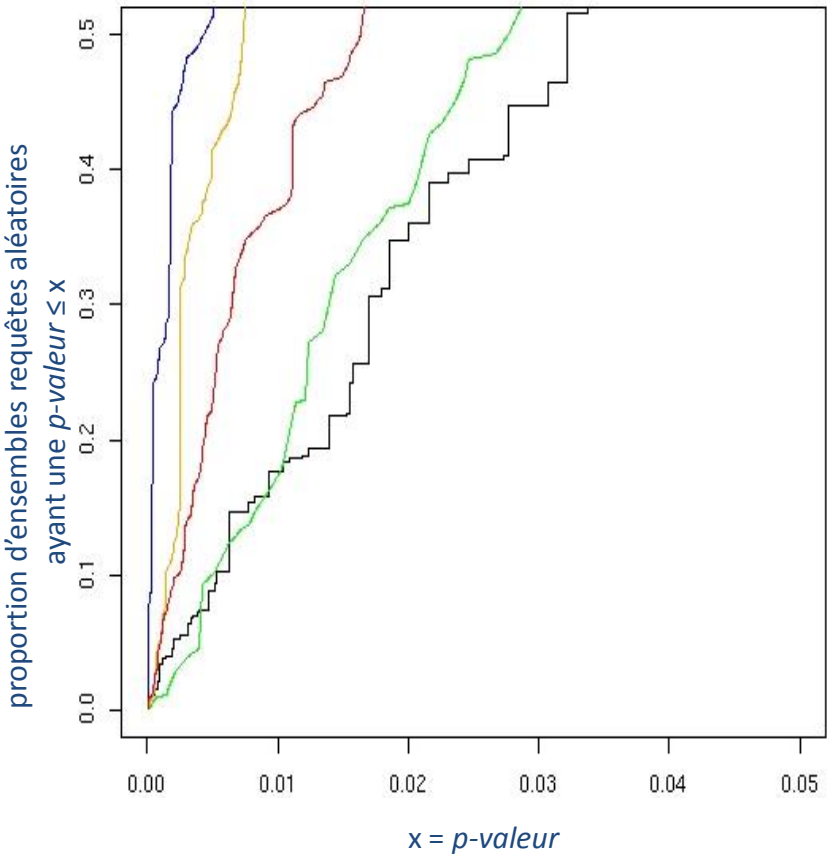
Étant donnée une p-valeur p
Combien ont un meilleur score ?



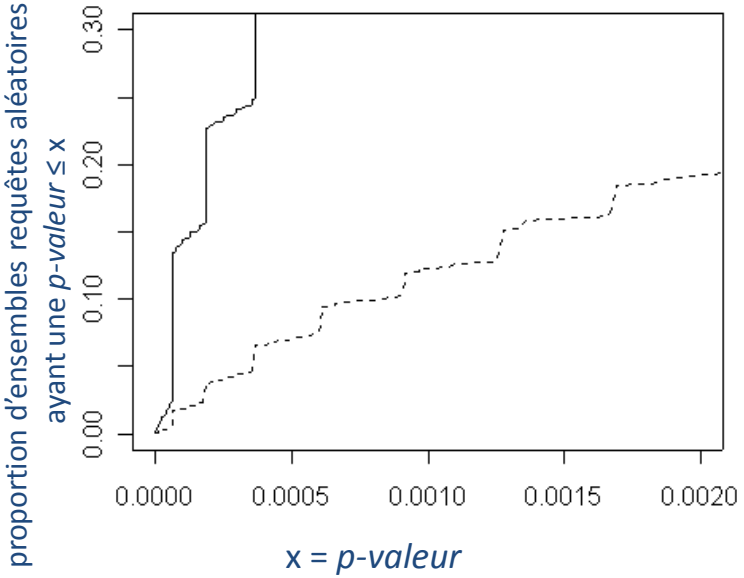
levure *Saccharomyces cerevisiae*
n=500, q=9, g=5786, KEGG Pathways

[Barriot *et al.* 2004]

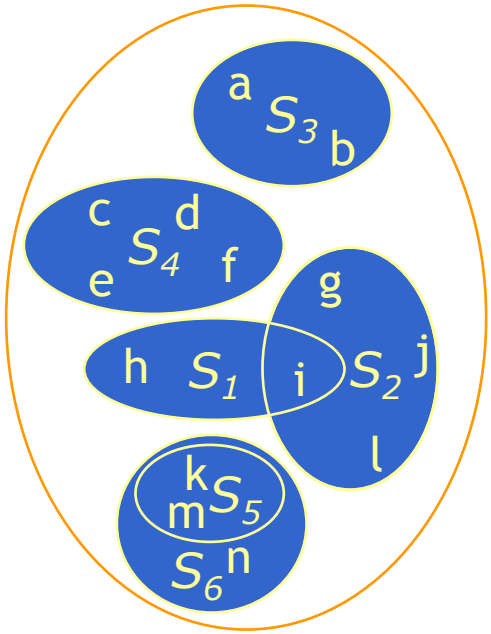
Significativité des p-valeurs obtenues

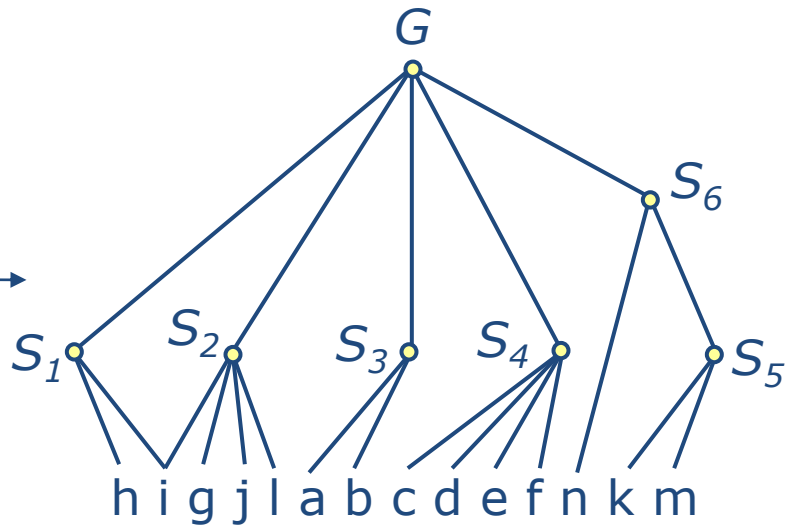


Saccharomyces cerevisiae
n=500, q=6-9-200-500-1000,
g=5786, KEGG Pathways



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





Hasse diagram of N

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

a target set T is **pertinent** if

$Q \cap T \neq \varnothing$

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

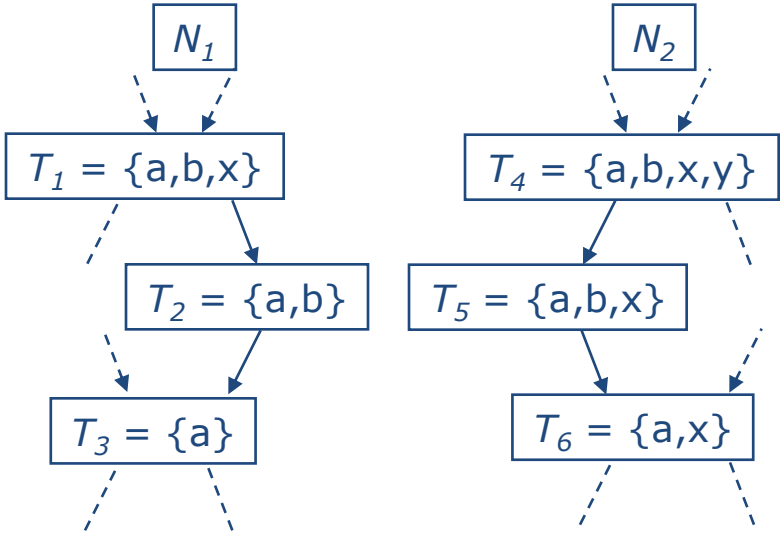
$$Q \cap T \neq \varnothing$$

and

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

and

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



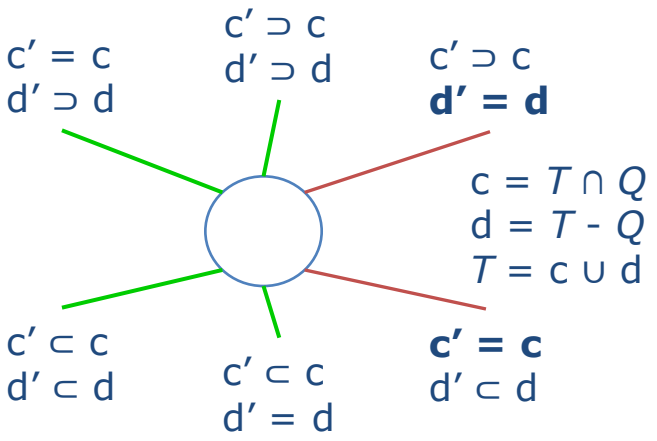
Q = {a,b,e}
Q' = {a,b}

Local decision

$|c| > 0$

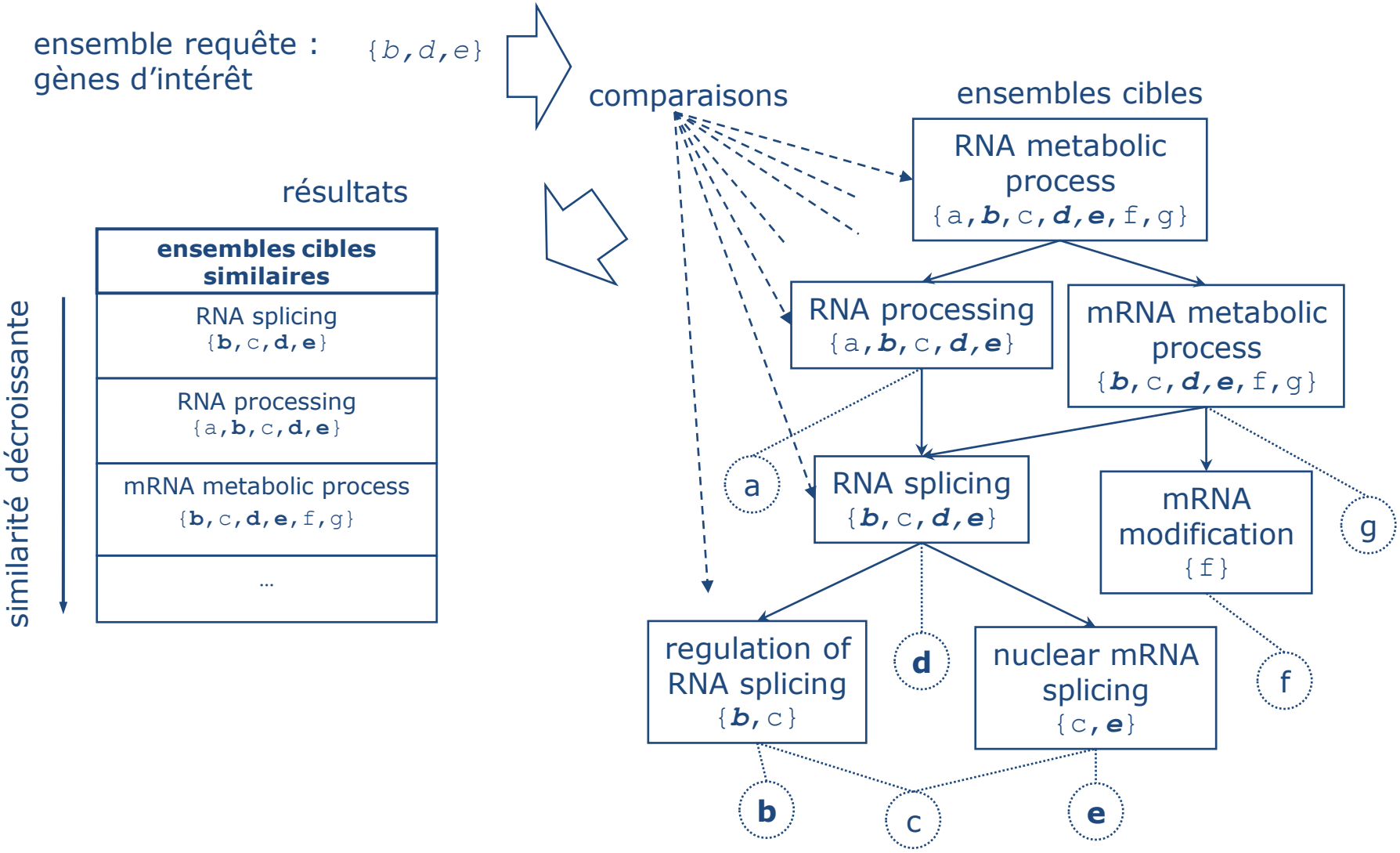
$|d| < \min(\{d_{parents}\})$

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

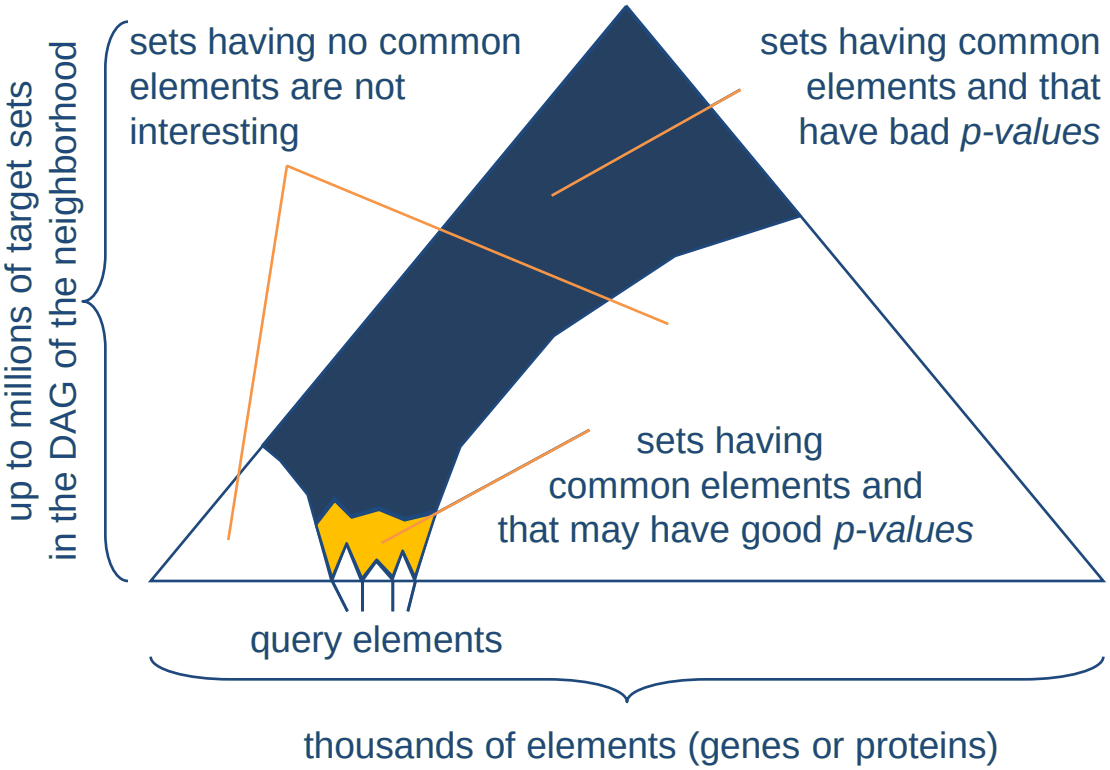


[Barriot, Dutour, Sherman, 2007, *BMC Bioinformatics*]

- Pertinence des comparaisons & redondance des résultats

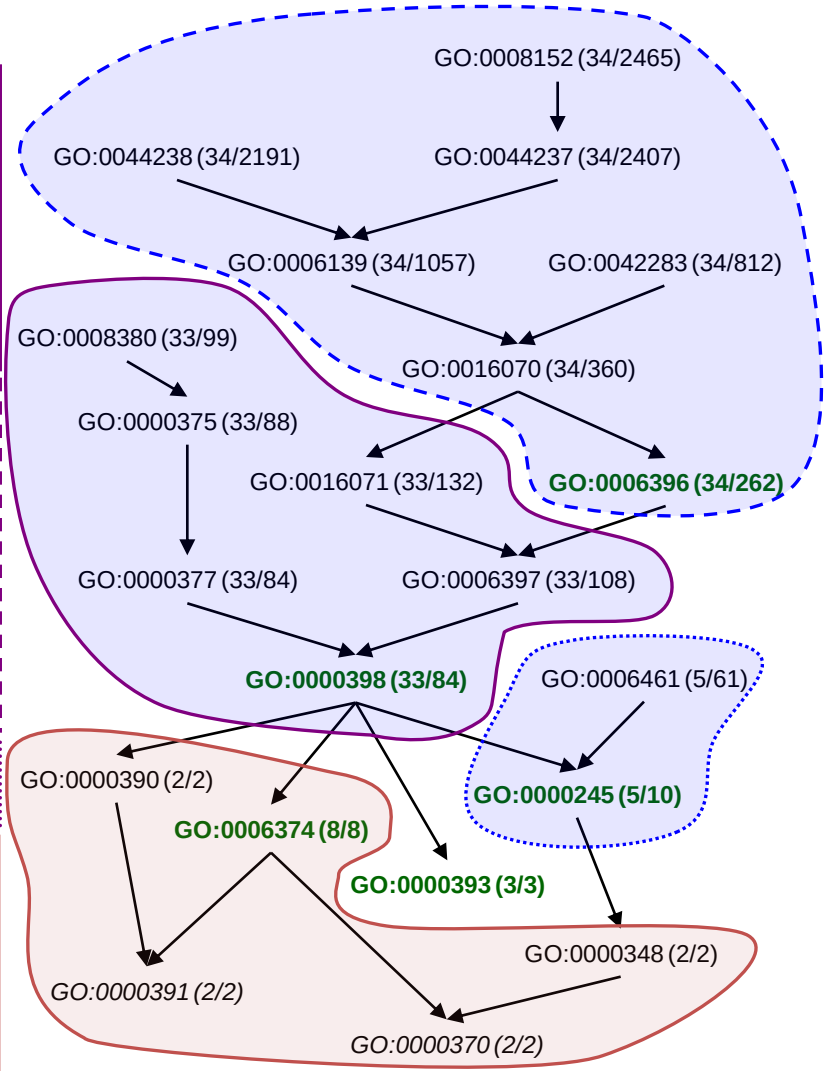


A small portion of the DAG is searched



Complex 440.30.10 mRNA splicing

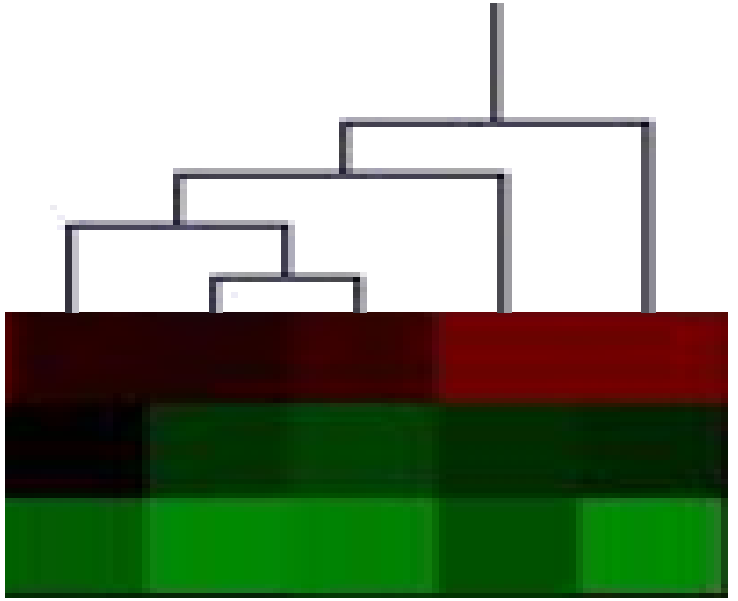
GO Term	Description	Target size	Common elements
GO:0000398	**nuclear mRNA splicing, via spliceosome**	**84**	**33**
GO:0000377	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	84	33
GO:0000375	RNA splicing, via transesterification reactions	88	33
GO:0008380	RNA splicing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



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

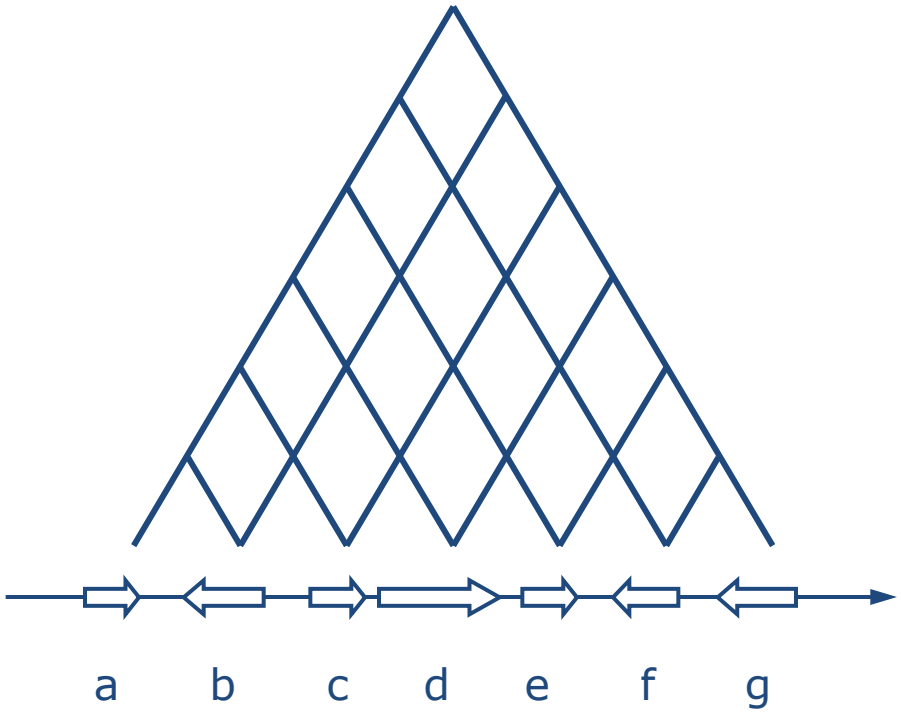


(((a (b c))d) e)

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

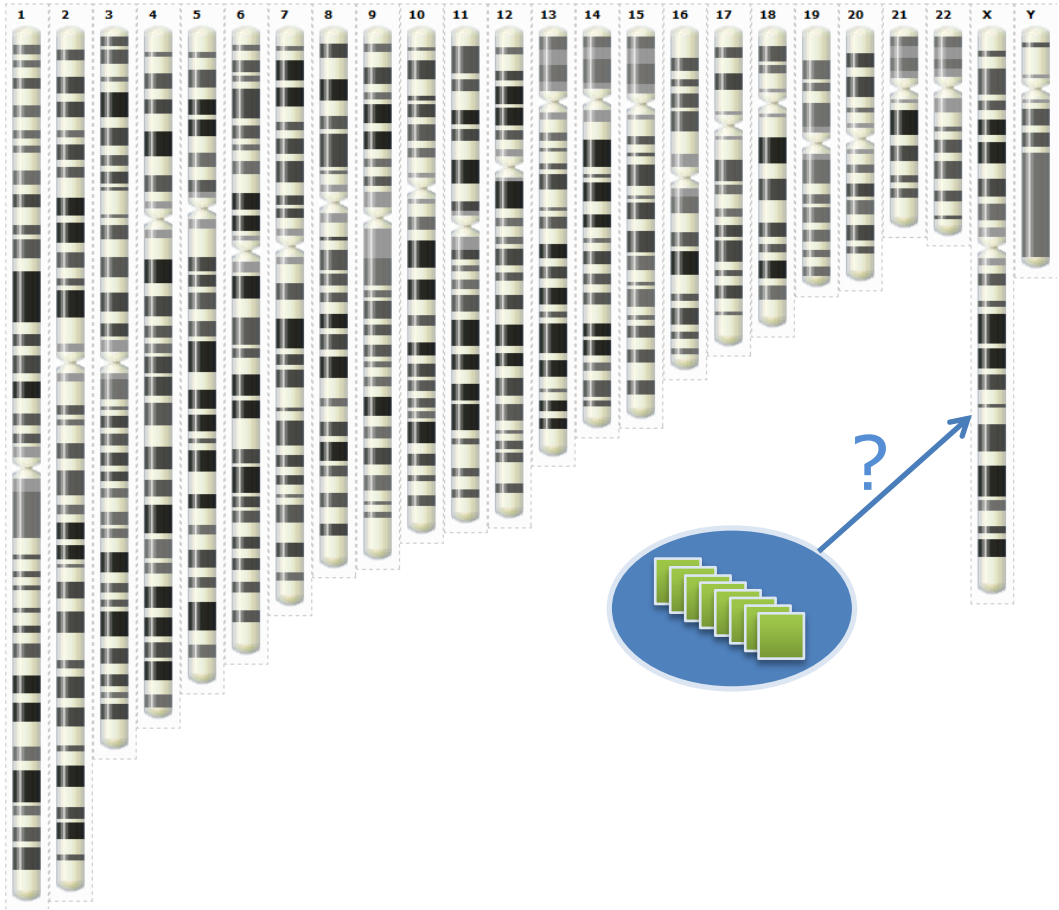


implicit

Set of genes of interest

Examples

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



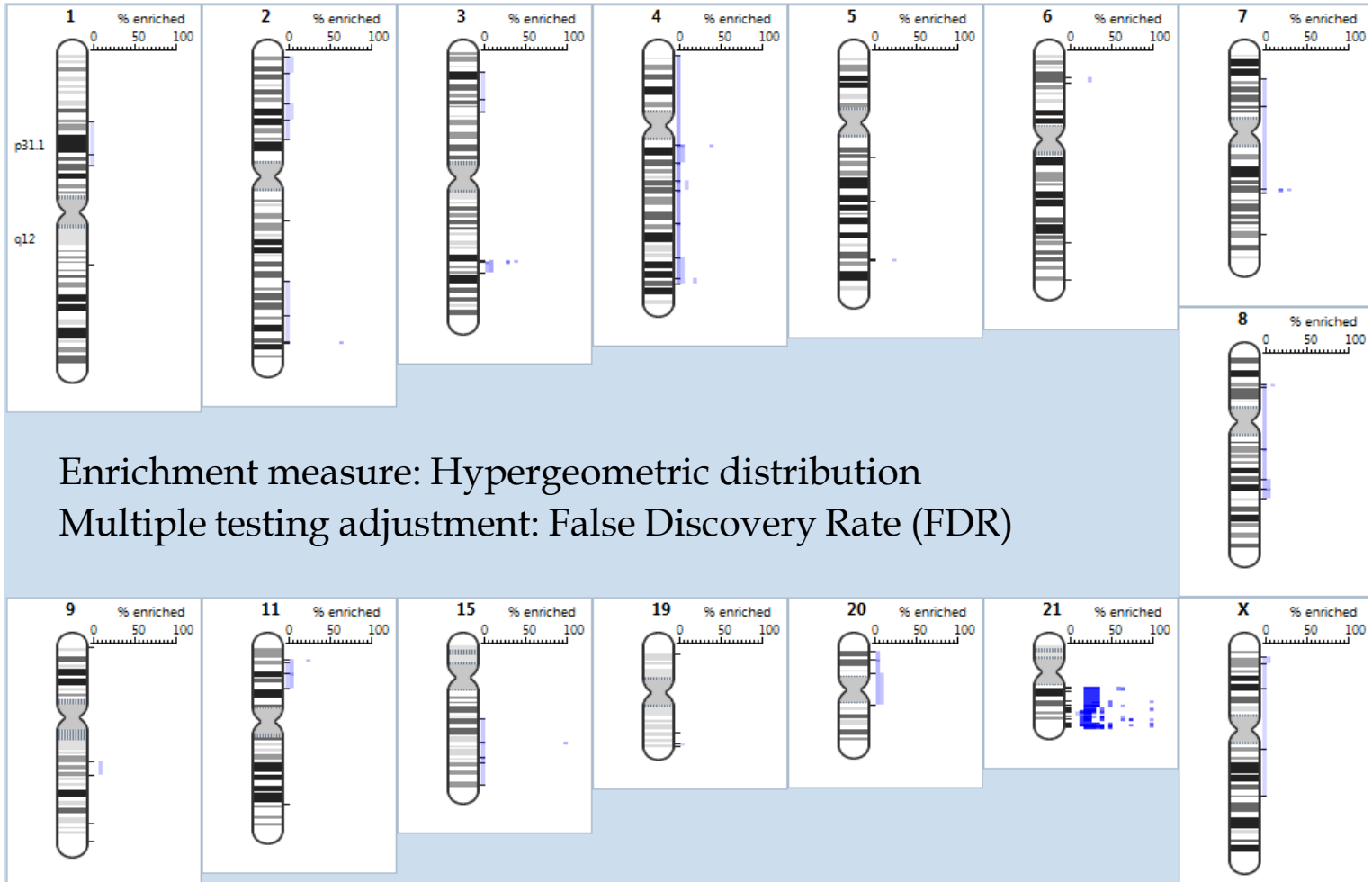
→ Question: Do those genes surprisingly cluster in the genome?

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)

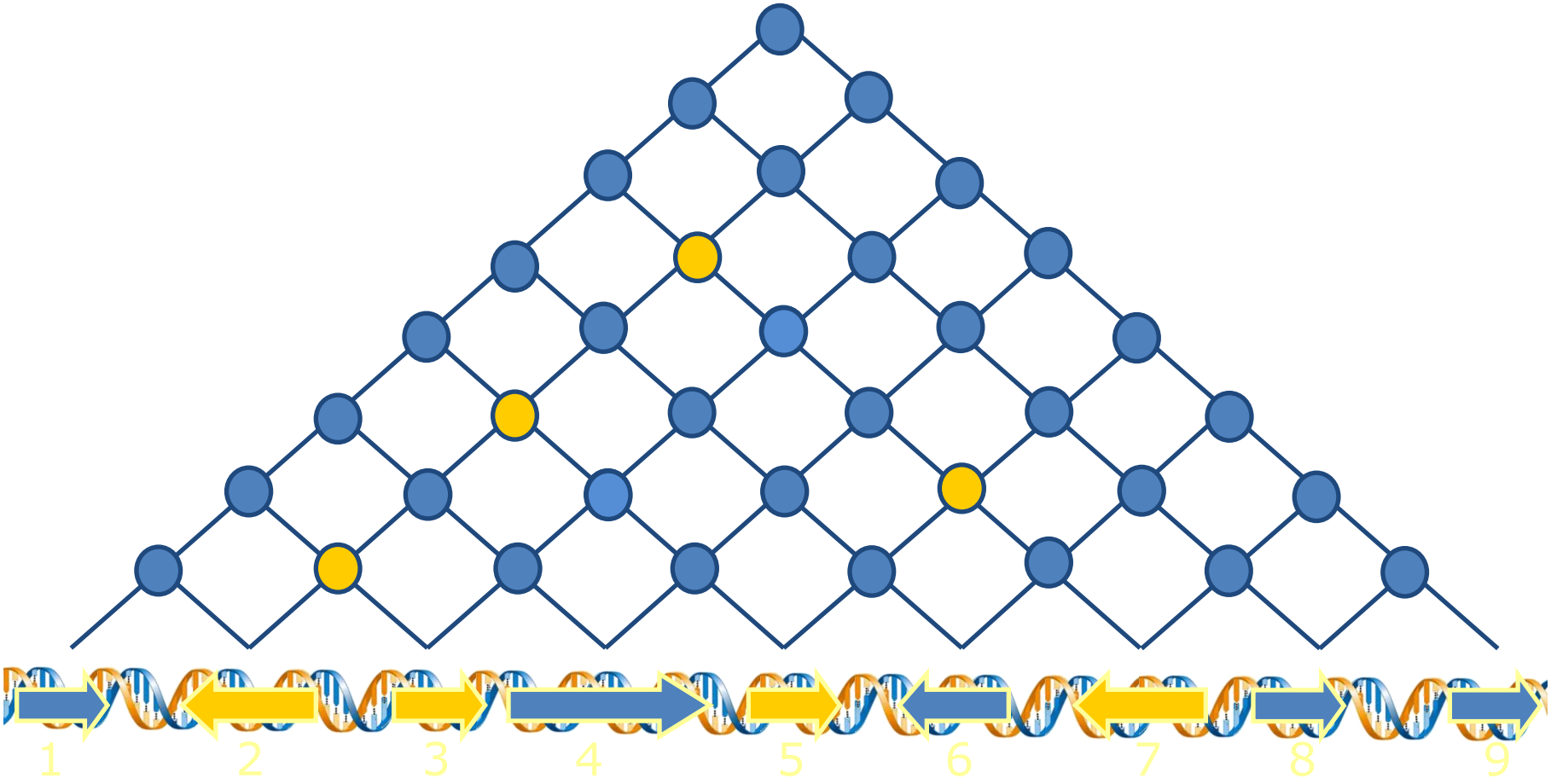
Issues:

- Number of regions to test
- False positives
- Redundancy

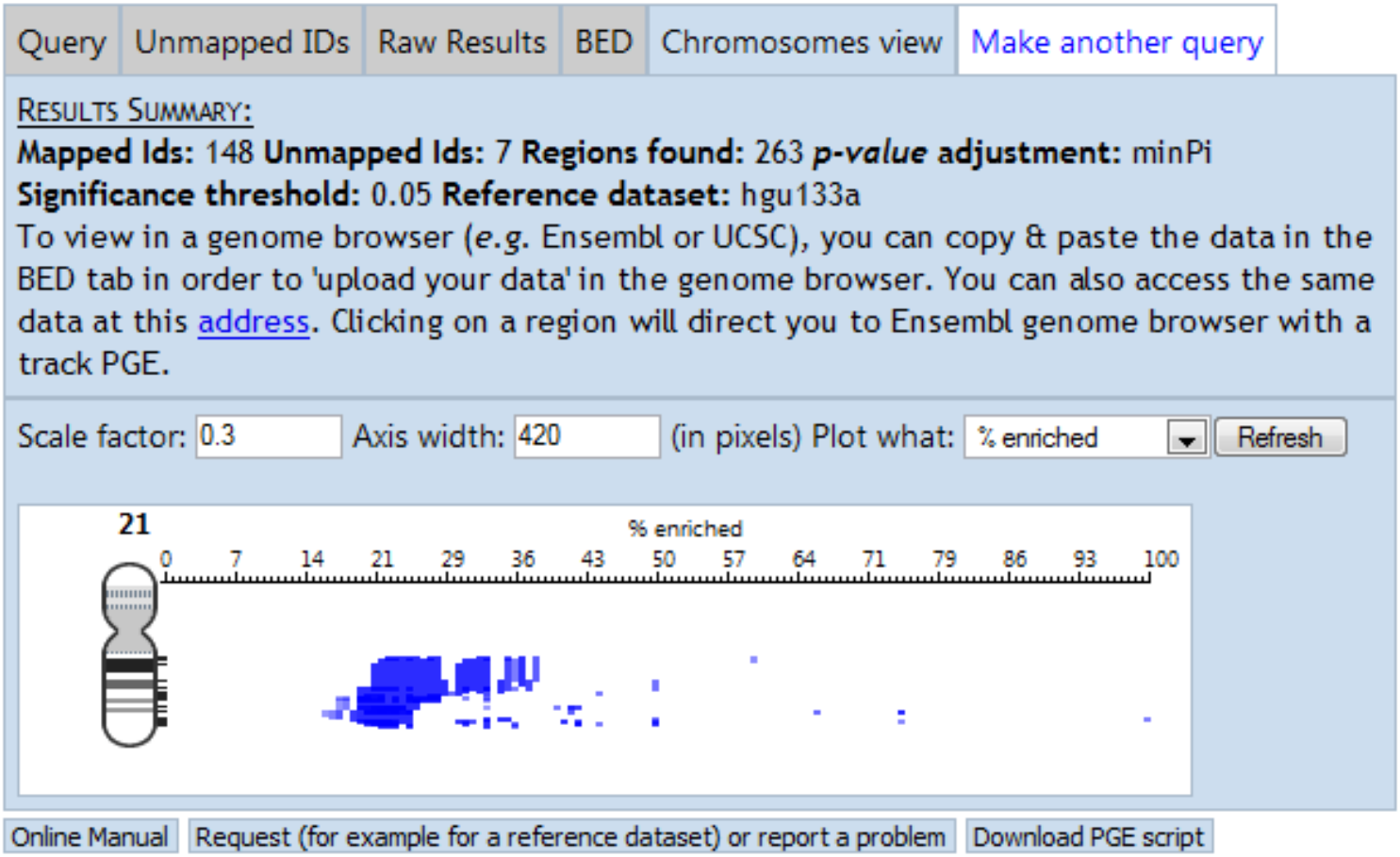
Pertinent Regions

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

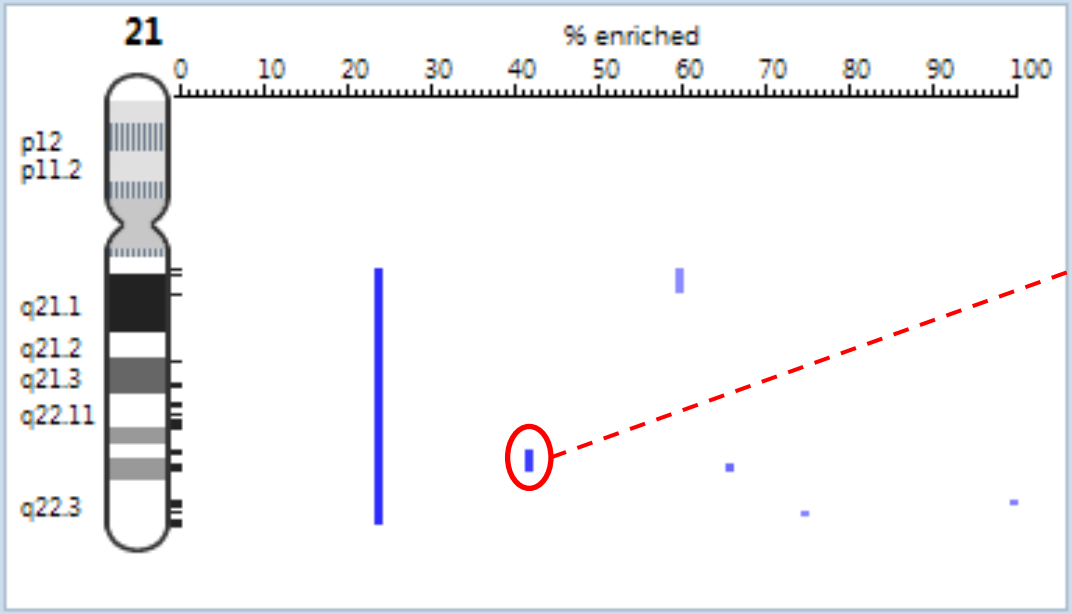
Query | Unmapped IDs | Raw Results | BED | Chromosomes view | Make another query

RESULTS SUMMARY:

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

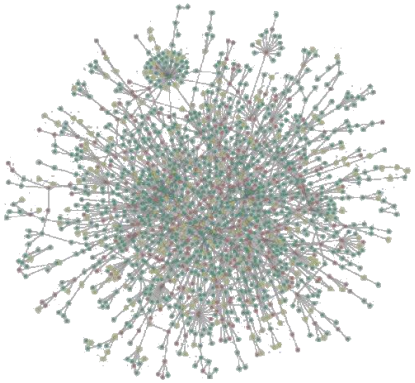
Scale factor: 1 Axis width: 300 (in pixels) Plot what: % enriched

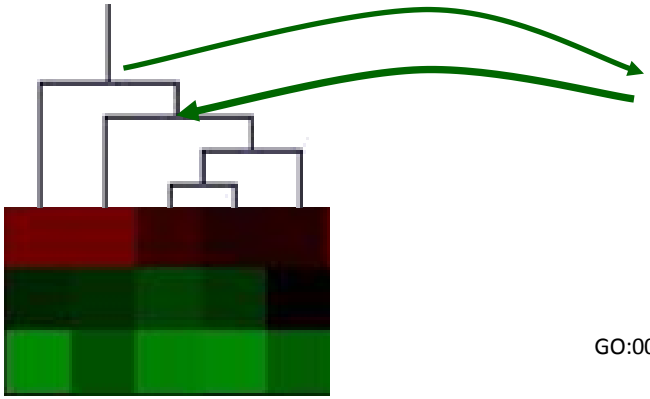
Refresh

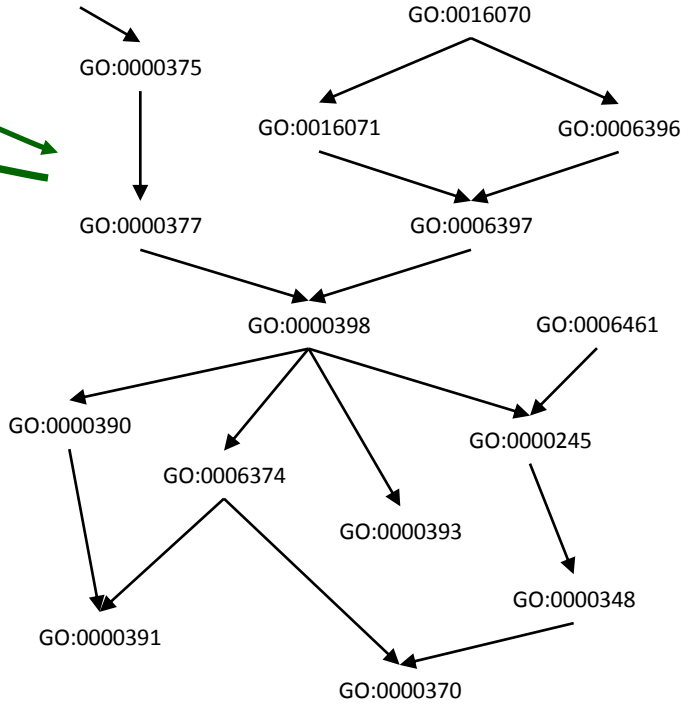


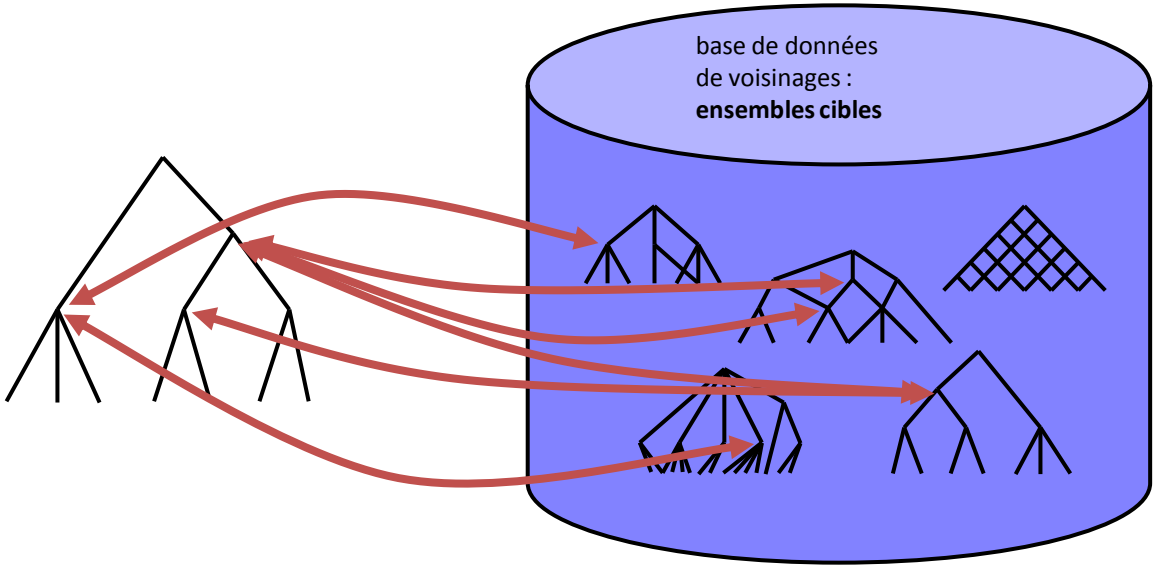
p-value:	7.87E-10
$p\text{-value}_{adj}$:	<0.002
Score:	91.039
$Score_{adj}$:	INF
Common elements:	6 / 14 (43%)
Overlaping regions:	1
Start of region:	37 359 546 bp
End of region:	40 223 183 bp
Genes:	
DSCR2, DYRK1A, PCP4, PIGP, SH3BGR	
WRB	

Challenges actuels

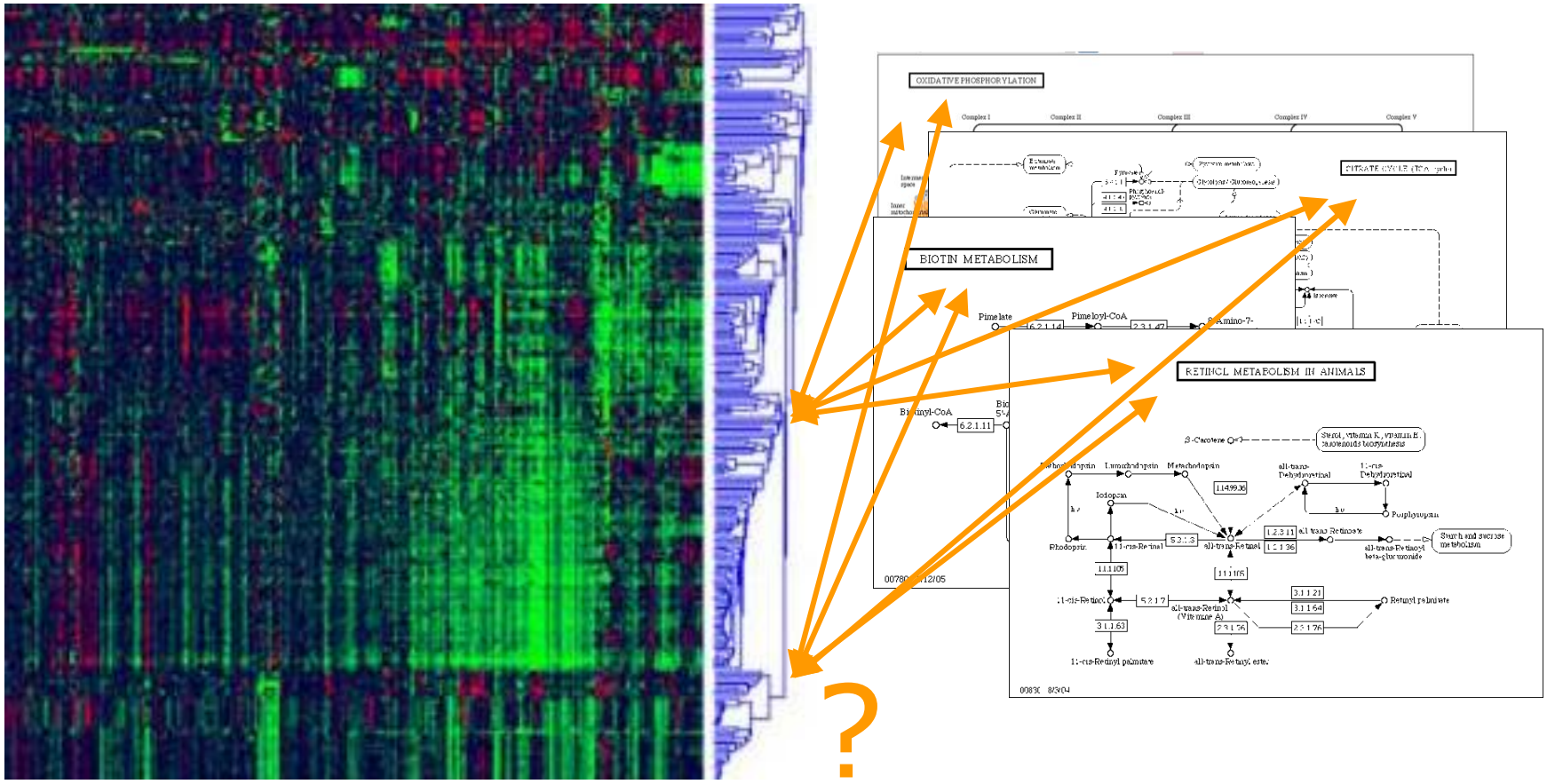








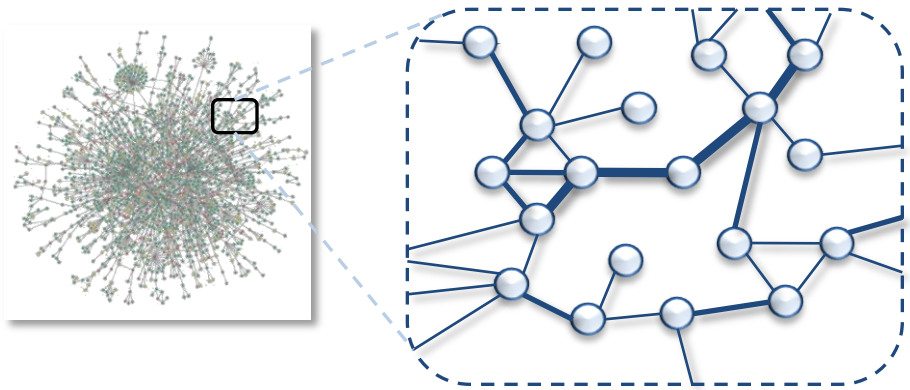
Analyse de données d'expression



[Ferea et al., 1999]

[Kanehisa et Goto, 2000]

Extraction de sous-graphe pertinent & visualisation



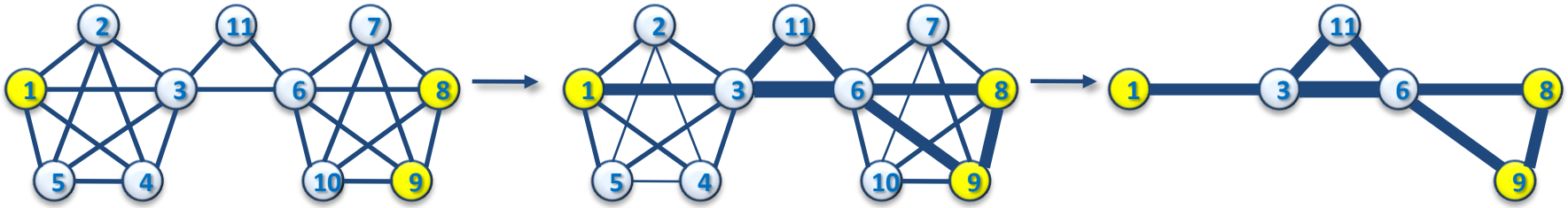
Idée :

- 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

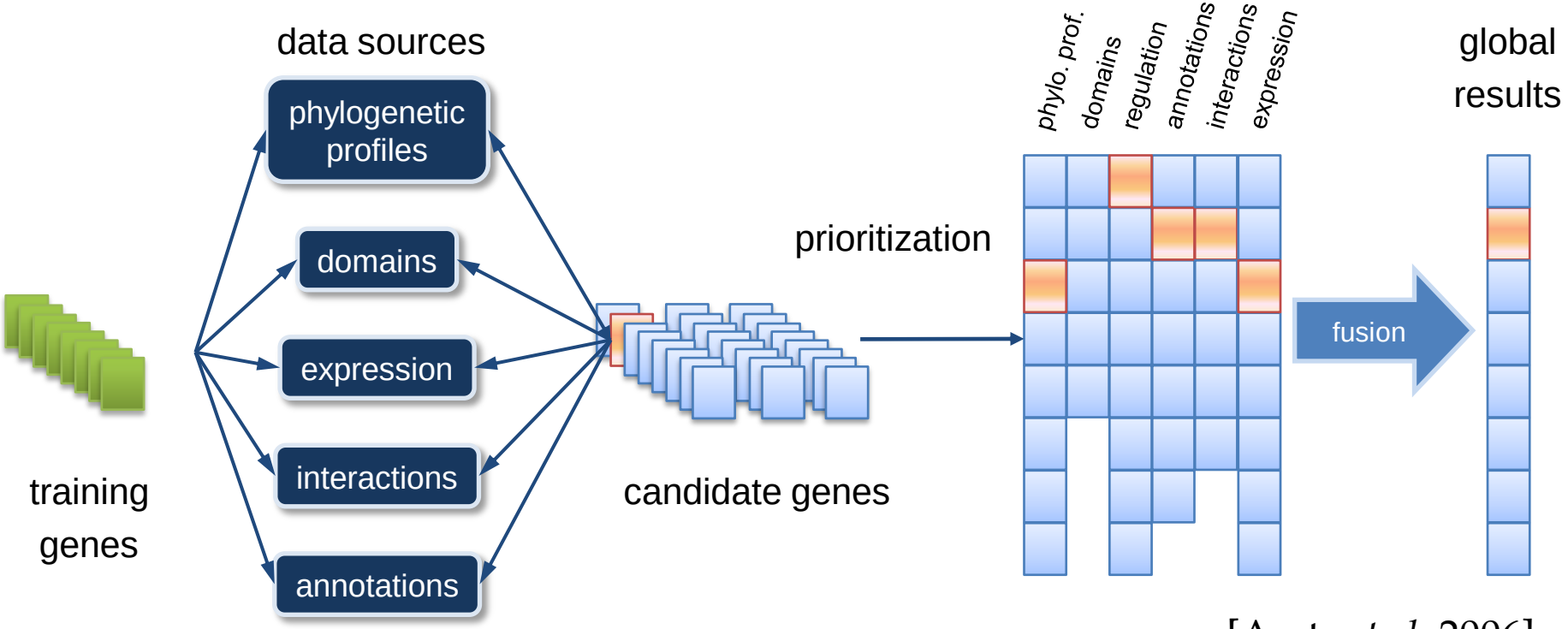
Surreprésentation :
sous-graphe pertinent



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

Candidate gene prioritization by genomic data fusion

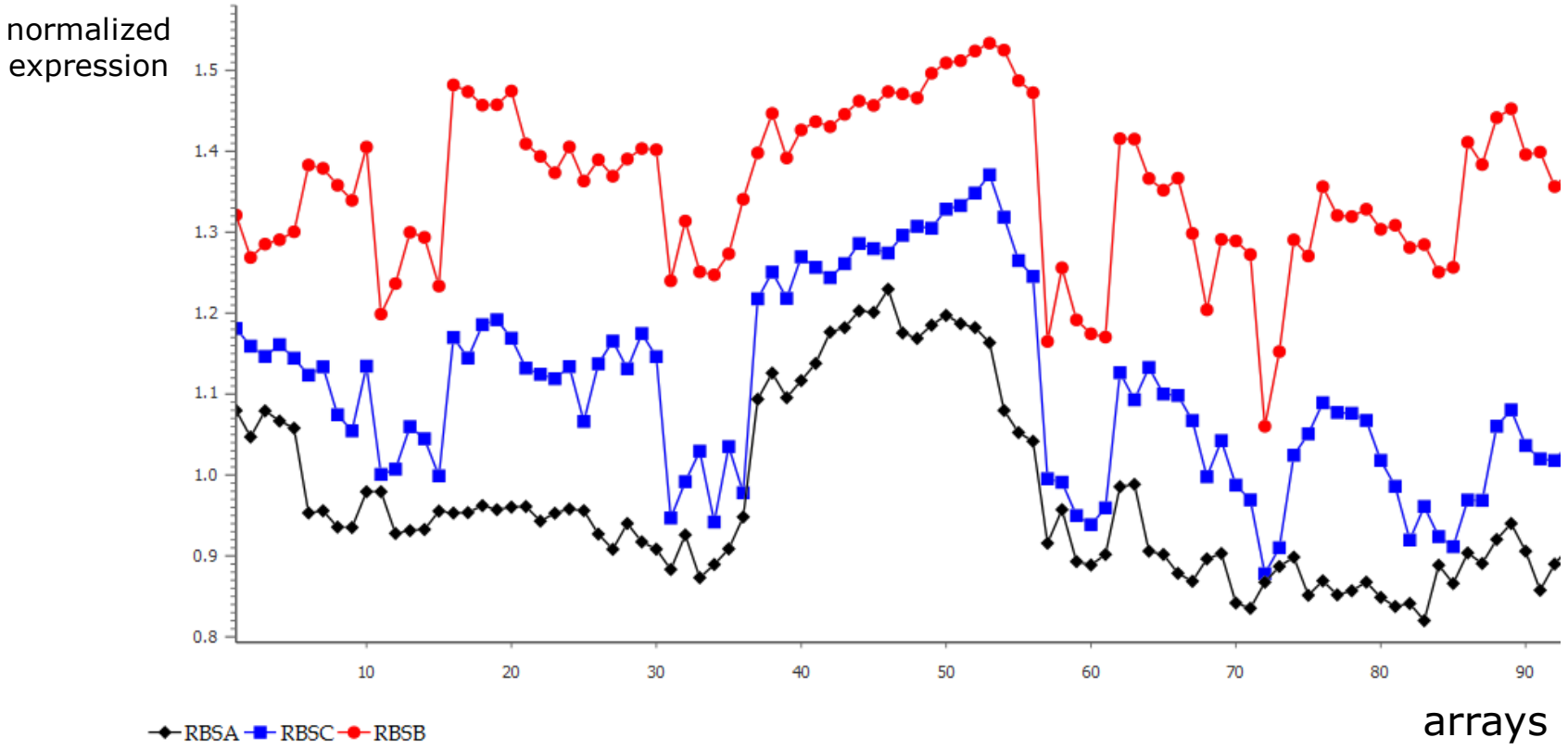
- Observation: more and more post-genomic data available
- Paradox: more difficult to select the best candidates
- Goal: objective and comprehensive evaluation of candidates



[Aerts *et al.* 2006]

Gene expression

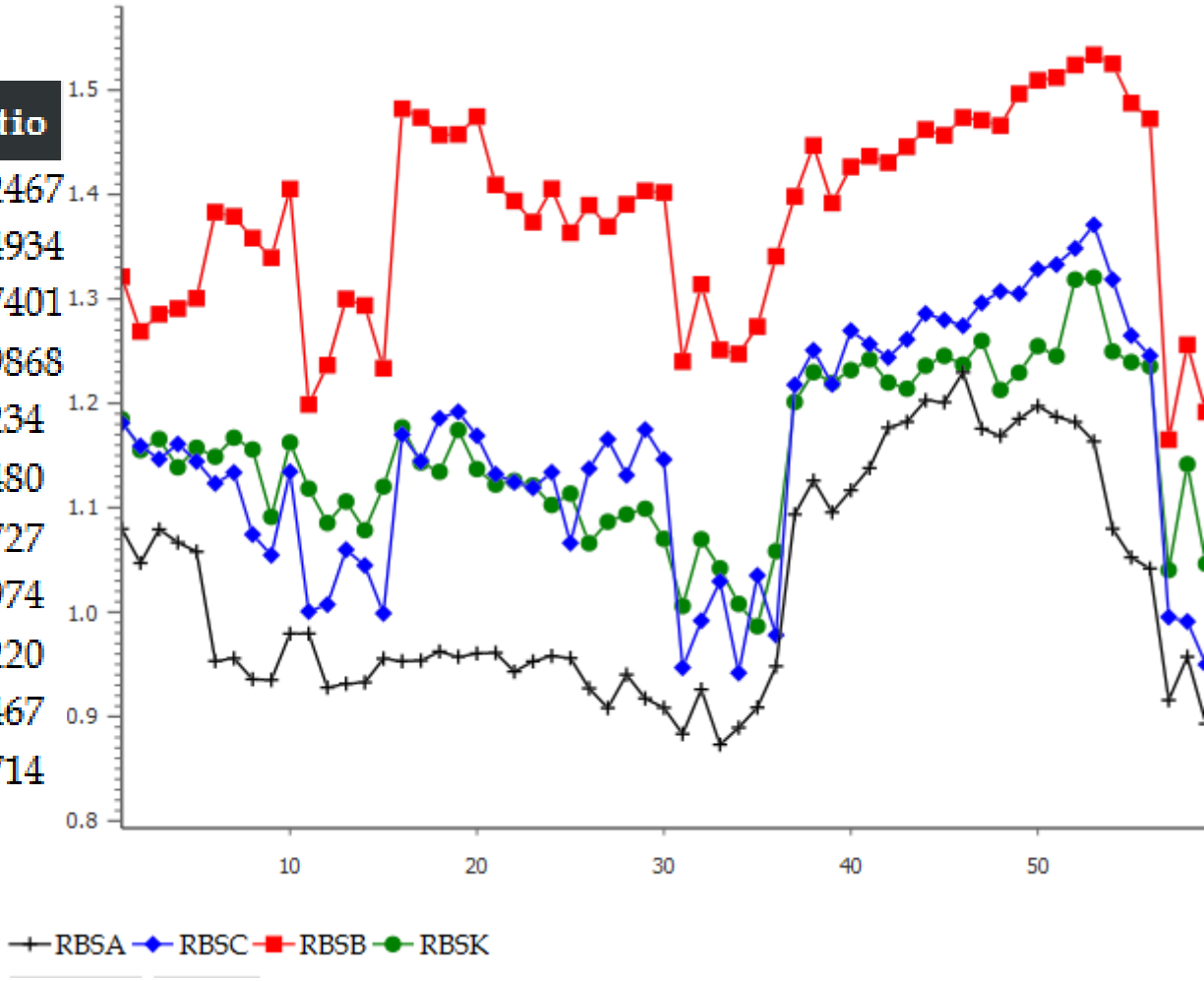
- a gene: set of expression values in various experimental conditions
- a pair of genes: dissimilarity index based on Pearson's correlation coefficient
- score : average dissimilarity



Gene expression example

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

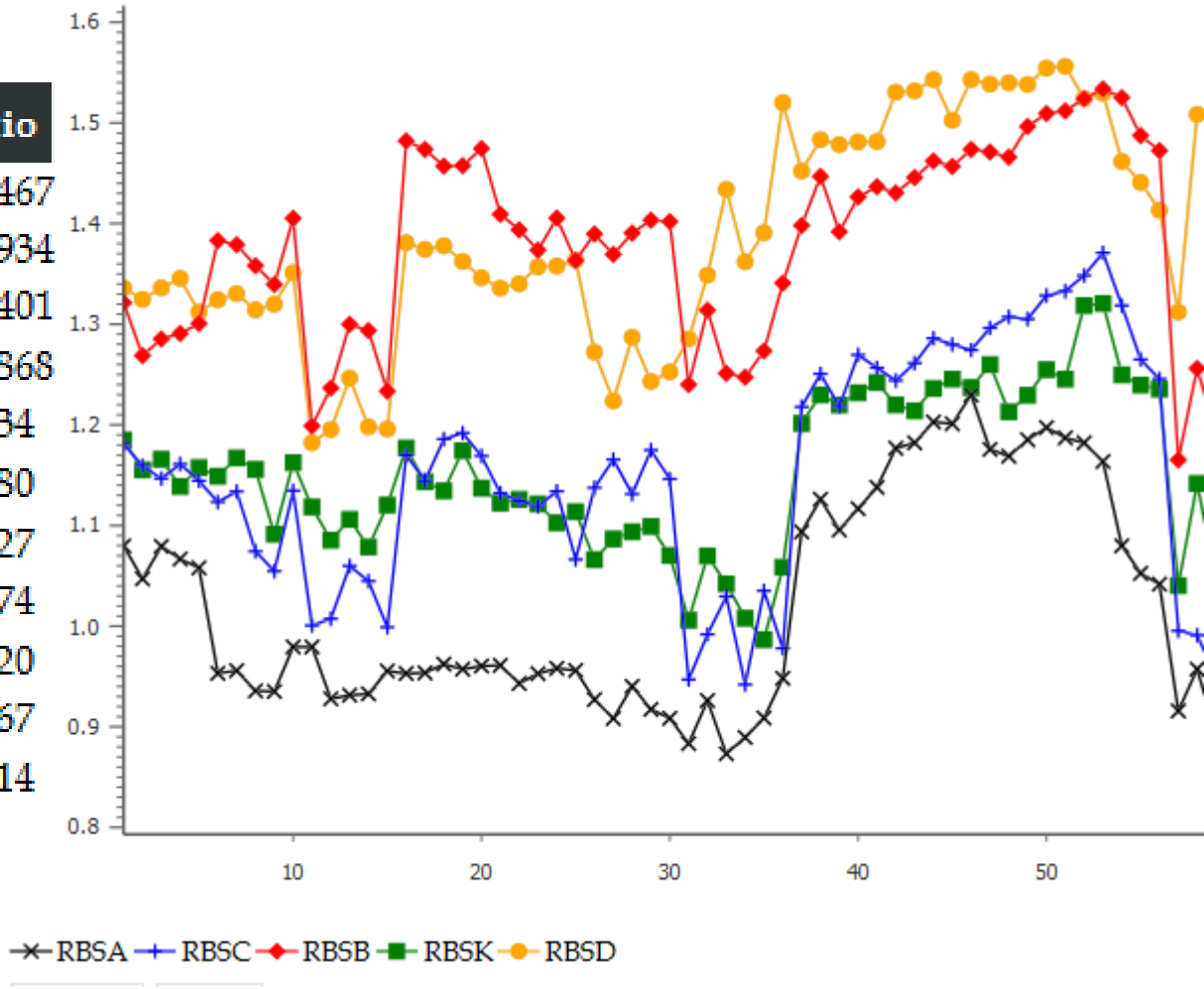
candidate	score	rank	rank ratio
RBSK	0.1870	1	0.0002467
RBSD	0.2695	2	0.0004934
FDOI	0.3288	3	0.0007401
MALE	0.3514	4	0.0009868
MALK	0.3537	5	0.001234
FDOG	0.3551	6	0.001480
FDOH	0.3670	7	0.001727
TREB	0.3679	8	0.001974
NUPG	0.3841	9	0.002220
LAMB	0.3850	10	0.002467
MALF	0.3933	11	0.002714



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

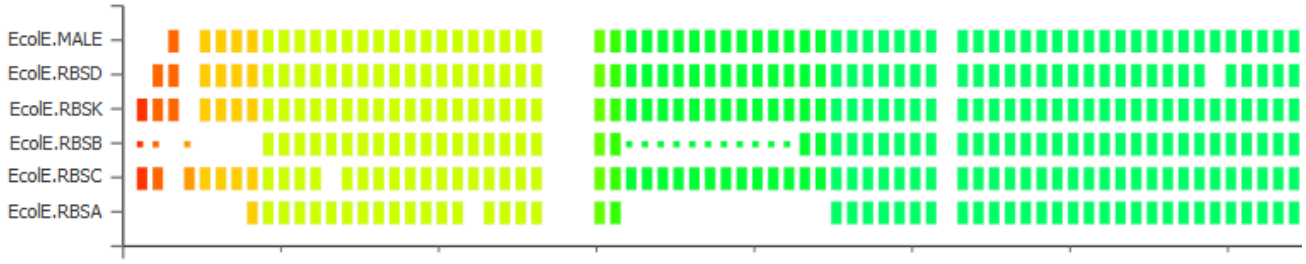


Phylogenetic profiles

- 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



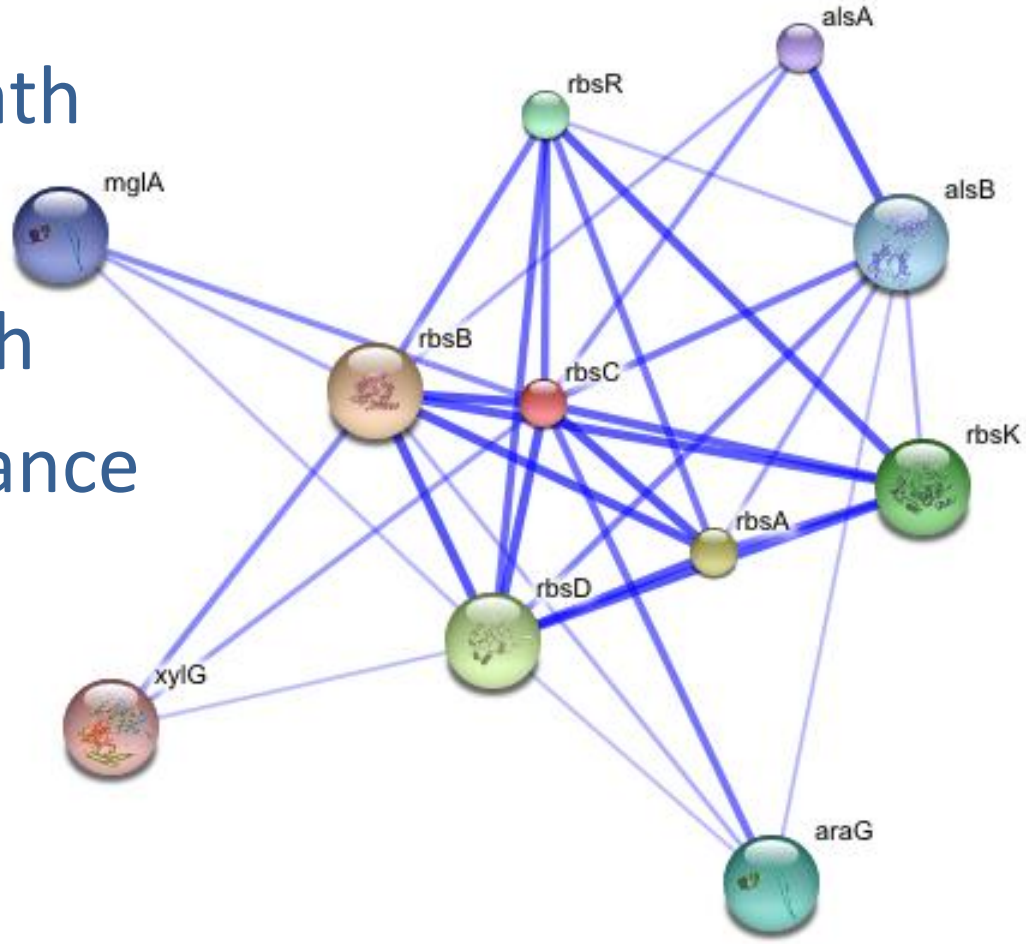
Shigella dysenteriae Iso · Shigella dysenteriae Ort · Shigella boydii Iso · Shigella boydii Ort · Sodalis glossinidius Iso · Sodalis glossinidius Ort · Shigella Iso · Salmonella Iso · Photorhabdus Iso · Pectobacterium Iso · Yersinia Iso · Yersinia Ort

- all pairs shortest path
- a pair of gene:

  shortest path length
- score: average distance

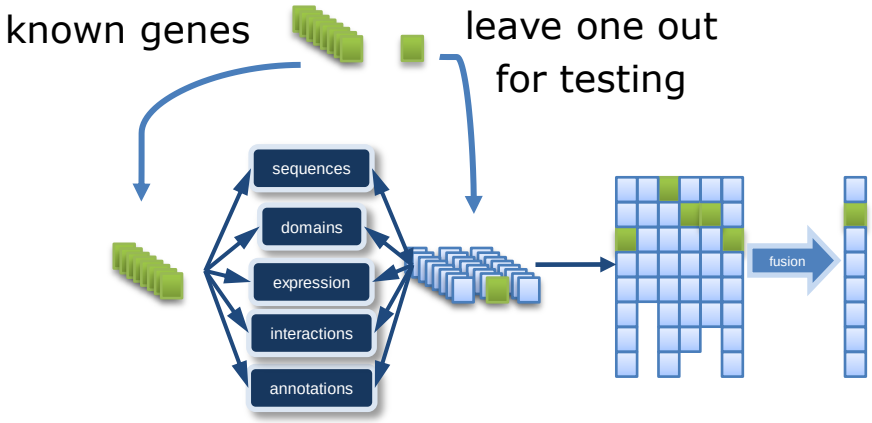
training: rbsA, rbsB, rbsC

candidate	score	rank	rank ratio
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

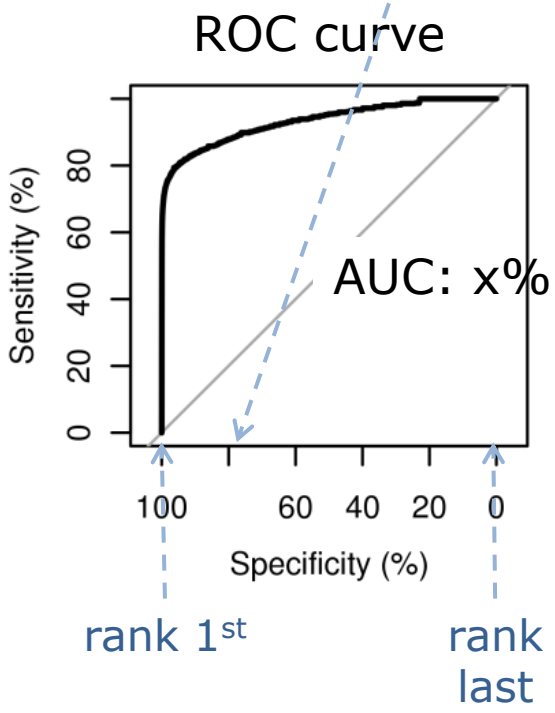
- Leave-one-out cross validation (LOOCV)



- 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)

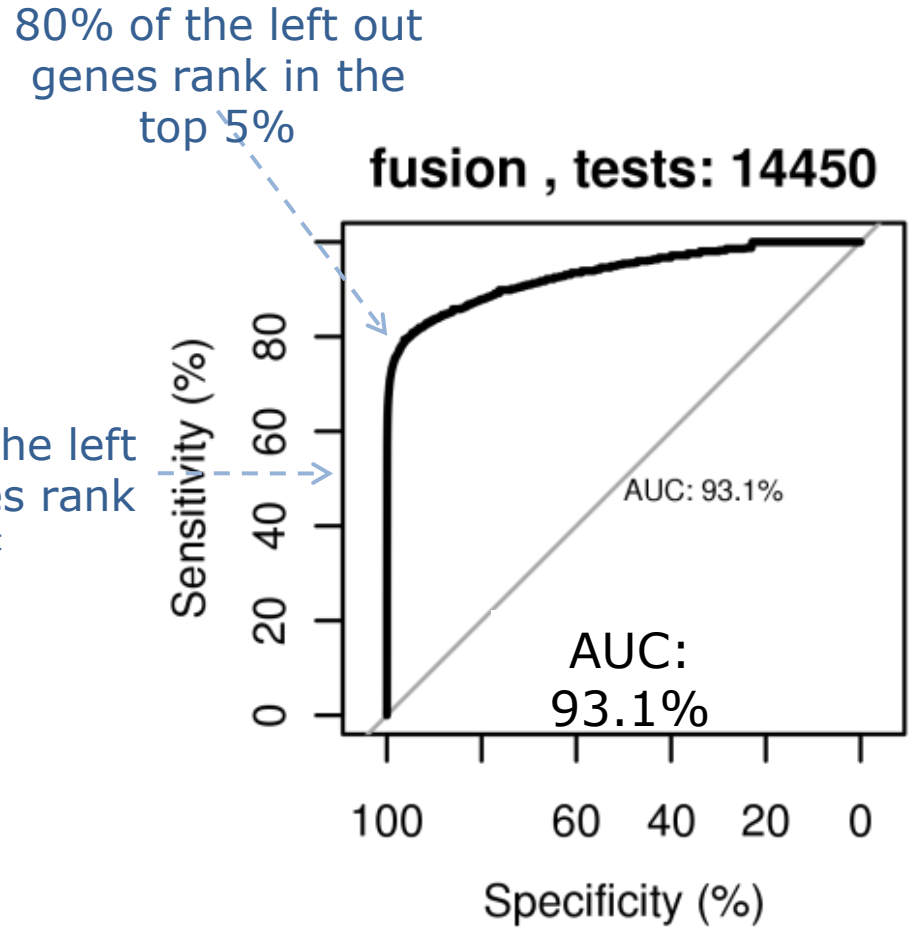
How well does it rank?

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



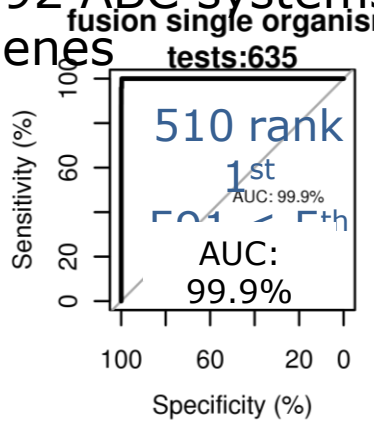
Gold standard

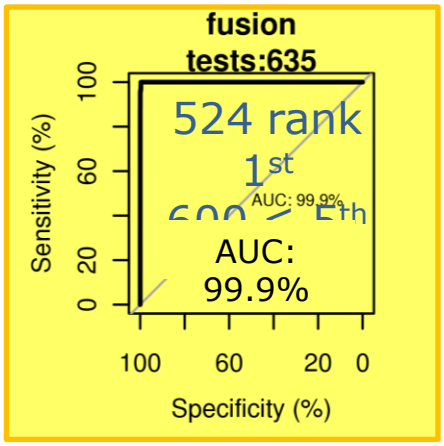
- ABCdb, manually curated ABC systems:
  - 135 genomes
  - 14,450 genes
  - 4,586 ABC systems

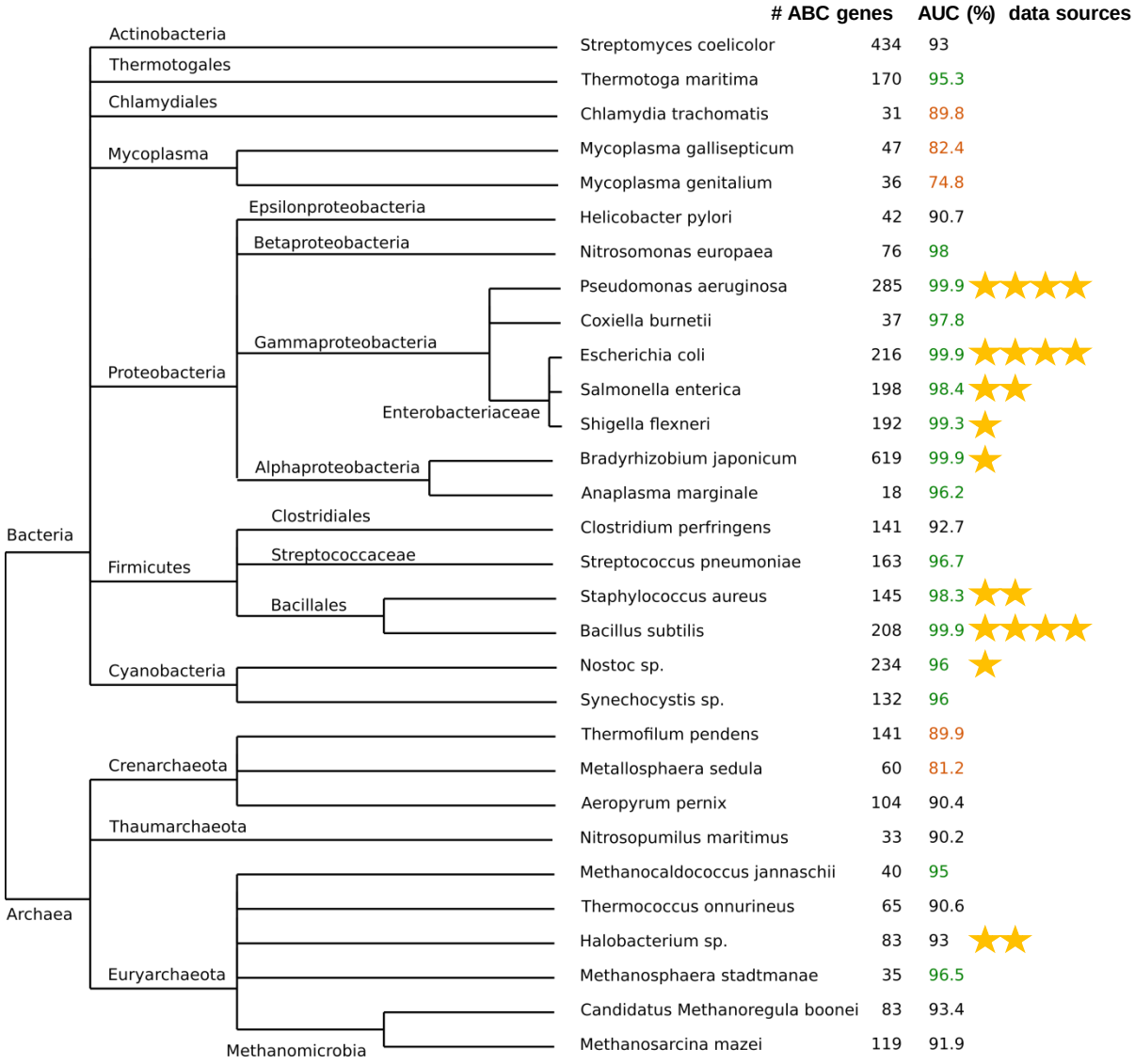


Performances: using other organisms data through orthology

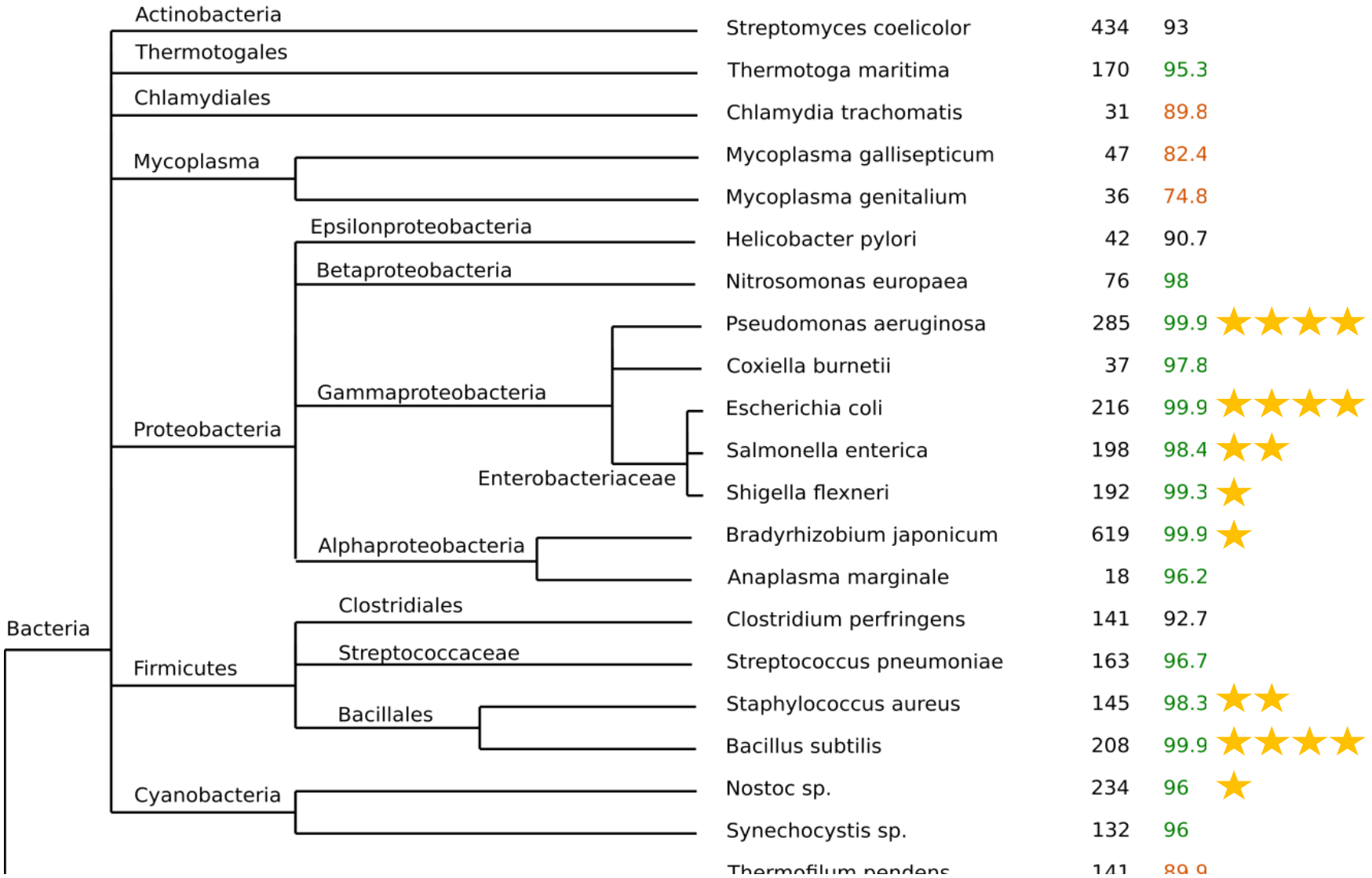
Organisms:
B. subtilis, E. coli, P. aeruginosa
192 ABC systems, 635 genes



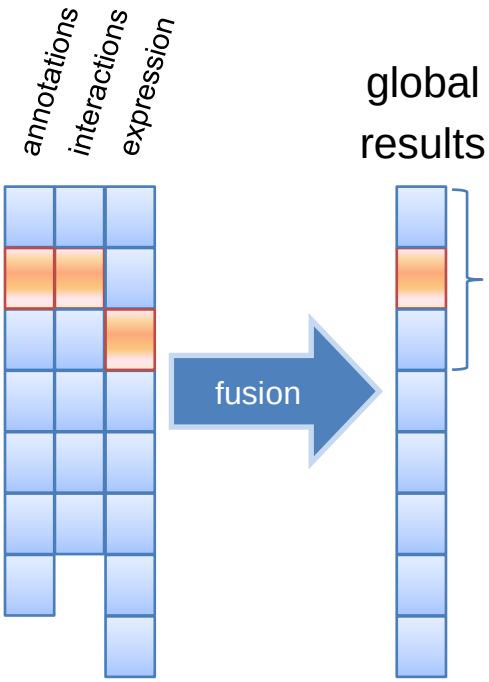




Performances: using other organisms data through orthology



Prioritization and ABC systems



Top candidates (*closest* to the training genes)

ABC systems reconstruction

- missing partner identification

Annotation and functional prediction of ABC systems

- genes associated to the system: hints on the biological process of the genes of interest

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