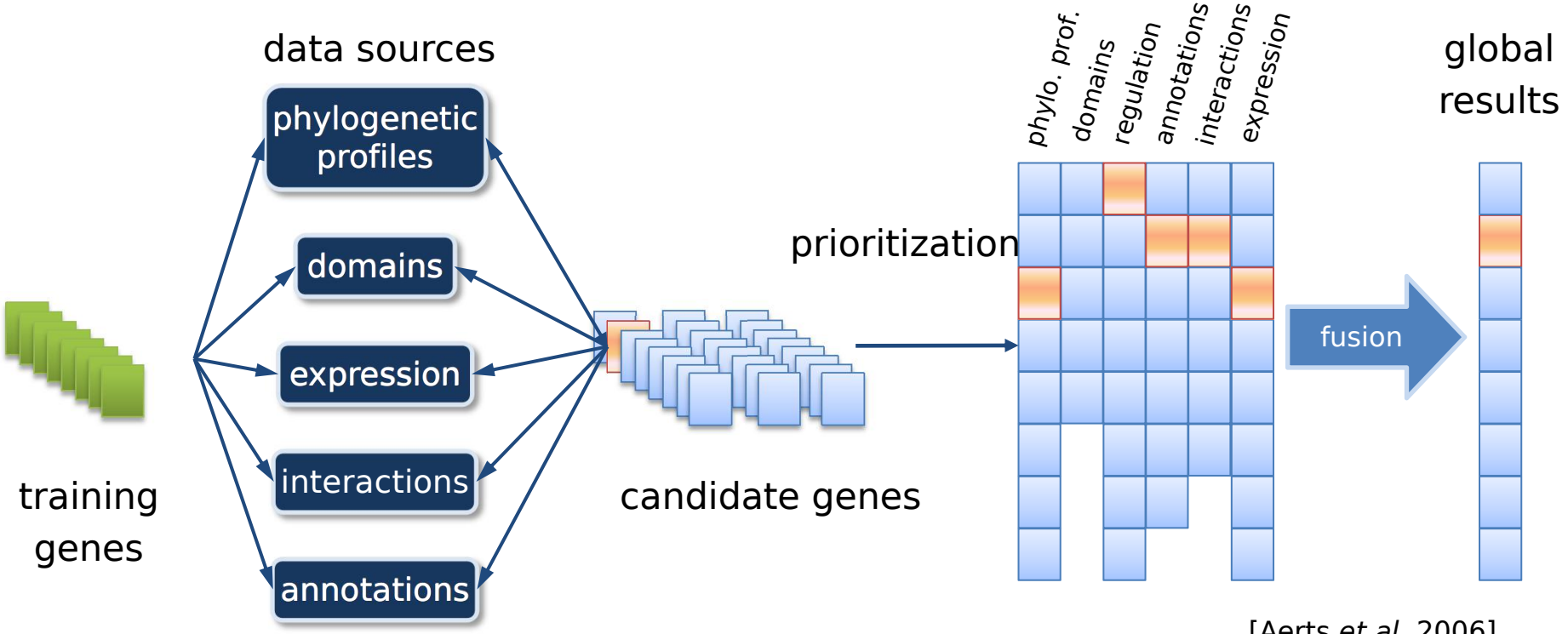
Intégration de données hétérogènes
Priorisation

Master 2

Bioinformatique et Biologie des Systèmes

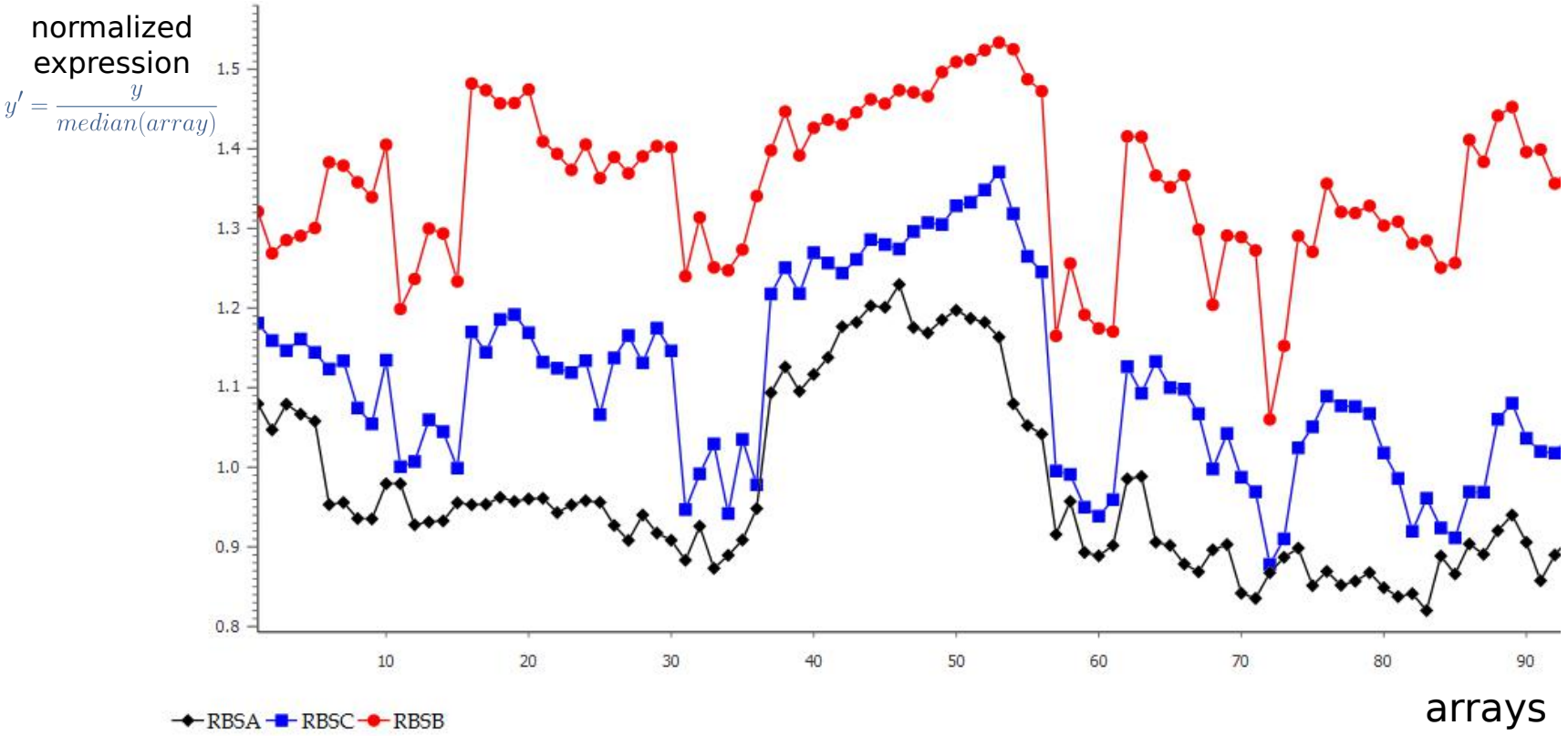
Candidate gene prioritization by genomic data fusion

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

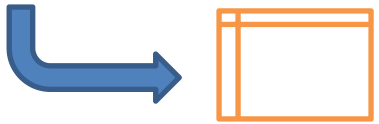


[Aerts *et al.* 2006]

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

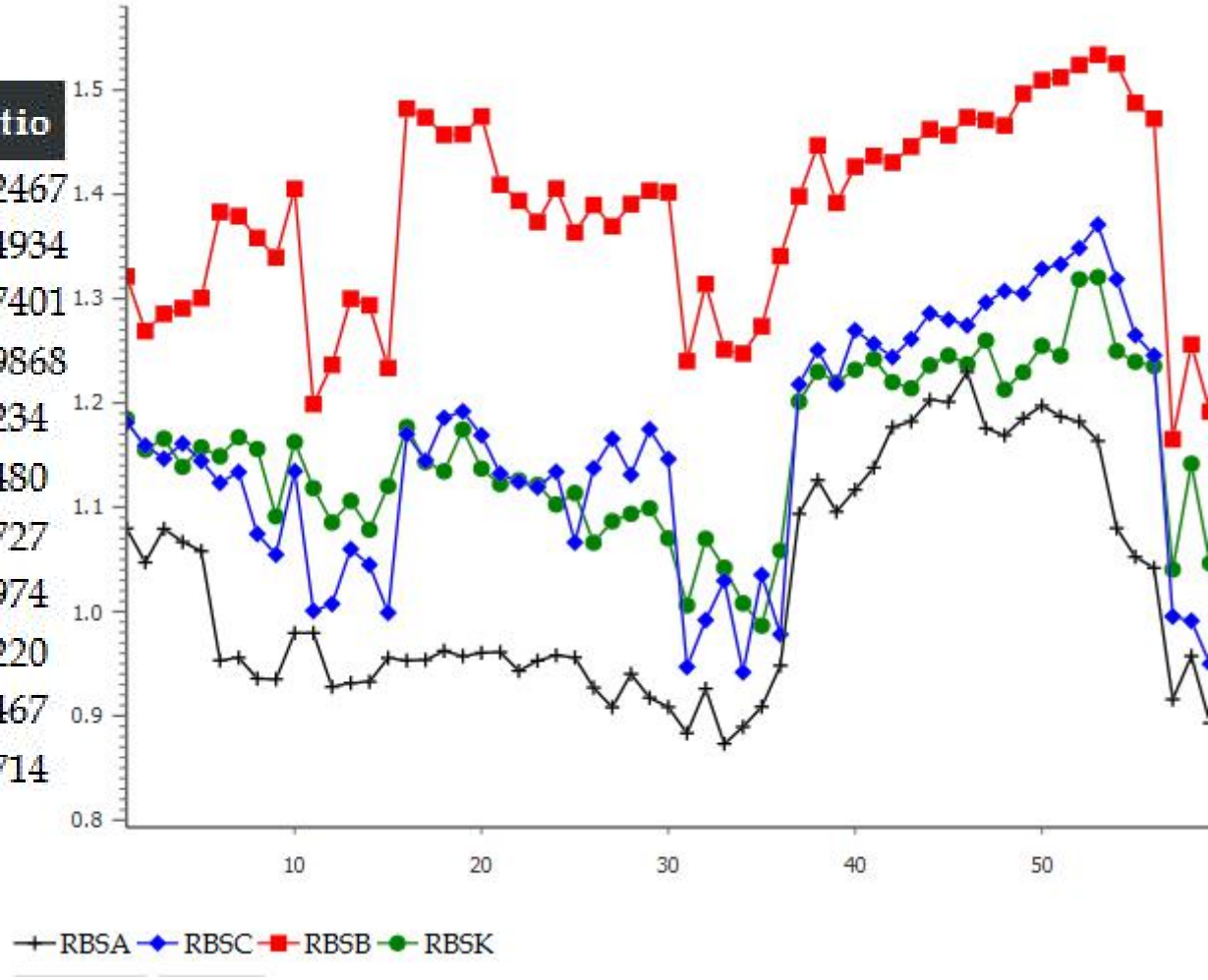


gene pairwise dissimilarity matrix

Gene expression illustration

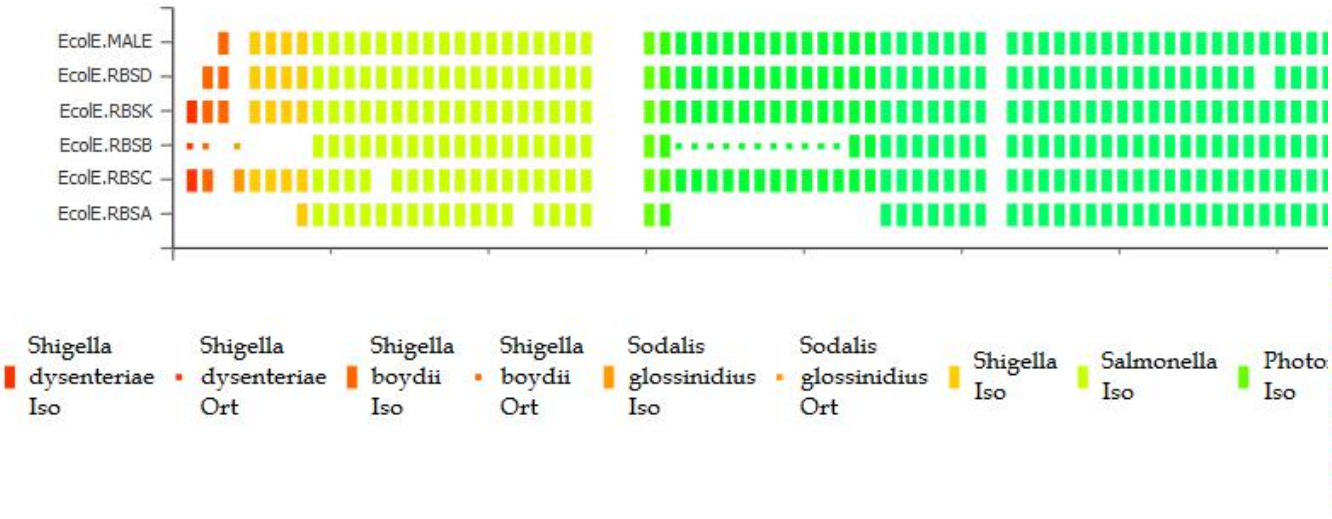
- training: rbsA, rbsB, rbsC in *E. coli* K-12

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



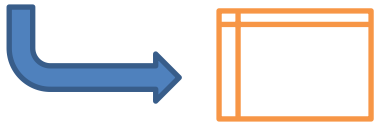
Phylogenetic profiles

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

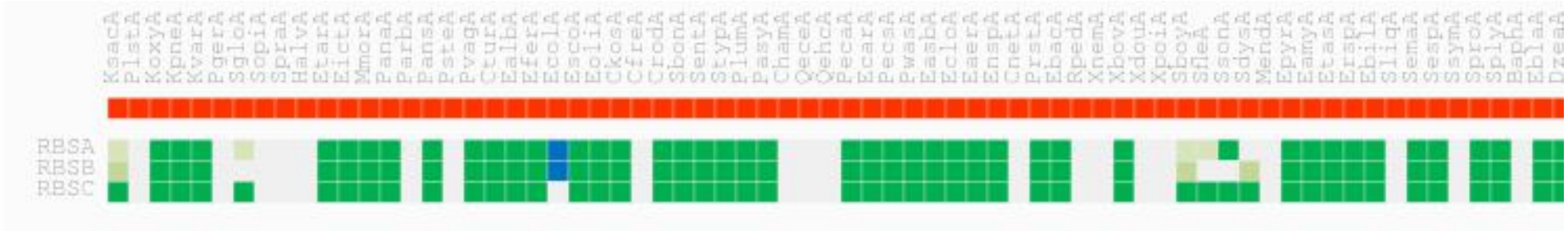
- a gene: presence/absence of orthologs 1:1 in other genomes
- pair of genes: dissimilarity index based on the Jaccard index
- score: average dissimilarity



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

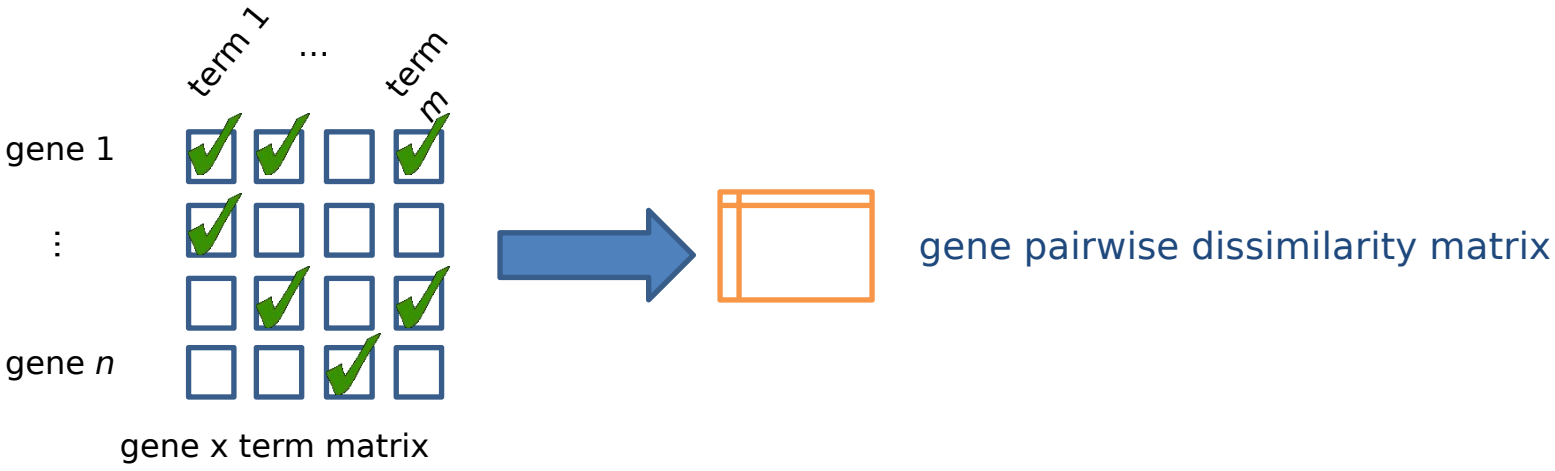
- Need to estimate the relevance of a genome with respect to
  - A genome of interest: is it not too closely related? is it informative?
  - A set of already selected reference genomes: redundancy *vs.* additional signal

- 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



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]

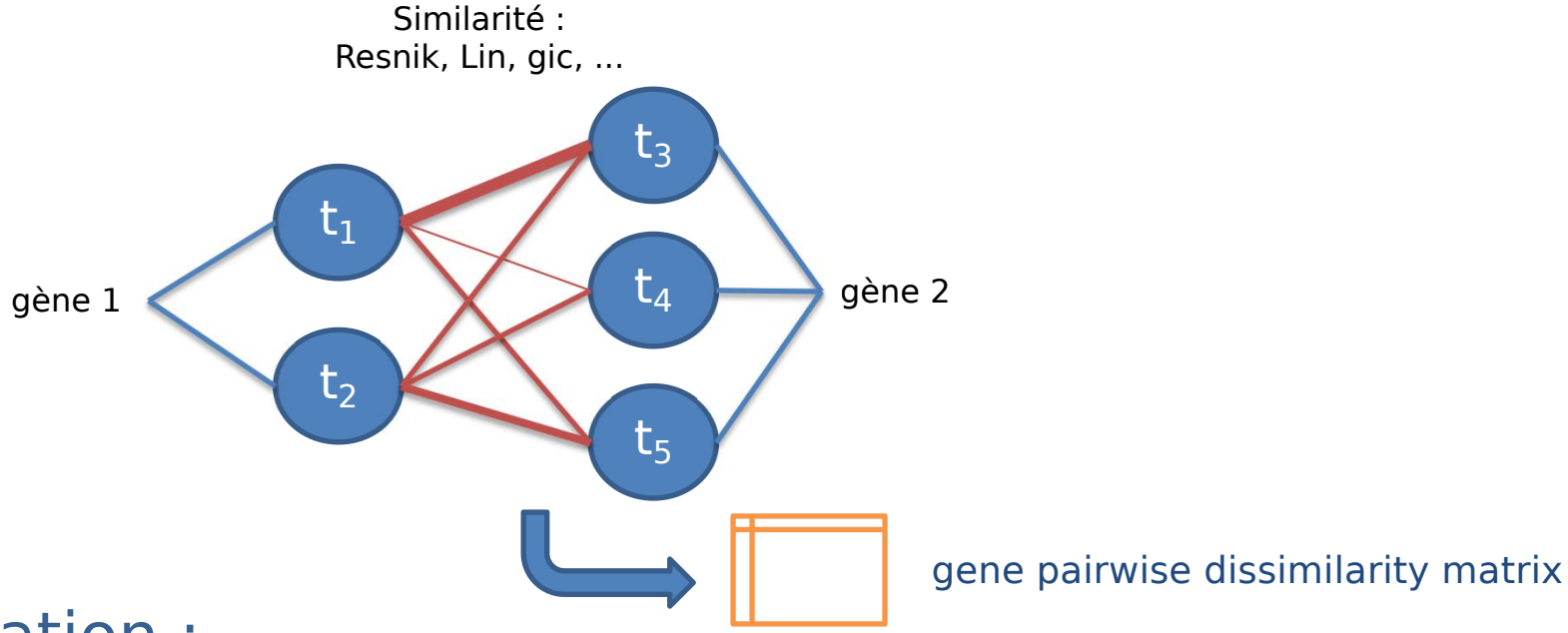
- $sim_{lin}(t_1,t_2) = IC(MICA(t_1,t_2))/(IC(t_1)+IC(t_2))$ [Lin, 1998]

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

Simimlarité entre gènes basée sur la similarité entre termes GO

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)

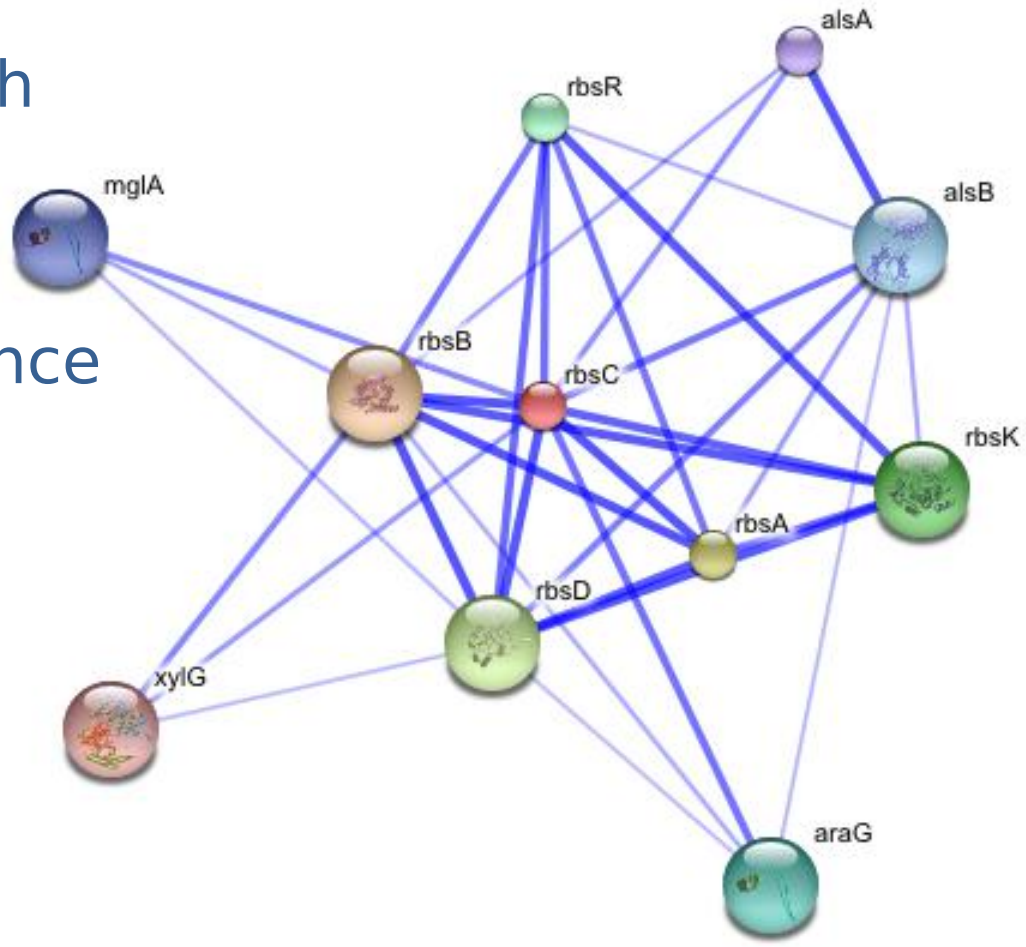


Application :

- Performances légèrement meilleures obtenues que les autres avec la combinaison Resnik + similatité 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



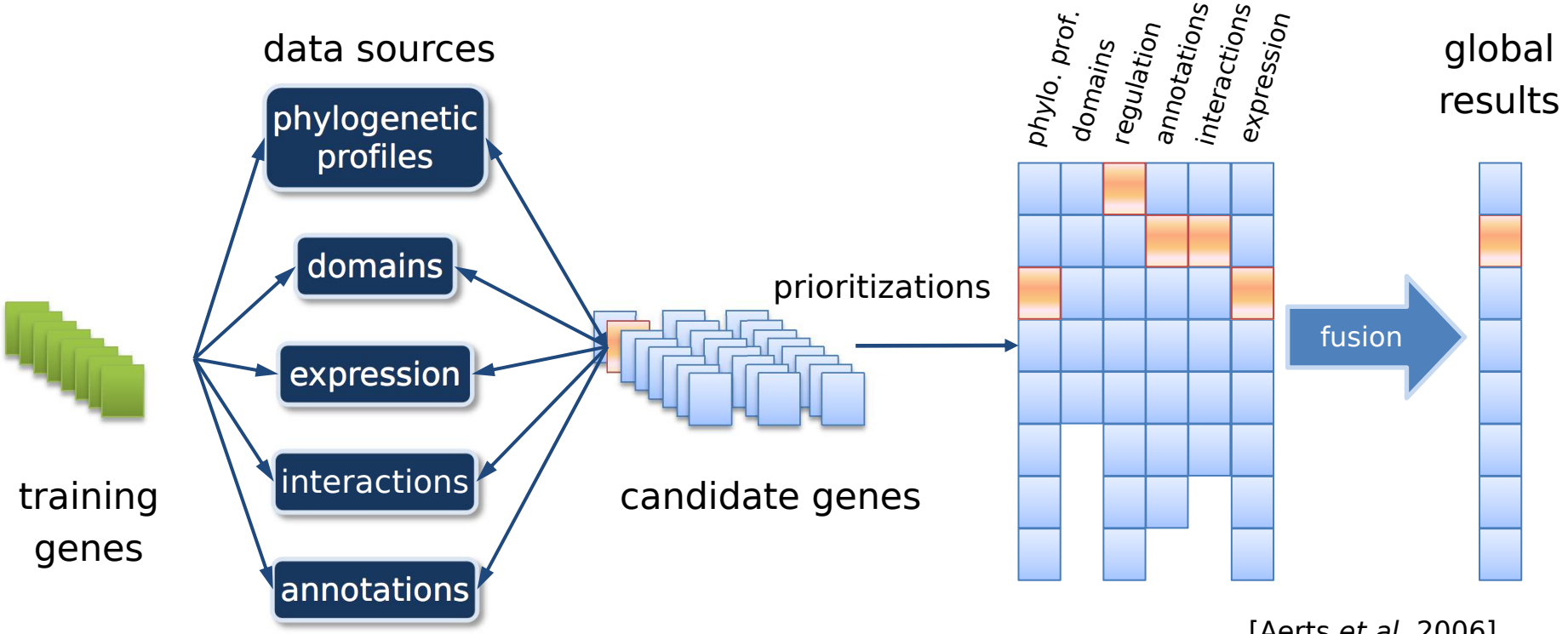
from STRING
http://string-db.org

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

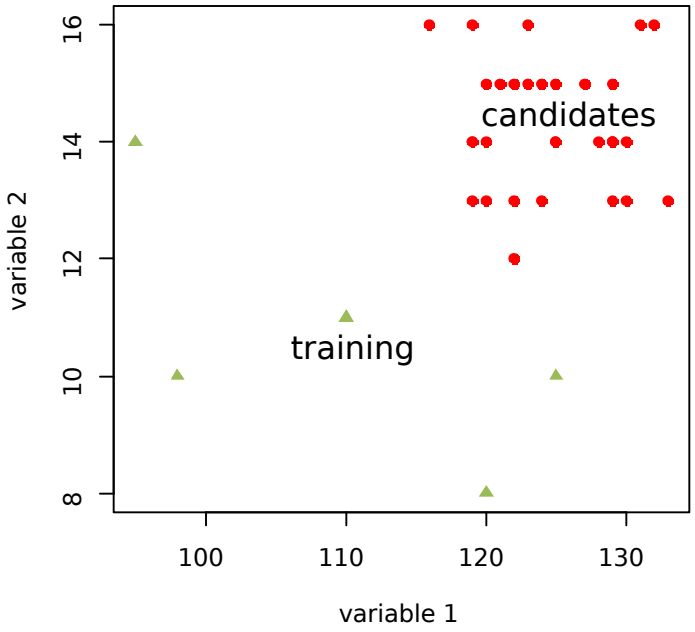
- Observation: more and more post-genomic data available
- Paradox: more difficult to select the best candidates
- Goal: objective and comprehensive evaluation of candidates



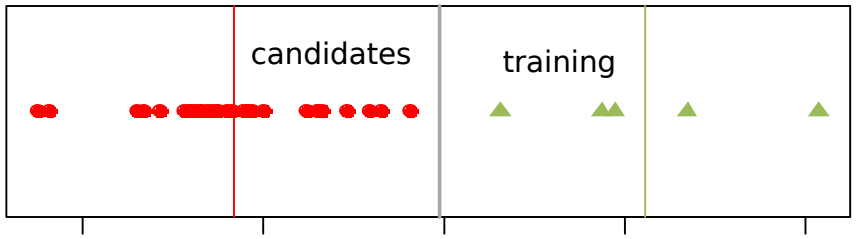
[Aerts *et al.* 2006]

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)

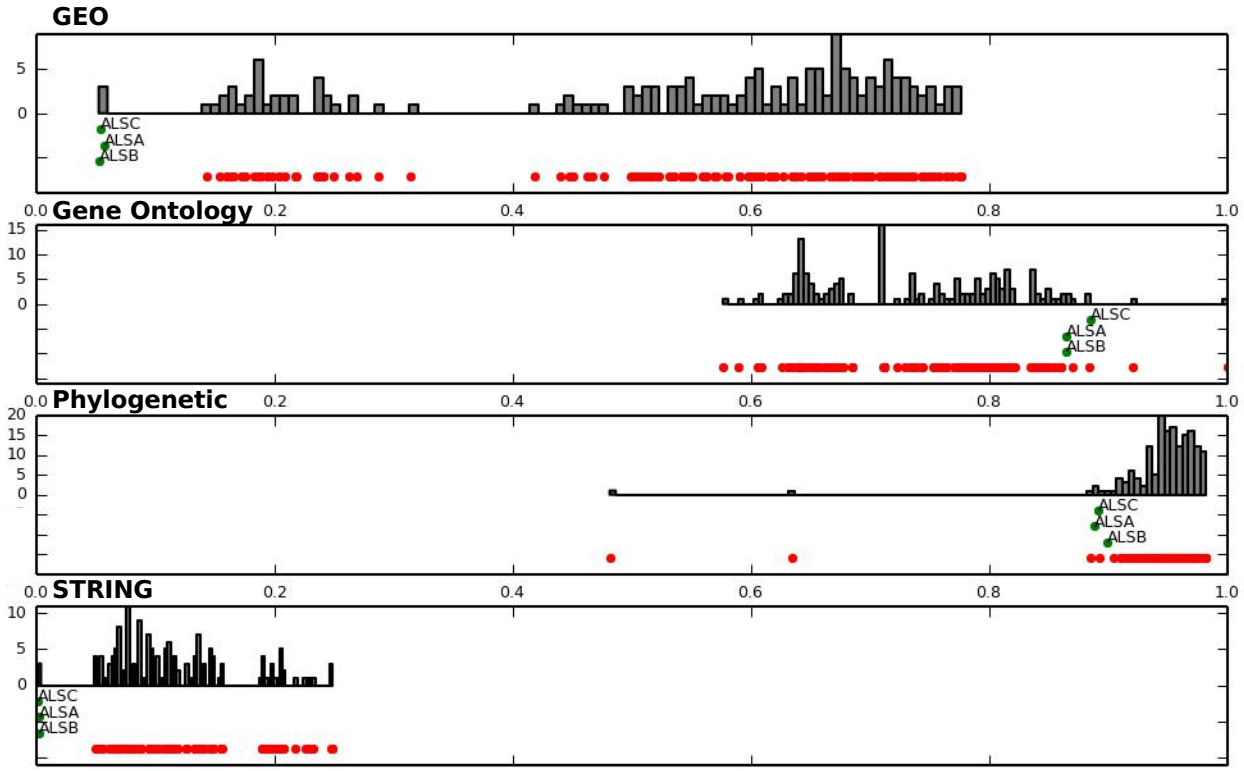


dimension	weight
variable 1	-0.1307346
variable 2	-0.7031850



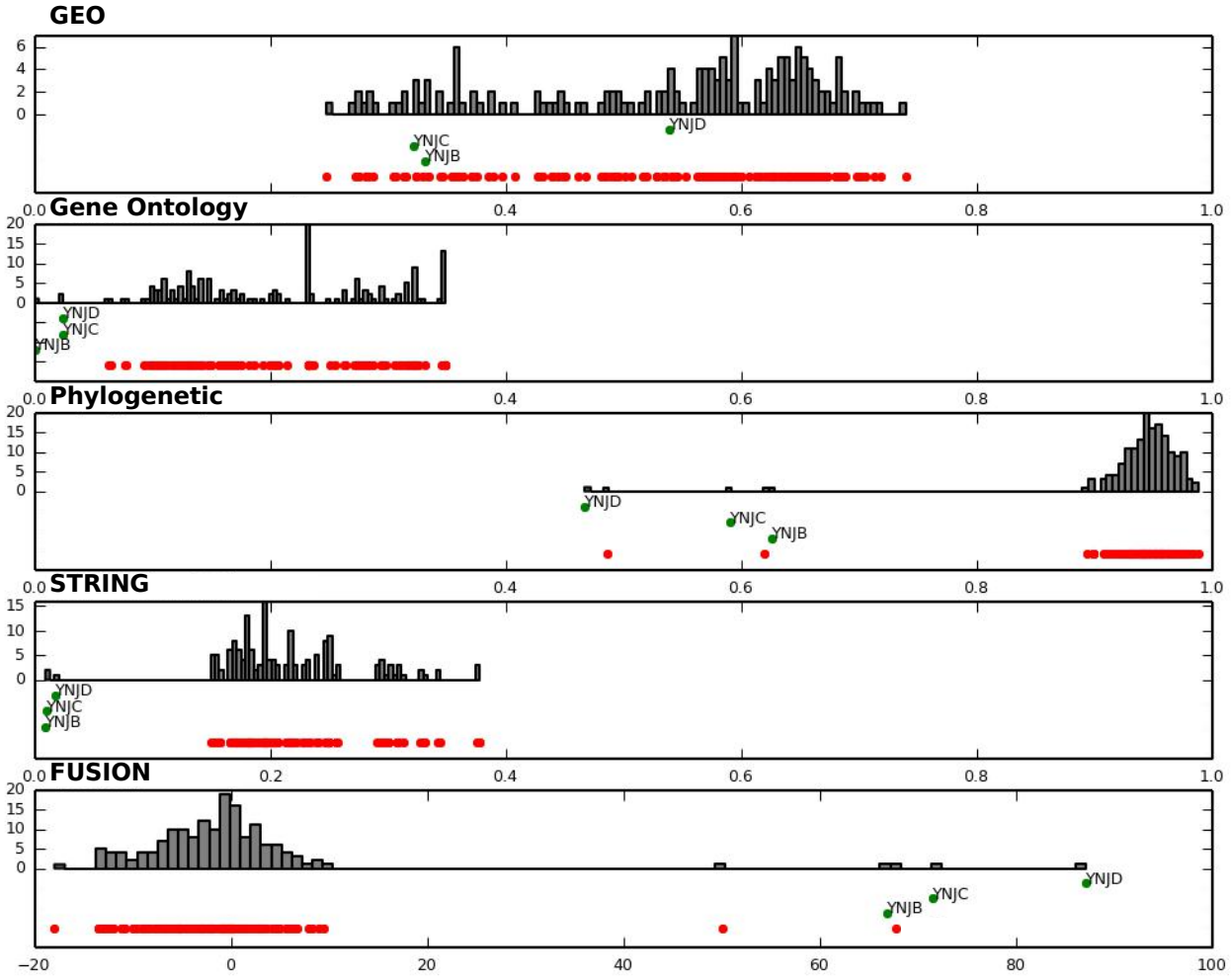
Application to *E. coli* alsA system: *alsA, alsB, alsC*

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



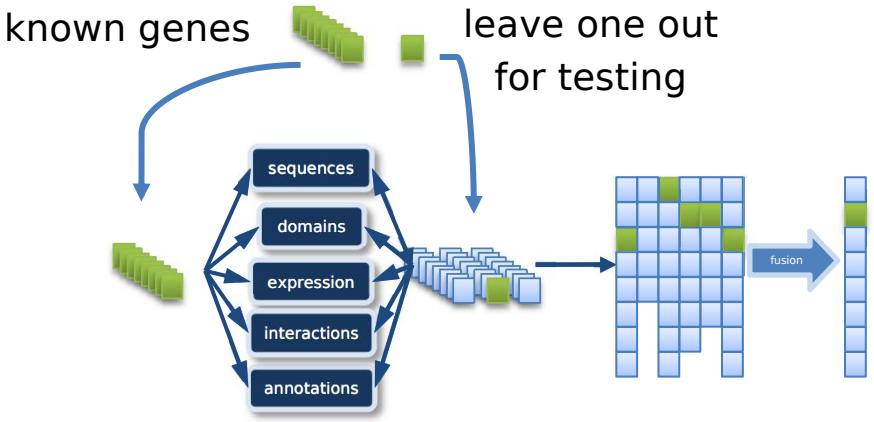
Application to *E. coli* ynjD system: *ynjB, ynjC, ynjD*

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



Evaluation methodology

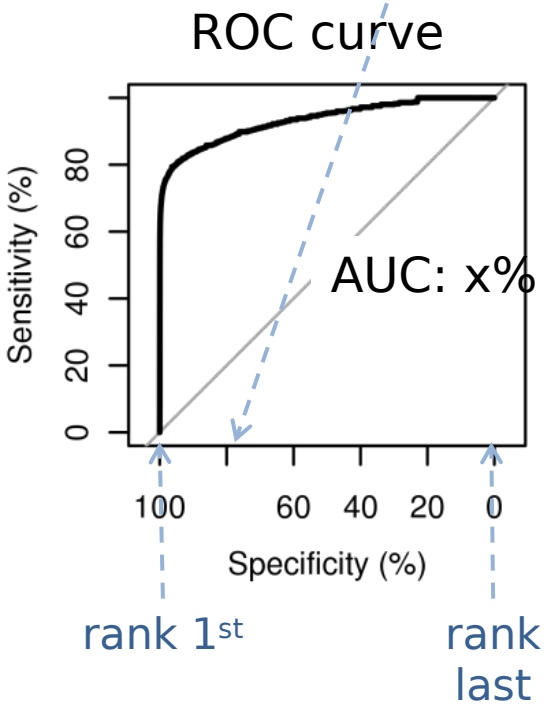
- Leave-one-out cross validation (LOOCV)



- 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

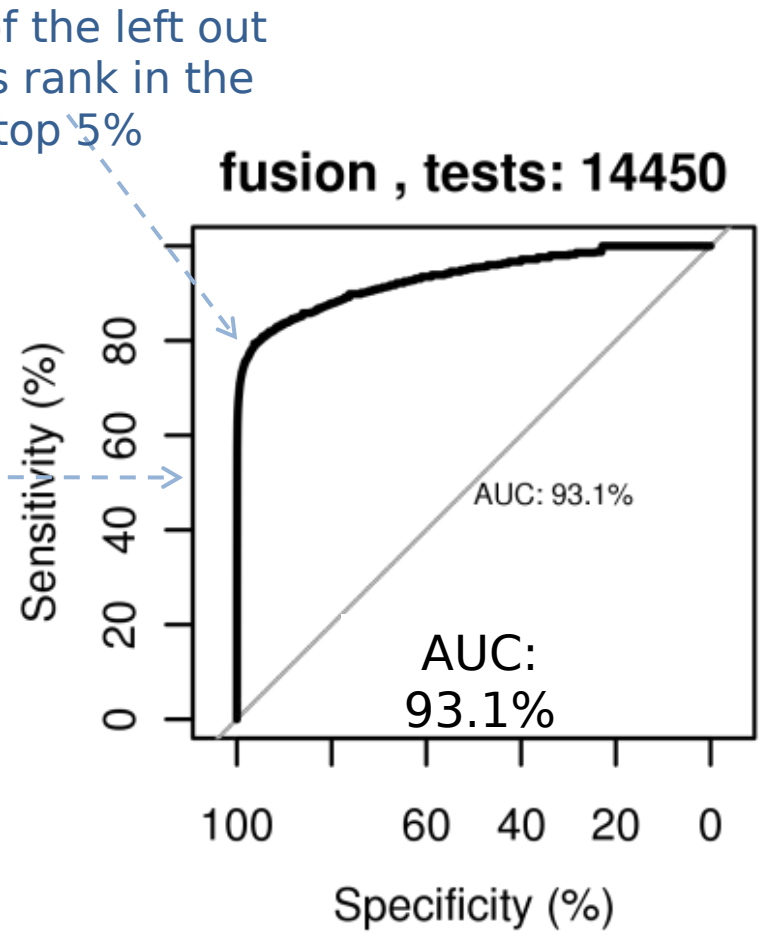


Evaluation: global results

Gold standard

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

53% of the left out genes rank 1st



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

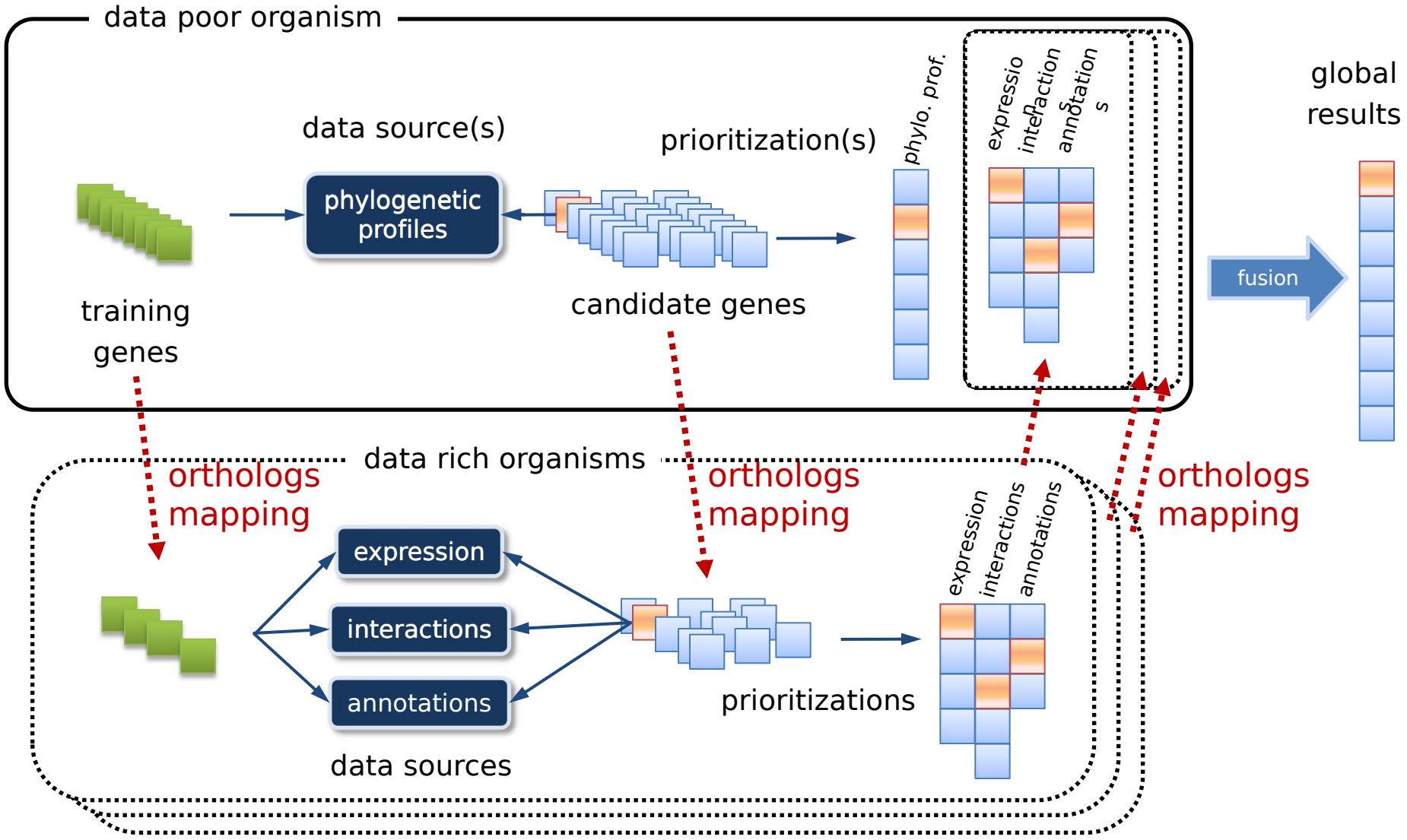
Prioritization Hide Hide

Run prioritization.

Show 10 entries Search:

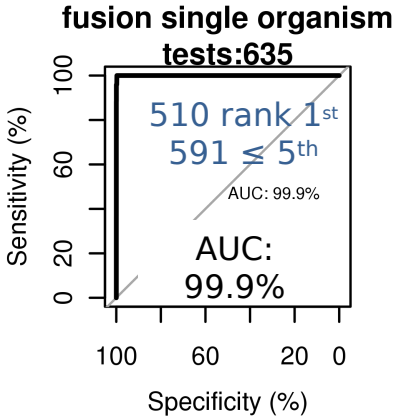
rank	Global results	pathways (fusion)	string (fusion)	transcriptome (fusion)	phylogenetic_profiles EcolE	go (fusion)	interactome EcolE
1	RBSD (1) S: 0, RR: 0	D-ribose pyranase					
2	RBSK (2) S: 0, RR: 0	Ribokinase					
3	MALE (3) S: 0, RR: 0.001	SBP of maltose/maltodextrin/maltoologisaccharide ABC transporter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

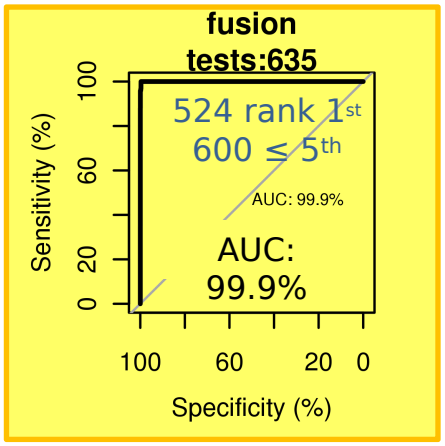
Extension of the method to data poor organisms

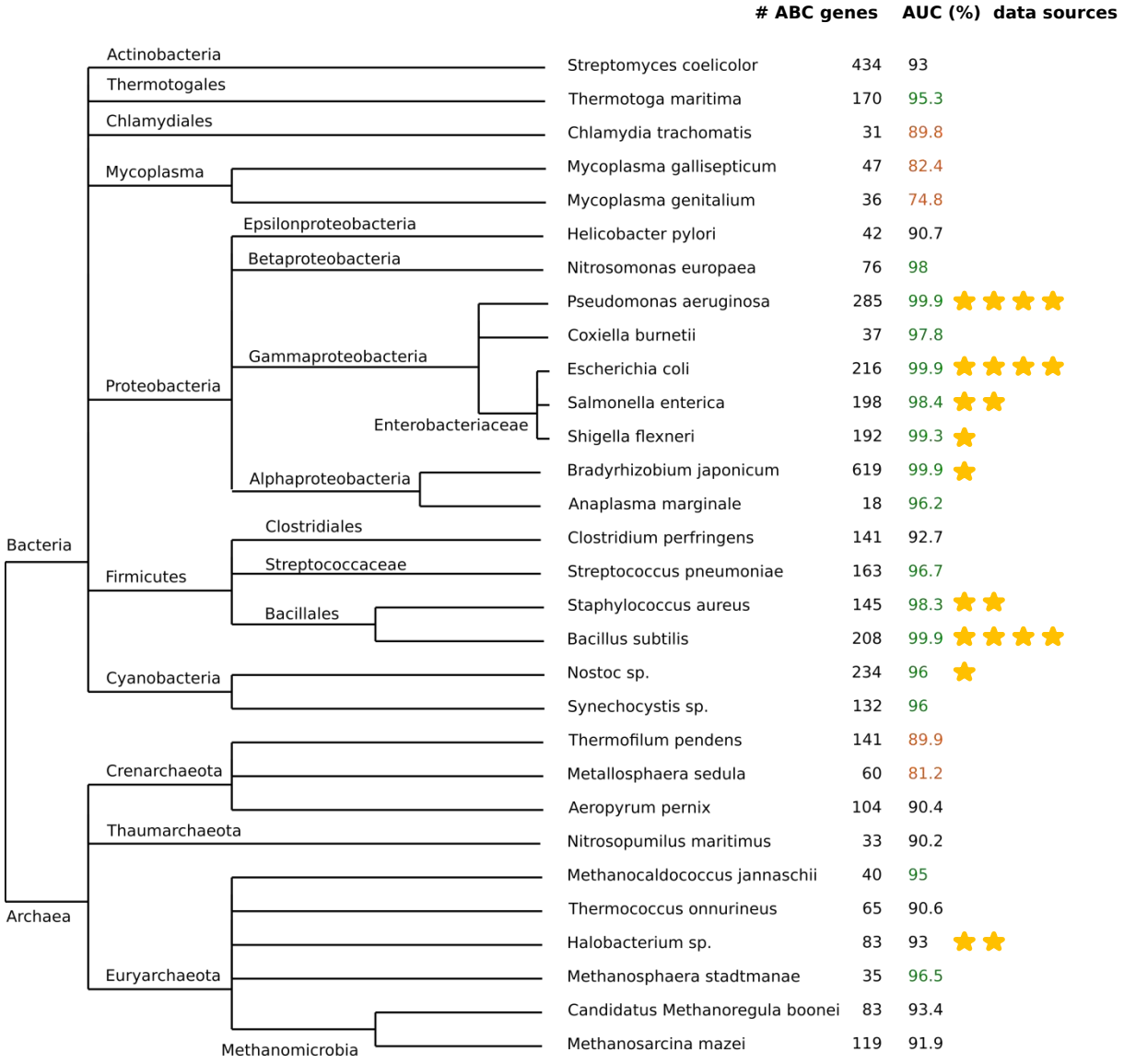


Performances: using other organisms data through orthology

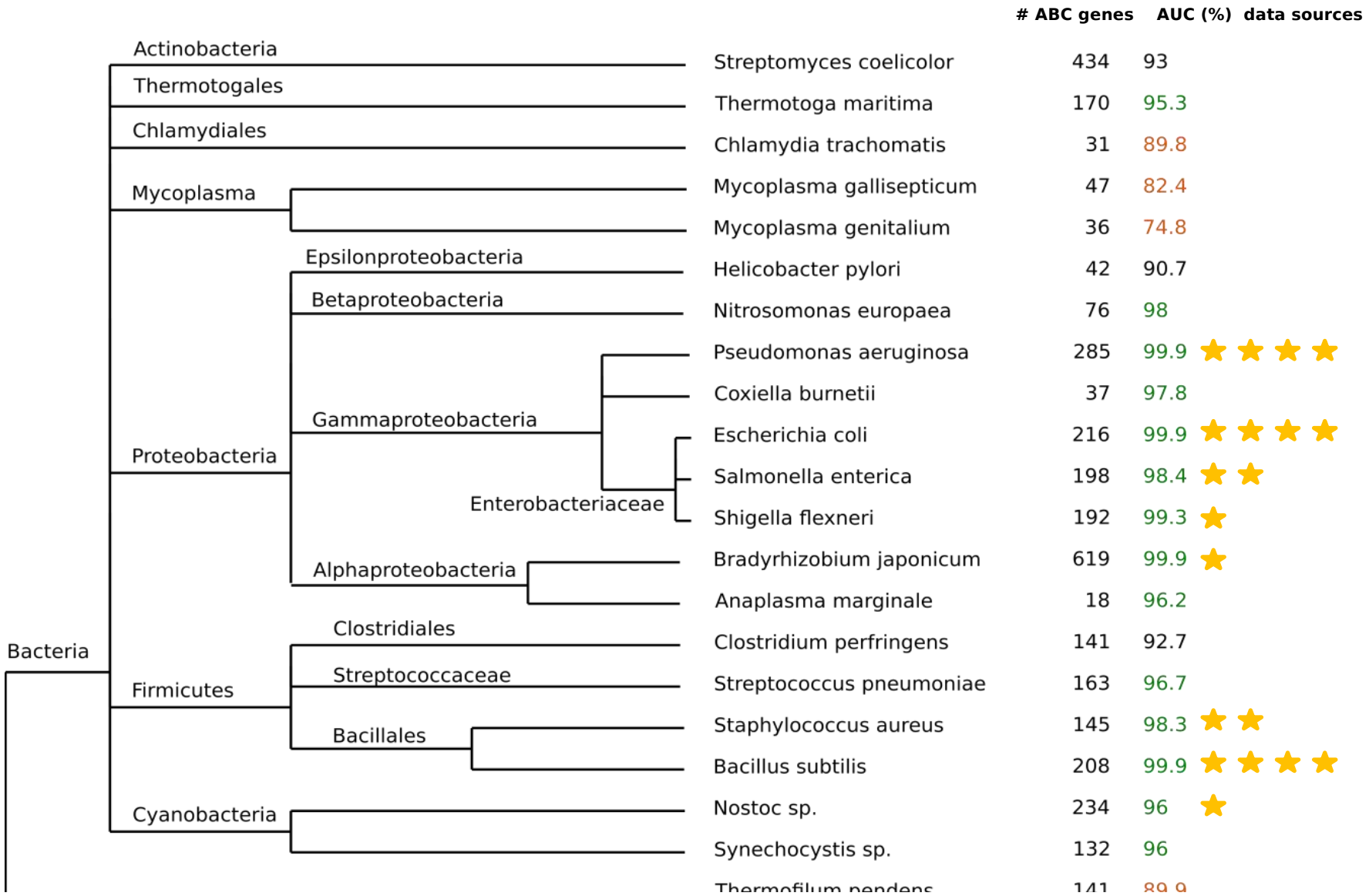
Organisms:
B. subtilis, E. coli, P. aeruginosa
192 ABC systems, 635 genes





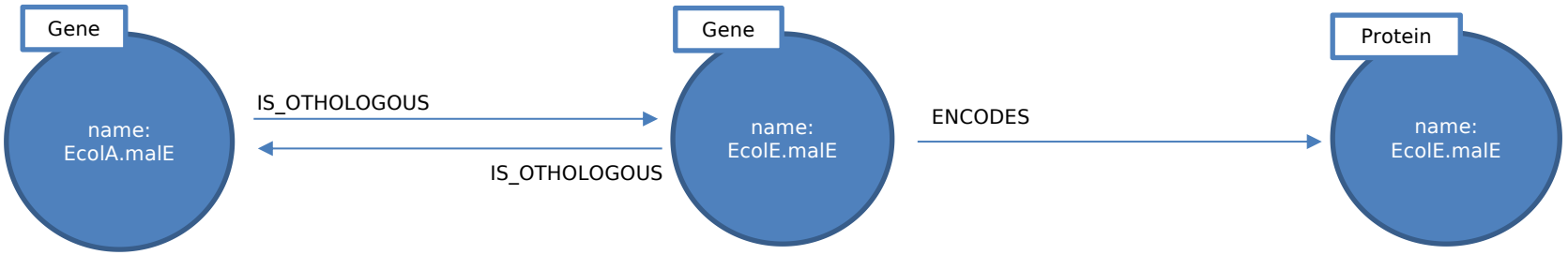


Performances: using other organisms data through orthology



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

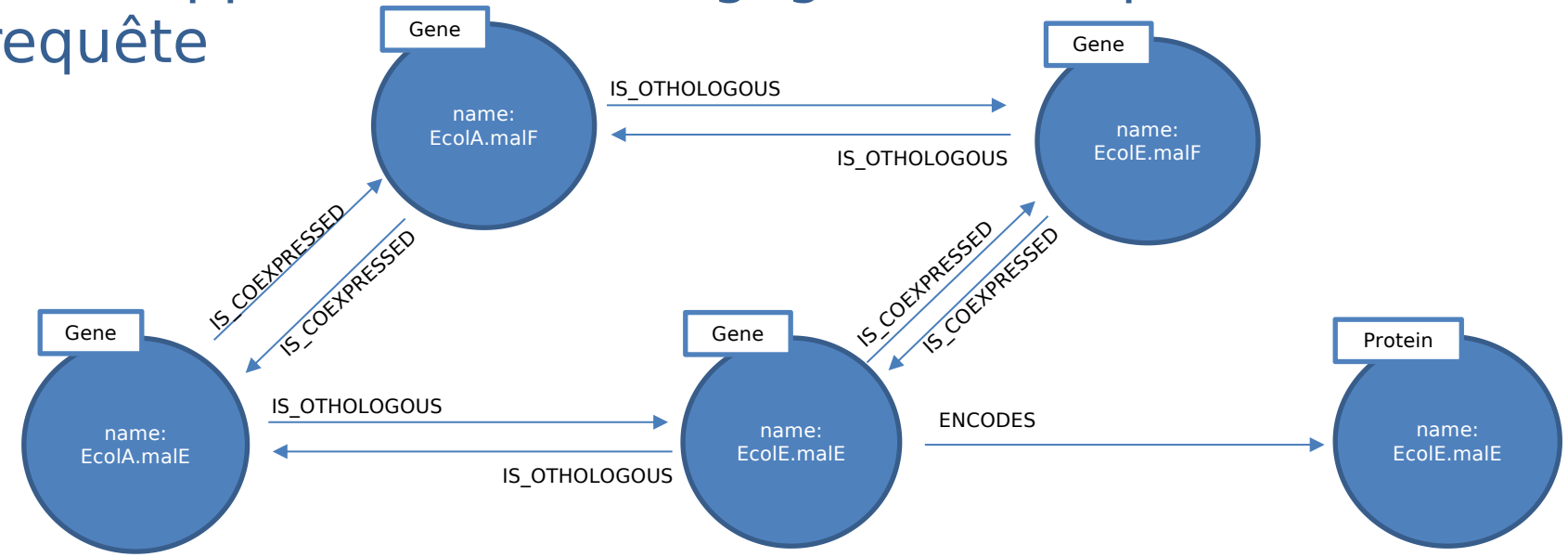


Labeled Property Graph

```
(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

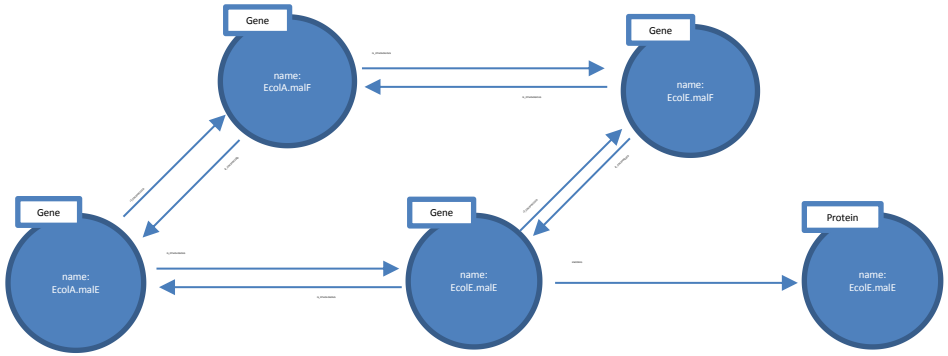


Labeled Property Graph

```
(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
```