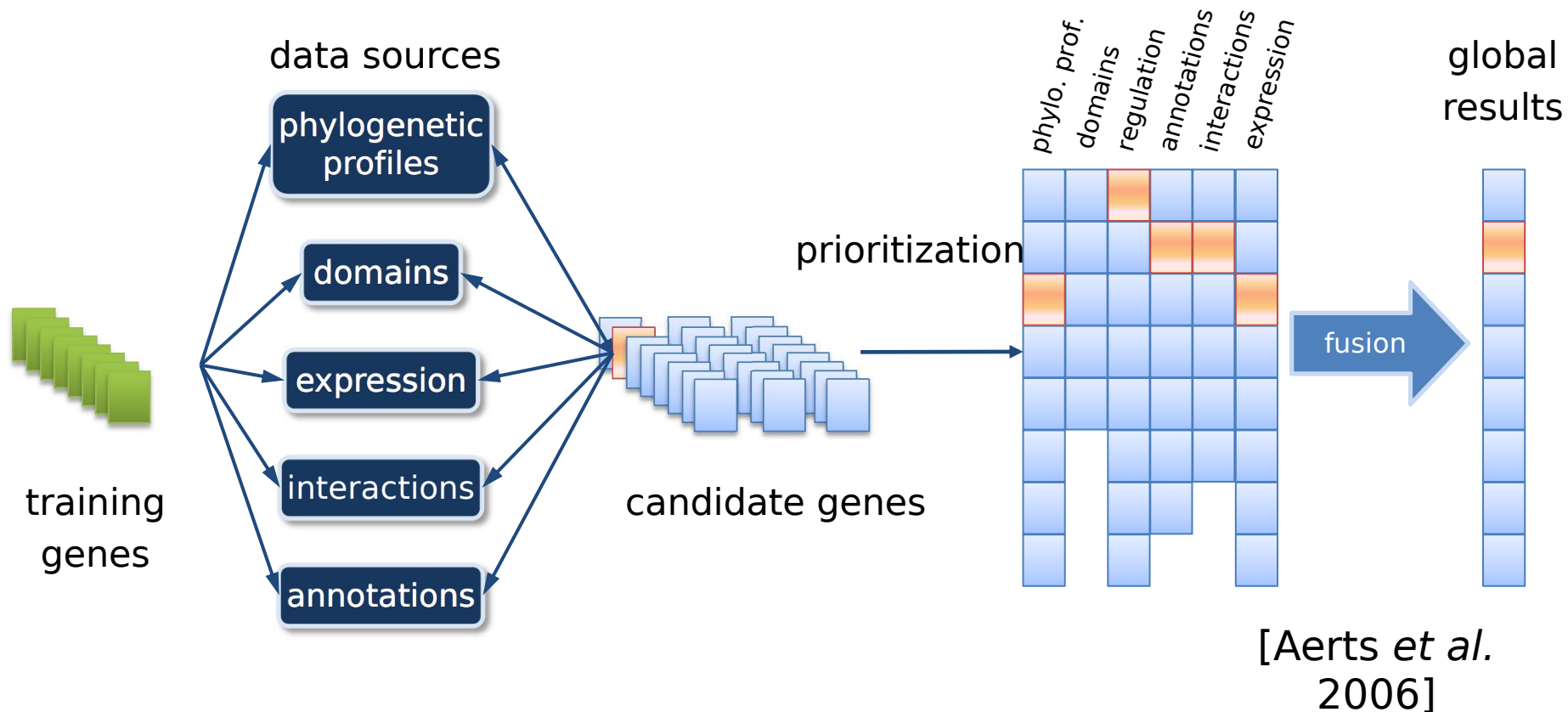# Intégration de données hétérogènes Priorisation

Master 2
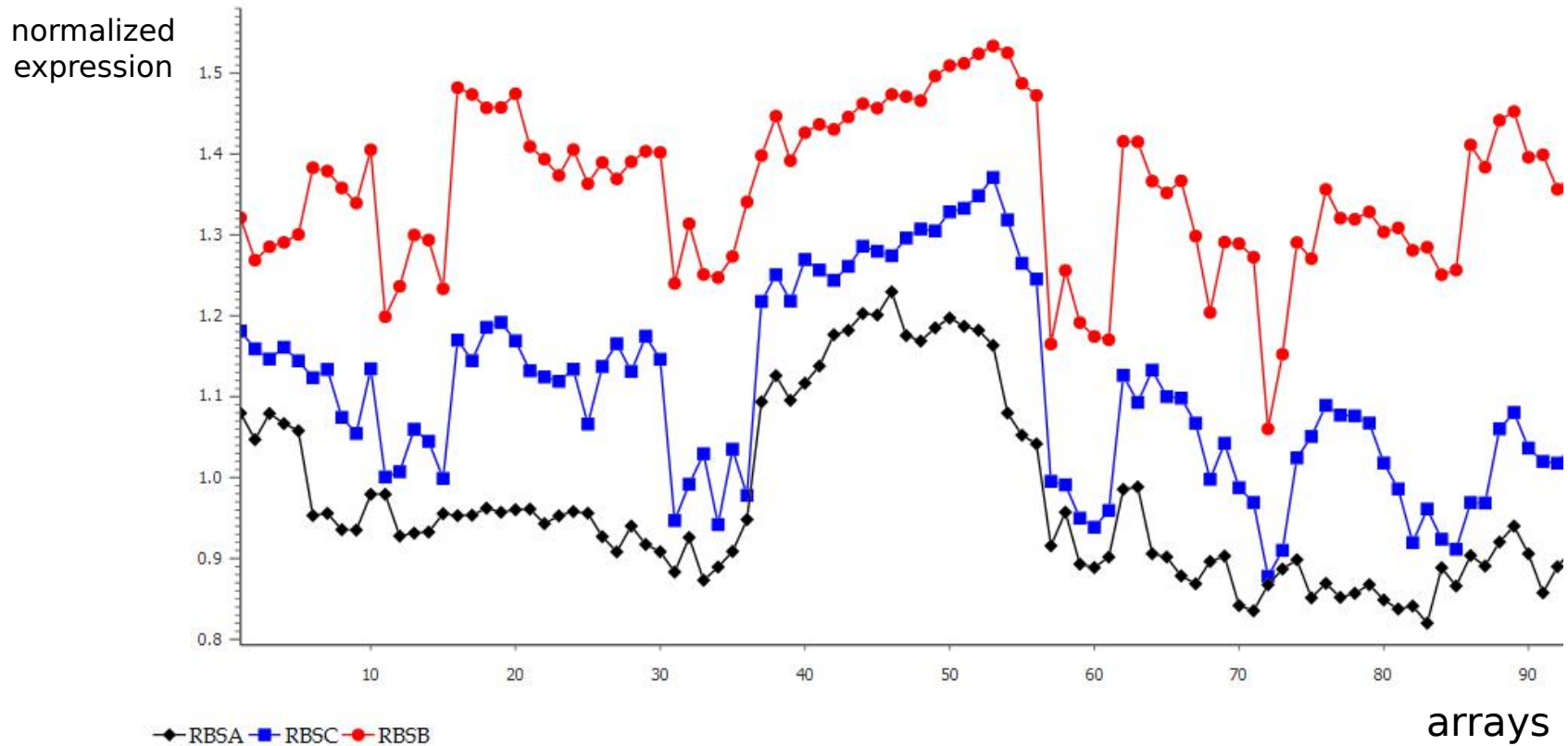
Bioinformatique et Biologie des Systèmes

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
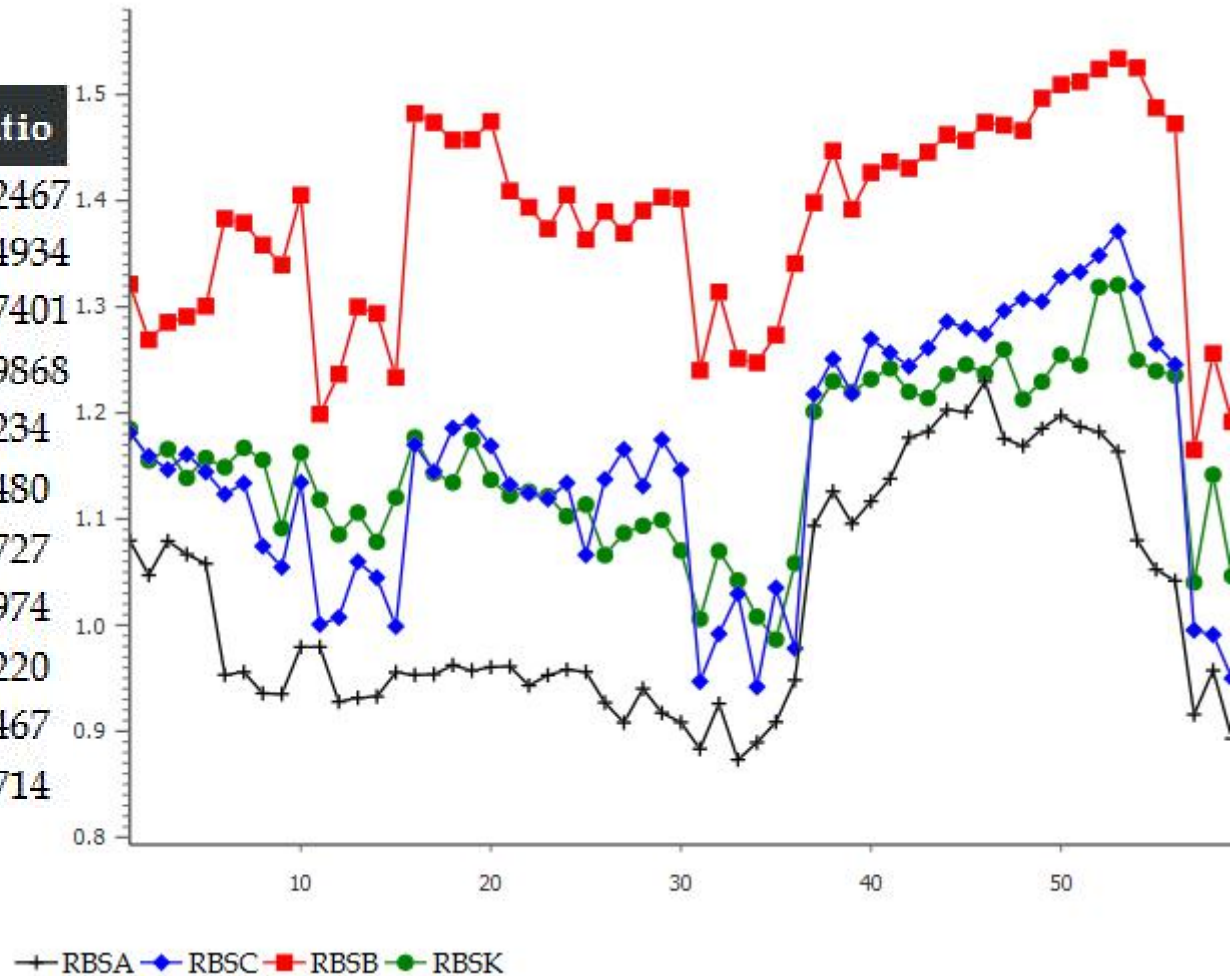

[Aerts *et al.* 2006]

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

normalized expression

arrays

- 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

RBSA   RBSC   RBSB   RBSK

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