Integration of functional genomics data

Laboratoire Bordelais de Recherche en Informatique (UMR)

Centre de Bioinformatique de Bordeaux (Plateforme)

Observations and motivations

- □ Genomics and functional genomics have expanded the focus of cellular biology from individual biomolecular entities towards relationships between those entities.
  - Entities : genes, ORFs, proteins…
  - Relationships : interactions, complexes, pathways, networks…
- □ This raises new types of questions and new requirements in terms of data integration.

How to make sense out of new experimental data ?

1. Purification of protein complexes.
   Is there a biological knowledge, or information, which significantly groups together the components of a complex ?
2. Large scale expression profile analysis.
   Are there clusters of co-regulated genes that significantly correspond to known biological processes ?

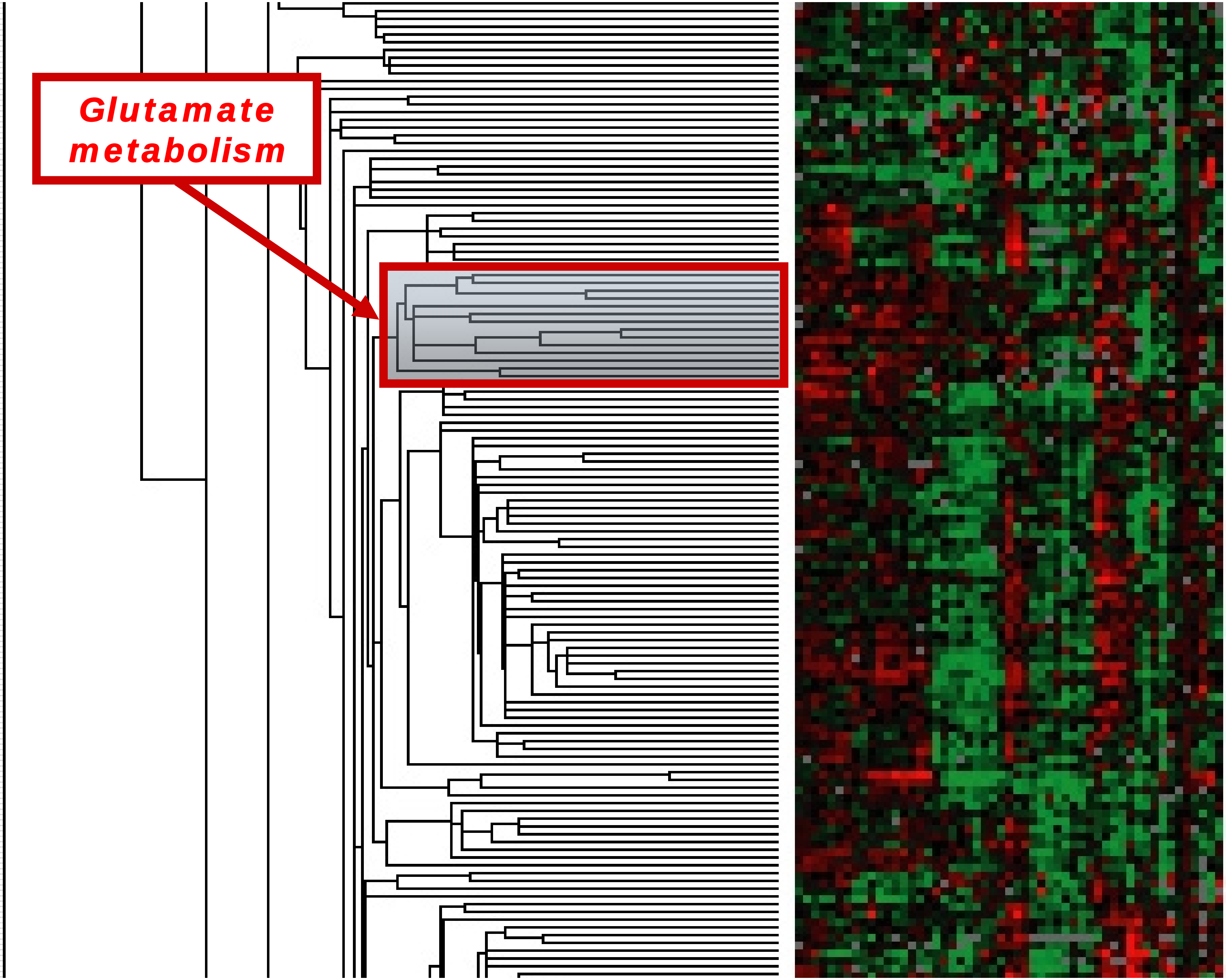
Characteristics of the “new” questions

- □ The questions start with a set of biomolecular entities (query set).
- □ The answer should go further than collecting information attached to all members of the query set (what <u>significantly</u> groups members of the set ?).

<u>Example</u> : Analysis of a 13 proteins yeast complex

2 Valine, leucine and isoleucine biosynthesis (16)
3 Biodegradation of Xenobiotics (137)
1 Glycolysis / Gluconeogenesis (47)
2 Oxidative phosphorylation (70)
5 Non-enzymes (4312)

KEGG Pathways


Glutamate metabolism

Difficulties related to biological information

- □ **Heterogeneity** : in terms of semantics (functional and structural information) and in terms of structures (numerical values, discrete attributes, natural language texts,…)

- □ **Dissemination** : annotated databases (UNIPROT, EMBL, KEGG,…), literature (MedLine, full text of articles), raw data sources (SMD, ArrayExpress,…).

How to identify a biological criteria which <u>significantly</u> groups components of my query set ?

Proposed strategy

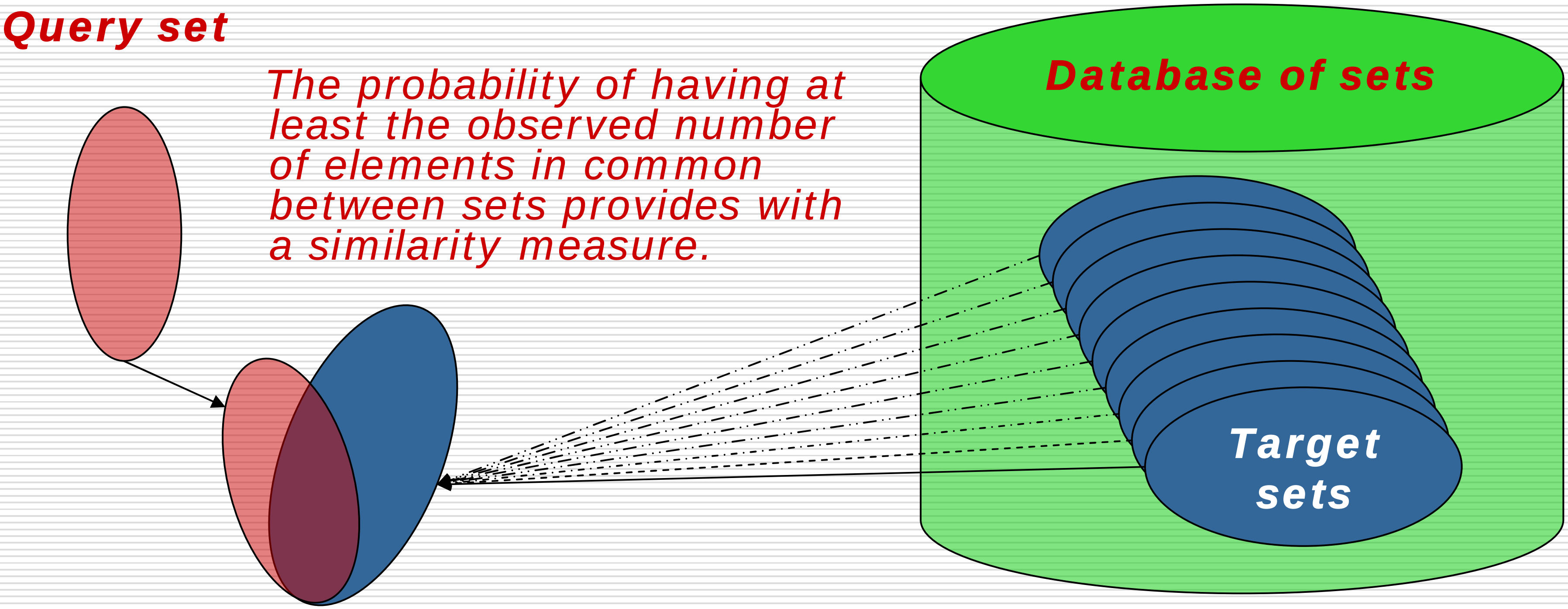
- Principles
  - Use sets of genes, or gene’s products, as a unified data structure
  - Convert as much as possible of available biological knowledge into sets (known / target sets)
  - Use a measure of similarity between sets in order to compare a query set with the target sets
- System
  - Store all the target sets in a database
  - Define a standard format to import new sets
  - Develop a system that supports queries: comparison of one or several sets against the content of the database in order to fetch similarities

Converting biological knowledge into sets

- Structural information (InterPro : 1 domain = 1 set)
- Functional classification (Kegg : 1 pathway = 1 set)
- Protein interactions (1 complex = 1 set)
- Cellular location (MIPS : 1 compartment = 1 set)
- Biliographical references (Pubmed : 1 article = 1 set)
- Expression data (GEO : 1 cluster = 1 set)
- Physico-chemical properties (a IP value range = 1 set)
- Genome structure (1 group of neighbors = 1 set)
- ...

Principles of sets comparison

The probability of having at least the observed number of elements in common between sets provides with a similarity measure.

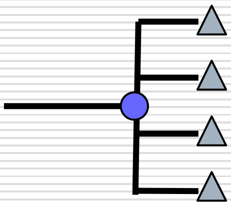


- Sets have to be taken from a define population (an organism).
- Due to multiple comparisons, statistical correction is necessary (i.e. Bonferonni) in order to compute an Evalue.

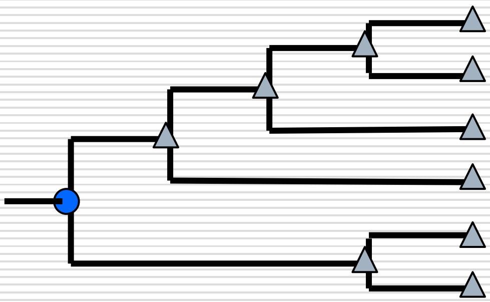
Organization of the sets

□ Each set belongs to a criteria (i.e. physical proximity, a given expression data experiment, GO, etc…)

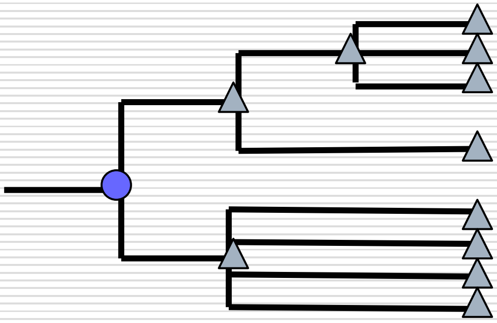
□ For a given criteria, there are relationships between sets that can be described in a graph



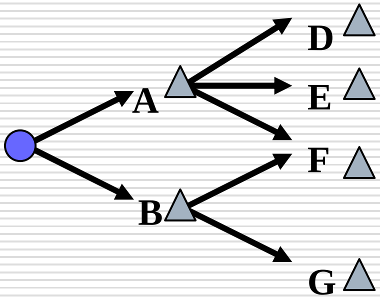
Star graph
(domains, biblio)



Binary tree
(Hierarchical clustering)

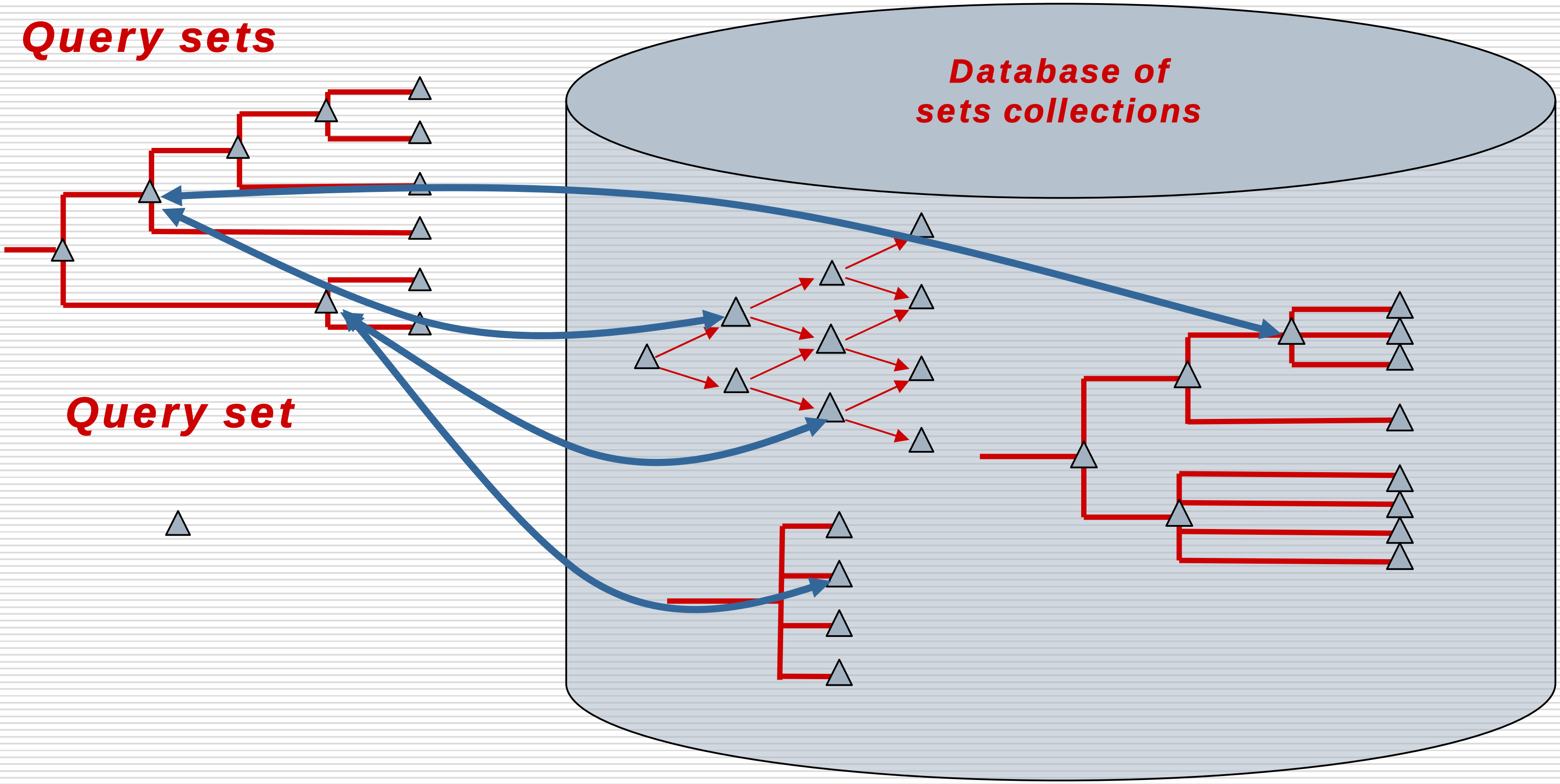


Tree
(Enzyme)



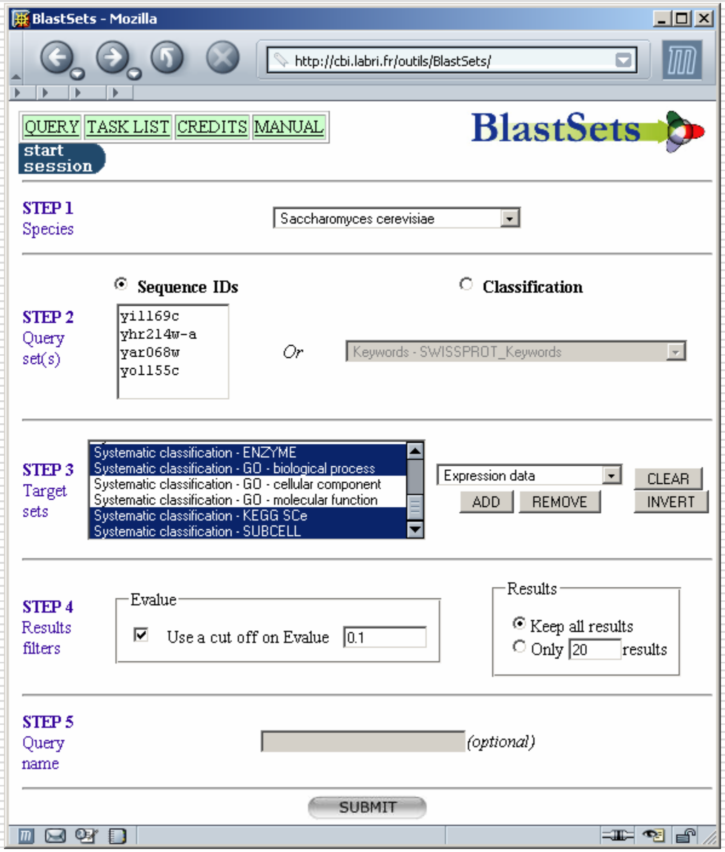
Directed acyclic Graph
(Go, physical proximity)

Single/Multiple query sets



BlastSets system

System up and running and publicly available at http://cbi.labri.fr/outils/BlastSets/



STEP 1 Species: Saccharomyces cerevisiae

STEP 2 Query set(s): ● Sequence IDs ○ Classification

```
yil169c
yhr214w-a
yar068w
yol155c
```

Or Keywords - SWISSPROT_Keywords

STEP 3 Target sets:
- Systematic classification - ENZYME
- Systematic classification - GO - biological process
- Systematic classification - GO - cellular component
- Systematic classification - GO - molecular function
- Systematic classification - KEGG SCe
- Systematic classification - SUBCELL

Expression data

STEP 4 Results filters:
Evalue: ☑ Use a cut off on Evalue 0.1
Results: ● Keep all results ○ Only 20 results

STEP 5 Query name: (optional)

Barriot, R., Poix, J., Groppi, A., Barre, A., Goffard, N., Sherman, D., Dutour, I. & de Daruvar, A. New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res*.

Can BlastSets be usefull ?

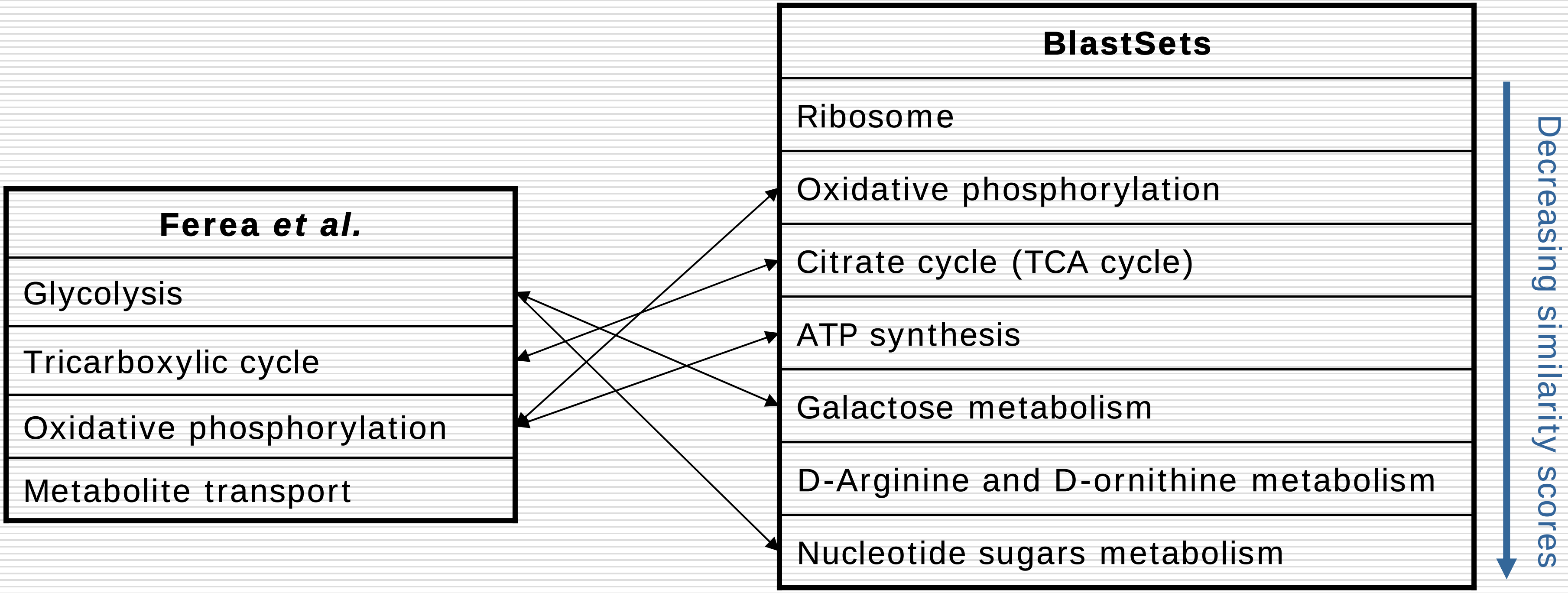
ex: Large scale expression profile analysis.

Are there clusters of co-regulated genes that significantly correspond to known biological processes ?

Interpretation of expression data

- □ Compute an automatic comparison of :
  - sets obtained by hierarchical clustering of real expression data
  - sets corresponding to metabolic pathways
- □ Compare BlastSets results (pathways that are found most significantly similar to a given node in the hierarchical tree) and published results (obtain by manual exploration of the hierarchical tree)

Results



Ferea *et al.*

Glycolysis

Tricarboxylic cycle

Oxidative phosphorylation

Metabolite transport

BlastSets

Ribosome

Oxidative phosphorylation

Citrate cycle (TCA cycle)

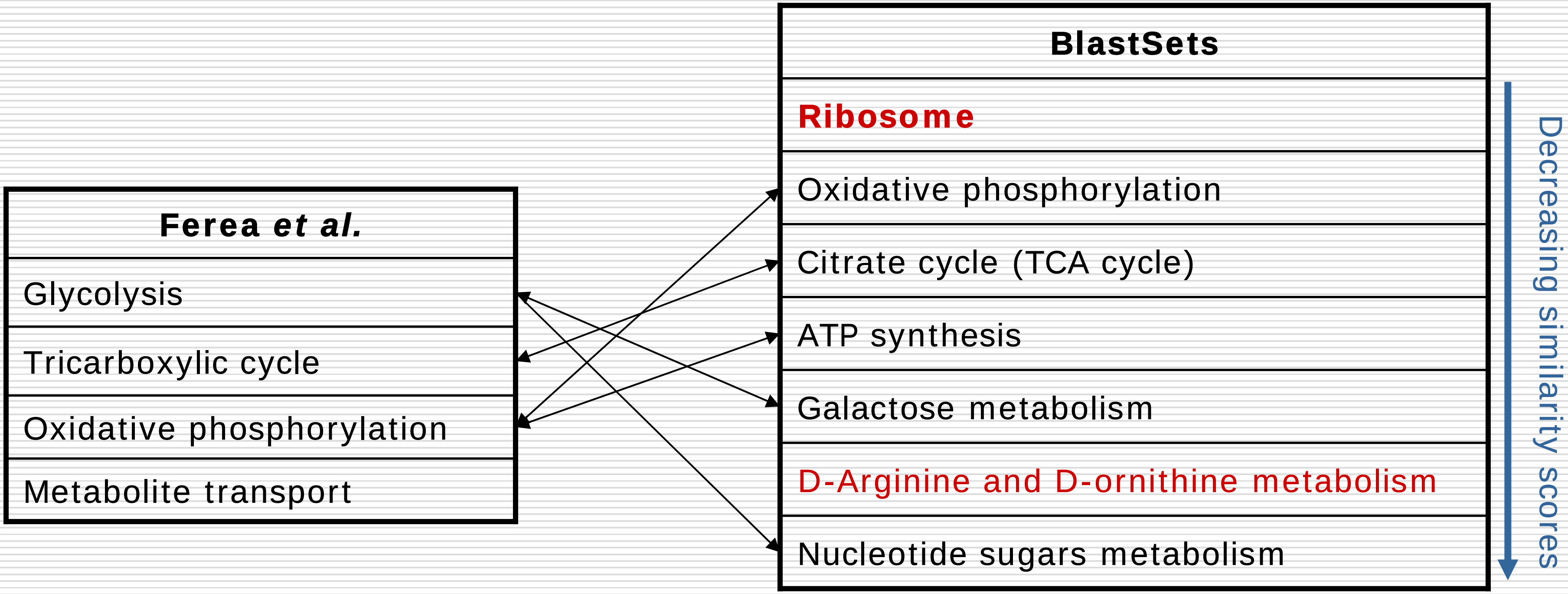
ATP synthesis

Galactose metabolism

D-Arginine and D-ornithine metabolism

Nucleotide sugars metabolism

Results

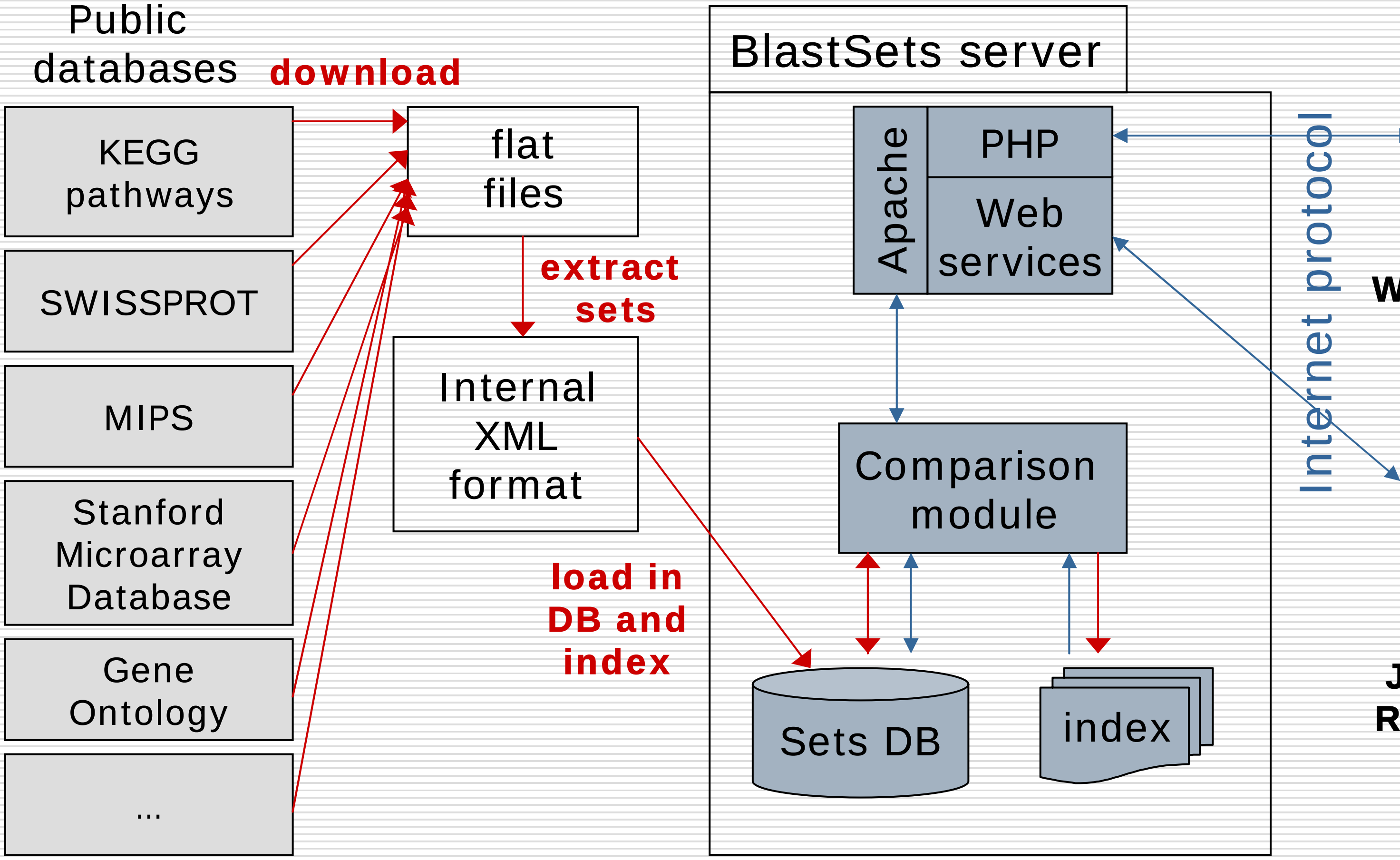


Ferea *et al.*
Glycolysis
Tricarboxylic cycle
Oxidative phosphorylation
Metabolite transport

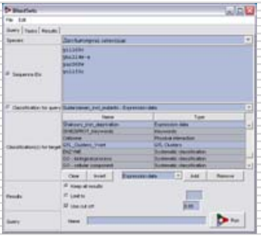
BlastSets
Ribosome
Oxidative phosphorylation
Citrate cycle (TCA cycle)
ATP synthesis
Galactose metabolism
D-Arginine and D-ornithine metabolism
Nucleotide sugars metabolism

Decreasing similarity scores

BlastSets architecture



Web browser



Java, Perl or Ruby program

Knowledge representation : how to define sets?

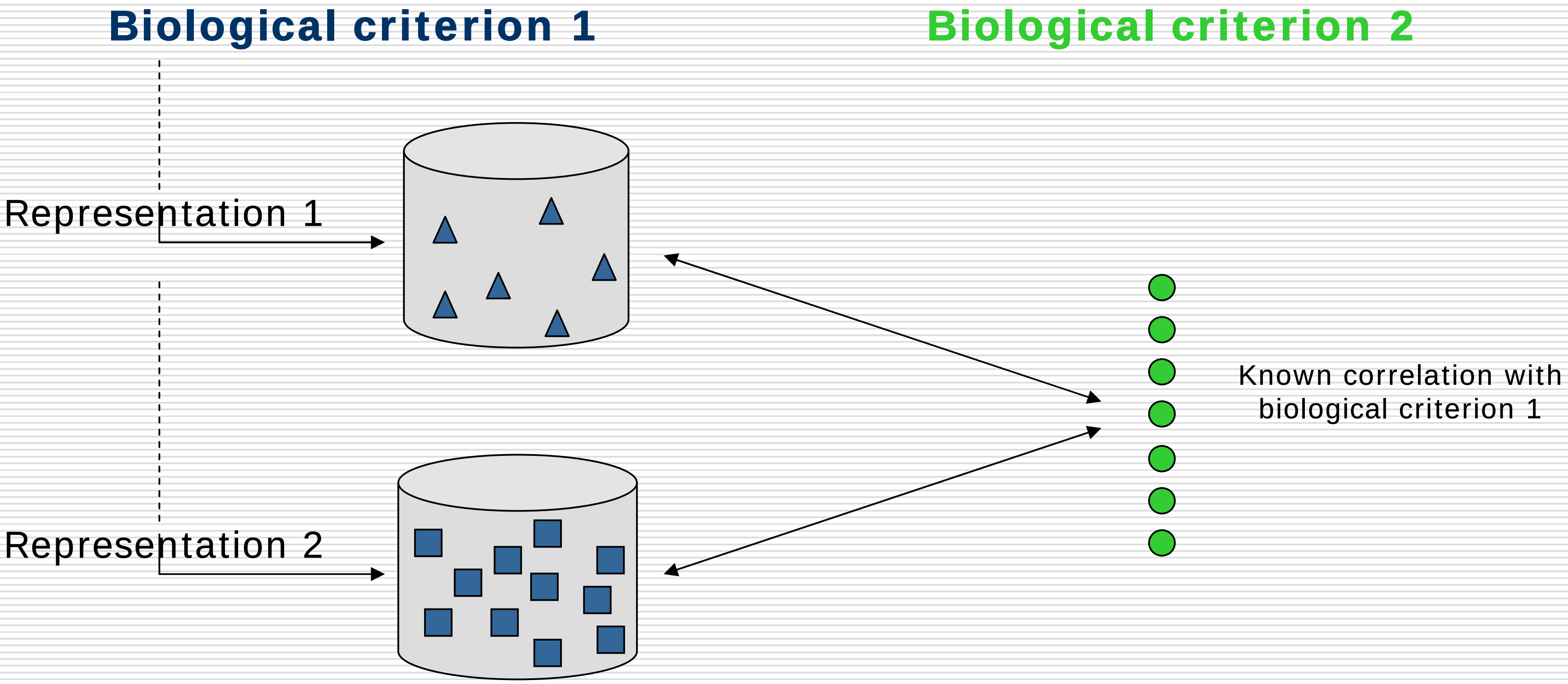
- Simple for discrete attributes:
  - Sub cellular compartments

    one compartment = one set
  - Metabolic pathways

    one pathway = one set
  - Multi-protein complexes

    one complex = one set
- Not simple otherwise... how to choose the most appropriate clustering method ?

Comparing different representations



Biological criterion 1

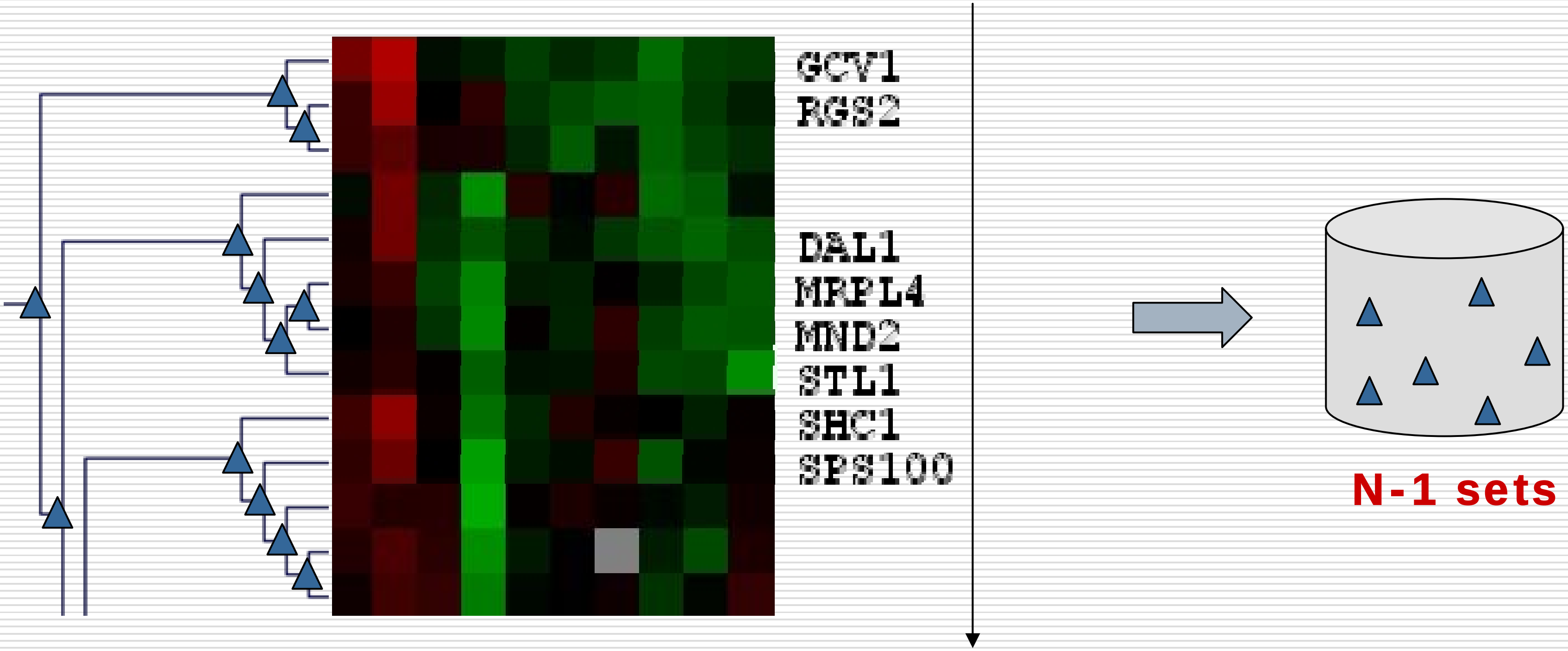
Biological criterion 2

Representation 1

Representation 2

Known correlation with biological criterion 1

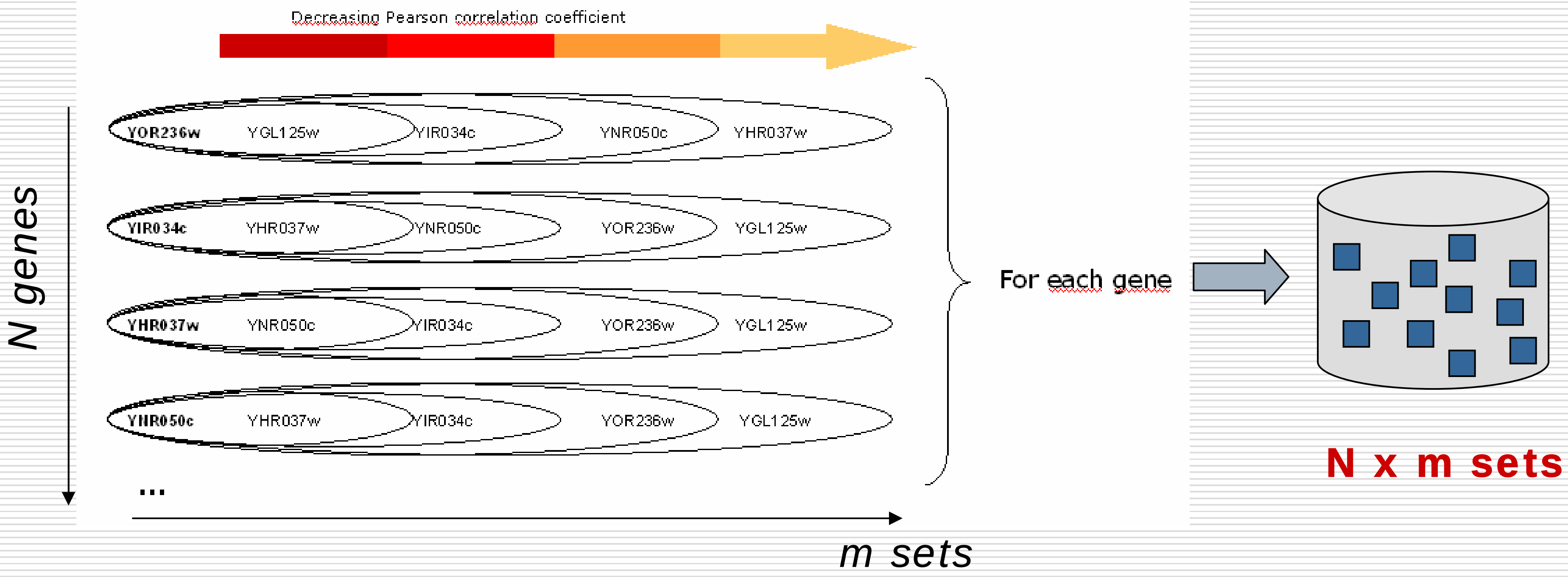
Clustering expression profiles : Hierarchical clustering



N-1 sets

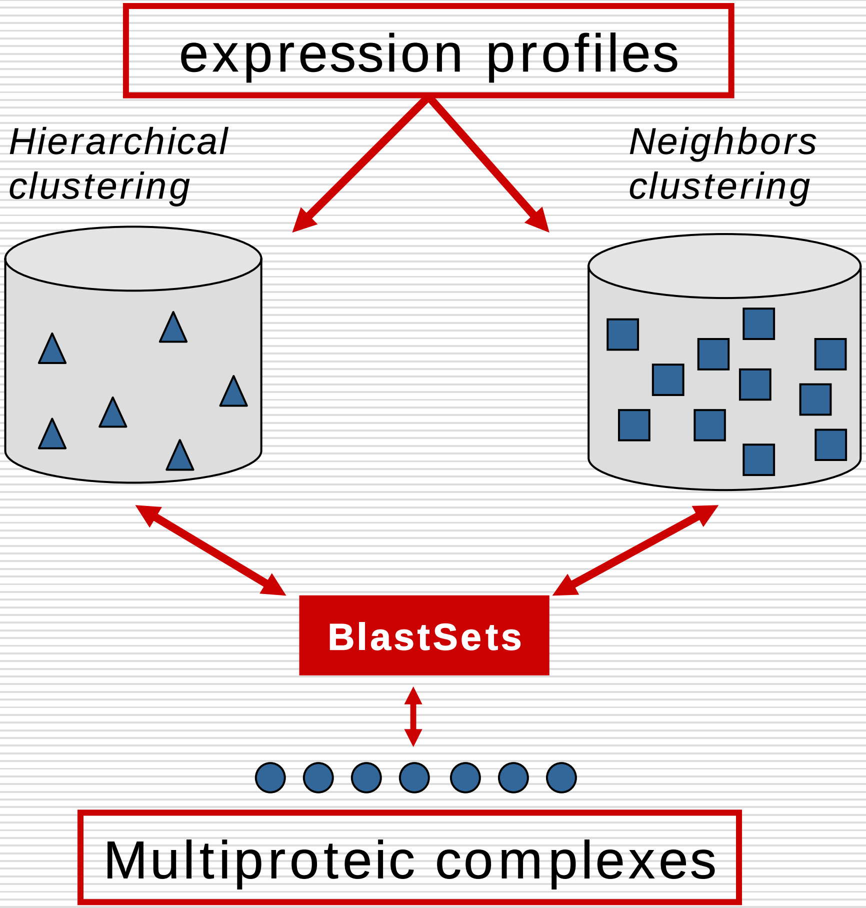
N genes

Clustering expression profiles : Best neighbors



More groups => more information captured… and more noise!

Assessment procedure using protein complexes



expression profiles

Hierarchical clustering

Neighbors clustering

BlastSets

Multiproteic complexes

Complexes	Hierarchical clustering	Neighbors clustering
Comp. 1	X	X
Comp. 2		
Comp. 3	X	X
Comp. 4		X
…		
Total	2	3

Results : nb. complexes similar to at least one expression cluster

	Spellman experiment[2]			Gasch experiment[3]		
	Hierarchical clustering	Neighborhood 60	Neighborhood 100	Hierarchical clustering	Neighborhood 60	Neighborhood 100
Number of sets	5629	56 300	78 820	5648	56 490	79 086
MIPS Complexes[1] (1059)	48	51	14	56	89	20
Random complexes (1059)	0	0	0	0	0	0

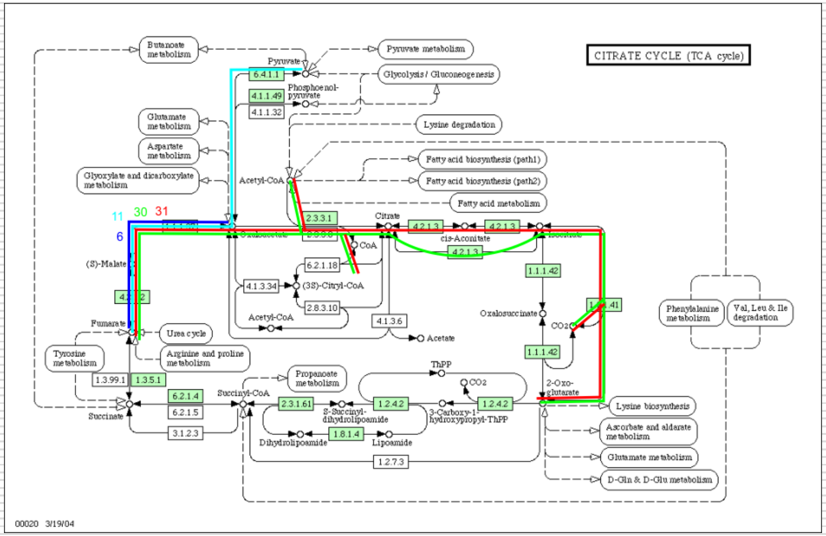
Obtained using Bonferroni correction

1. MIPS Database – Complex : http://mips.gsf.de/genre/proj/yeast/
2. Spellman PT et al. 1998. "Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization". *Mol Biol Cell* 9(12) : 3273-97
3. Gasch AP et al. 2000. "Genomic expression programs in the response of yeast cells to environmental changes". *Mol Biol Cell* 11(12) : 4241-57

Project on pathways (collaboration with the KEGG)

Assessment of various methods for representing metabolic pathways :

- One KEGG map = one set
- For each map : calculation of elementary modes each of which defines a set



Conclusion / perspectives

- BlastSets implements the concept of neighborhoods (A. Danchin) in order to reveal potential relationships between heterogeneous information.
- The strategy requires optimization of knowledge representation.
- Some computational problems remain to be solved.
- Can the method be implemented as a service provides by the each data source ?

Partners

ACI IMPBIO

- **Centre de Bioinformatique Bordeaux (A. de Daruvar, A. Groppi, A. Barré)**
- **Laboratoire Bordelais de Recherche en Informatique (A. de Daruvar, I. Dutour, D. Sherman, R. Barriot, C. Gaugain)**
- **Laboratoire de Statistique Mathématique et Applications (J. Poix)**
- **Unité de Génétique des Génomes Bactériens, Institut Pasteur (A. Danchin)**
- **UMR – INRA/UB2 Génomique Développement Pouvoir Pathogène (A. Blanchard)**

Other collaborations :

- **INIST (A. Zasadzinski)**
- **KEGG (M. Kanehisa, J.M. Schwartz, J. Nacher)**