Ca'Foscari University of Venice
Department of environmental sciences, informatics and statistics

Bioinformatics Day @ DAIS

# Comparison of Metabolic Networks: a two-level approach

Erboso Gianluca & Meggiato Alberto

Supervisors: Prof. Cocco Nicoletta & Simeoni Marta

07/07/2016

# Outline

Framework

State of the Art

KEGG database

The proposed method

Similarity indexes

Tool

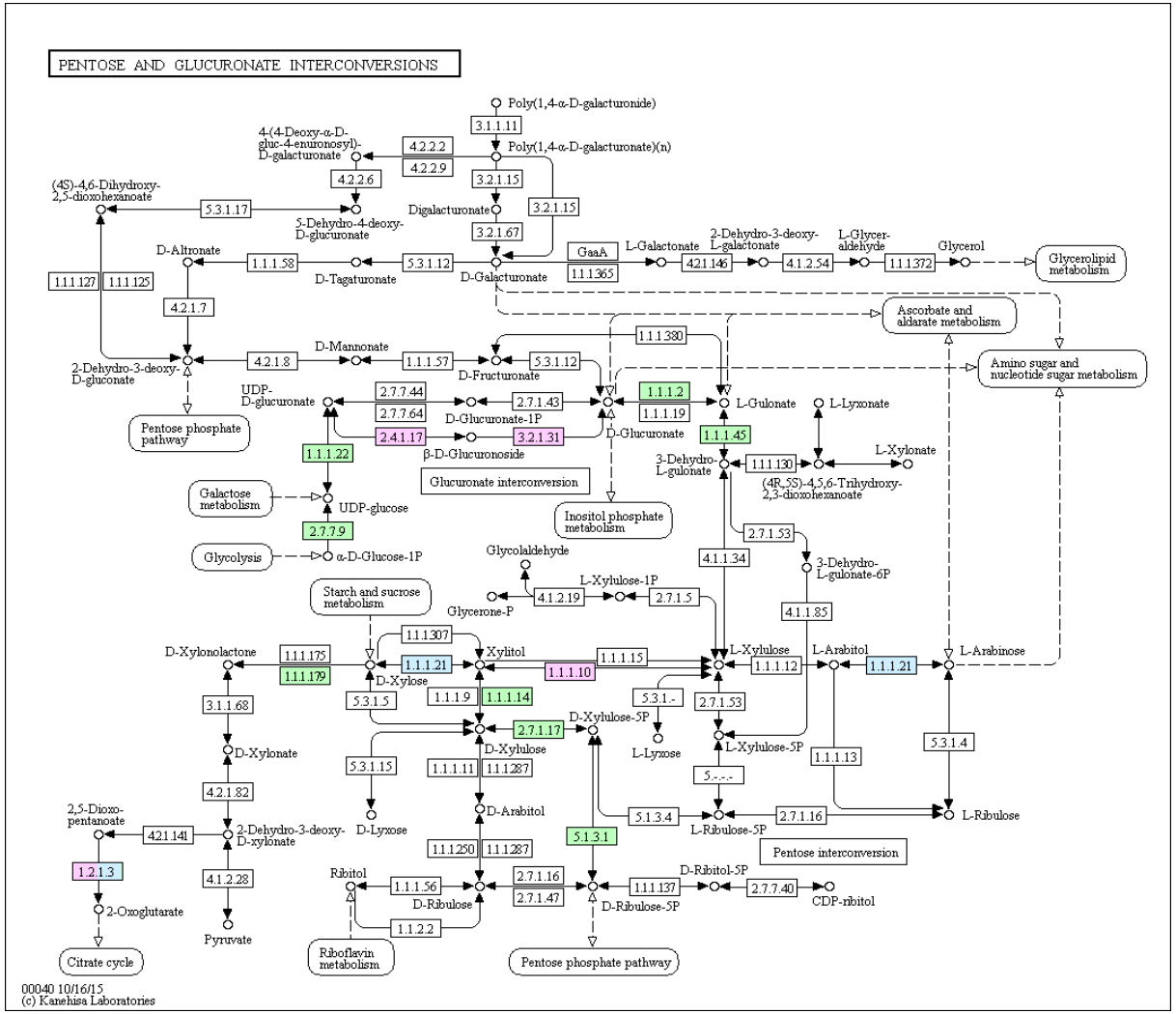
Experiments & Results

Conclusions

# Framework: metabolism

**Metabolism** [1, 2] is the network of all chemical and physical reactions that take place within the cells of the organisms.

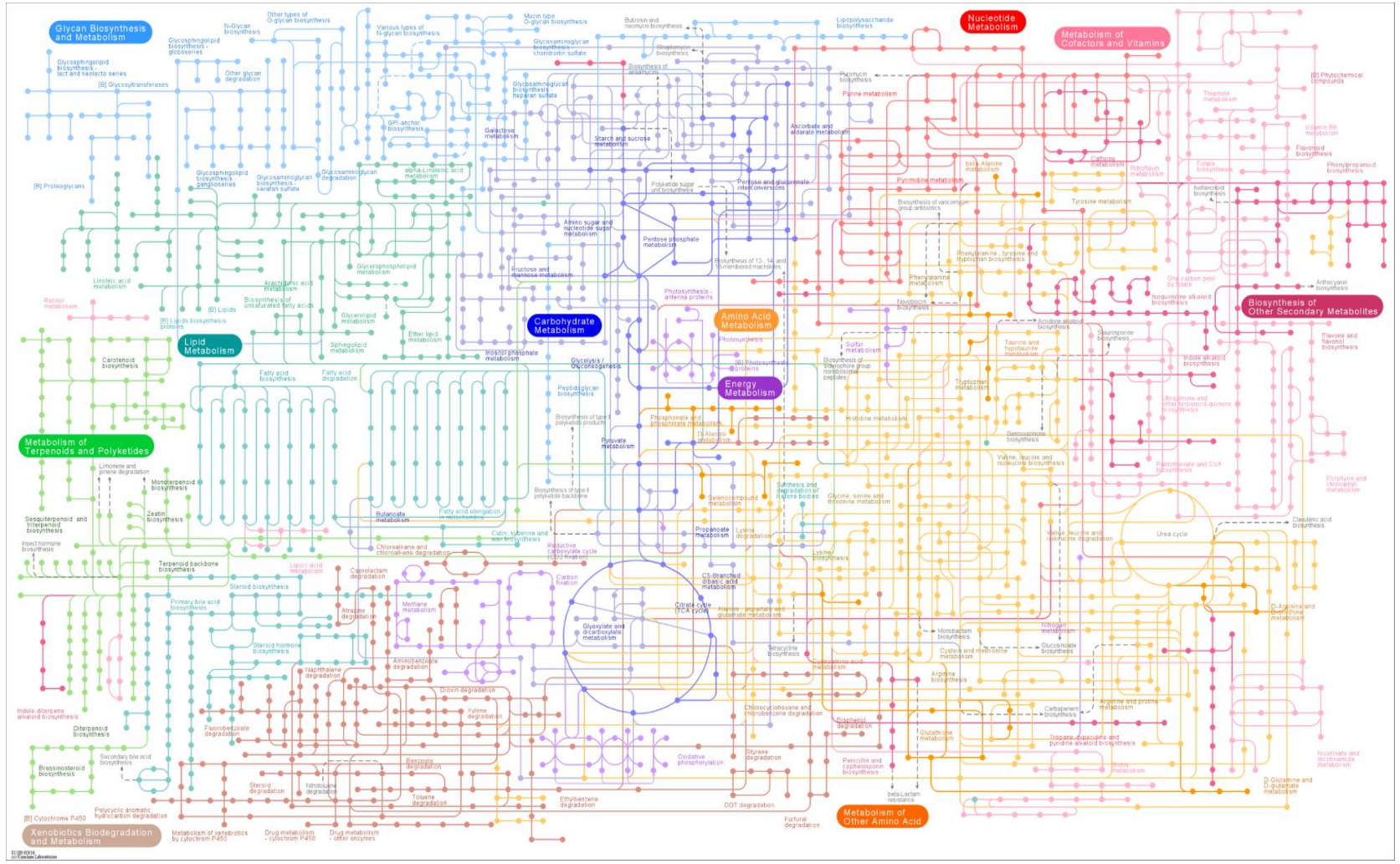
**Metabolic Pathways** [3] are a sequence of reactions such that the product of a single reaction can be used as reagent for another one.



# Framework: metabolism

The **Metabolic Network** [4] represents the complete set of metabolic functions and their interections that determine the structure and properties of the cells.

» Simplified version of a metabolic network.



# Framework: motivations

**Comparison of metabolic networks** is relevant for studying the evolutionary process, discovering drug targets and more in general for supporting medical science activities.

**Troubles:**

- In biology, the comparison of metabolic networks is really complex
- Graph based modeling system represents graphs of huge dimensions
- Graph matching is NP-Hard

Comparison of metabolic networks as well as of metabolic pathways is challenging from a computational point of view.

**Aim:** propose a new comparison method that consider the entire metabolic networks while avoiding the computational problems.

# State of the Art

The existing methods make use of different data structures keeping different level of detail:

- Sets (multisets)
- Sequences (Reactions profile)
- Graphs (including hypergraphs and Petri Nets)

Drawback: each of these approaches present a **computational problem** that is related to the **complexity of the data structure**

**Metabolisim Databases** (most popular)

- KEGG (Kyoto Encyclopedia of Genes and Genomes)
- BioCyc
- SEED
- EcoCyc (E. coli Database)
- SGD (Saccharomyces Genome Database)

# KEGG Database

It is one of the most important **collections of biological data**, containing information of different organisms on:

- metabolic pathways,
- genomic,
- chemical,
- health (i.e. human diseases).

Main advantages:

✓ 4290 cataloged organisms (Eukaryotes: 333, Bacteria: 3729, Archaea: 228)

✓ Standardized representation of the data

✓ Good modularization

✓ Integration of graphical and textual information

✓ Freely available and costantly updated

# KEGG: metabolic pathways

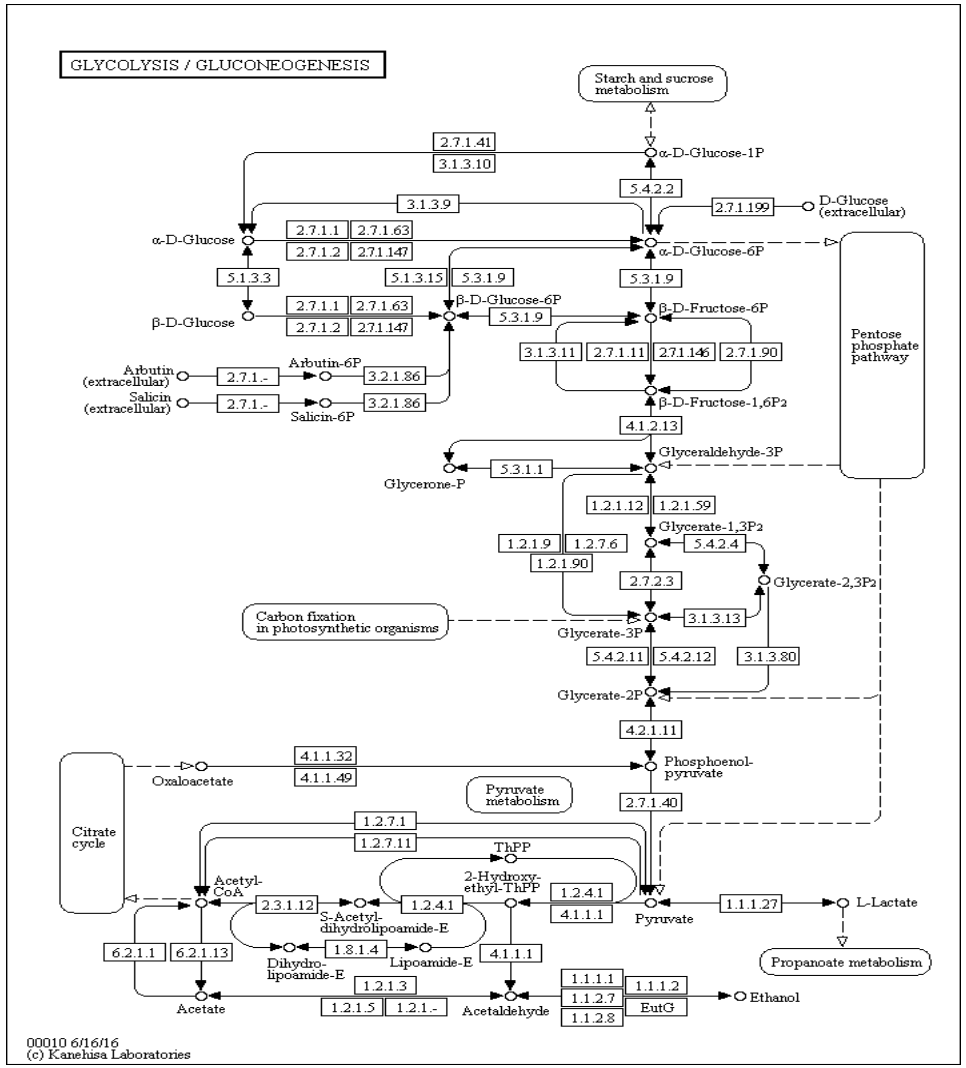
KEGG associates to each metabolic function, a unique **reference pathway** which corresponds to the union of the corresponding pathway in different organisms. (**unique modularization**)

Data representation:

- graphical (**pathway map** ) ⟶ all the KEGG knowledge of a metabolic pathway
- textual (**KGML file**) ⟶ the organism-specific info for the corresponding pathway map

**Aim:**

- use the KGML files for metabolic pathway comparison
- exploit the KEGG API for data retrieval



# KEGG: metabolic pathways

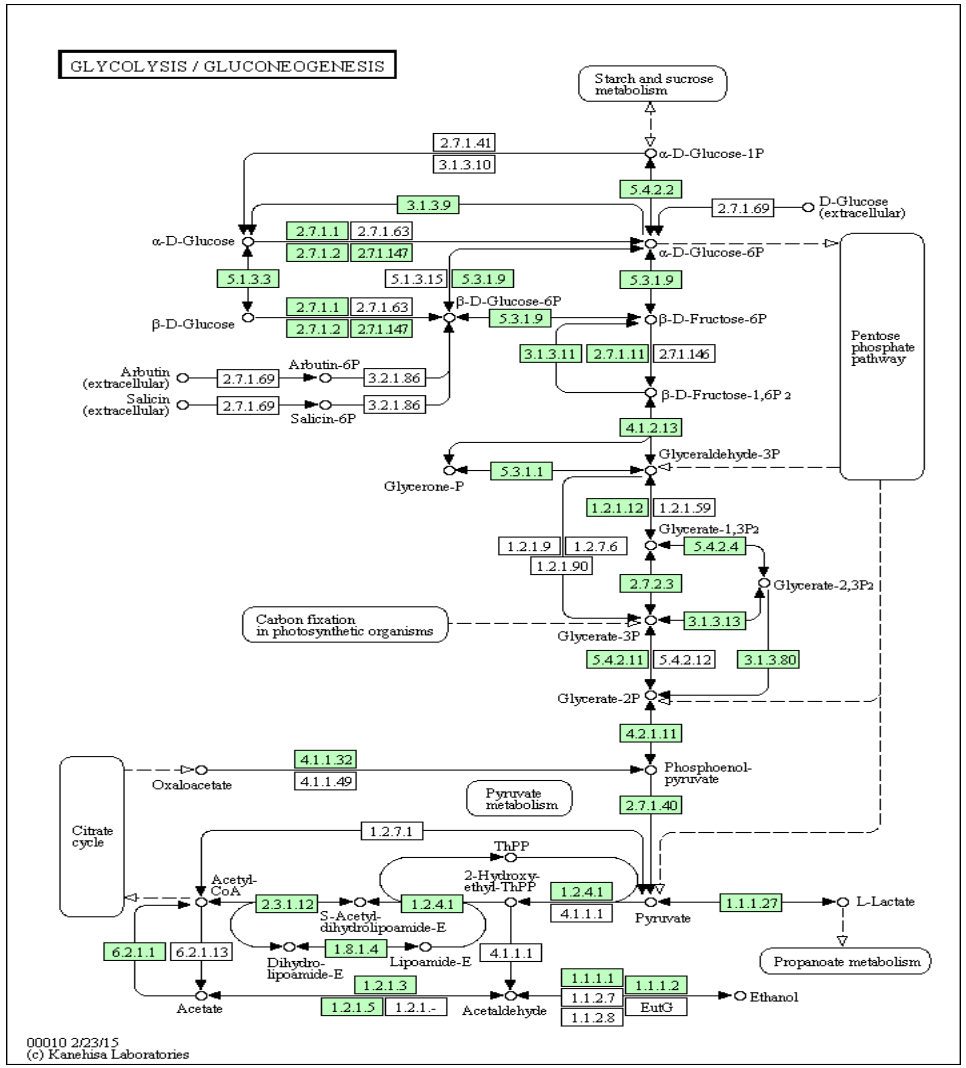
KEGG associates to each metabolic function, a unique **reference pathway** which corresponds to the union of the corresponding pathway in different organisms.

Data representation:

- graphical (**pathway map** ) ⟶ all the KEGG knowledge of a metabolic pathway
- textual (**KGML file**) ⟶ the organism-specific info for the corresponding pathway map

**Aim:**

- use the KGML files for metabolic pathway comparison
- exploit the KEGG API for data retrieval

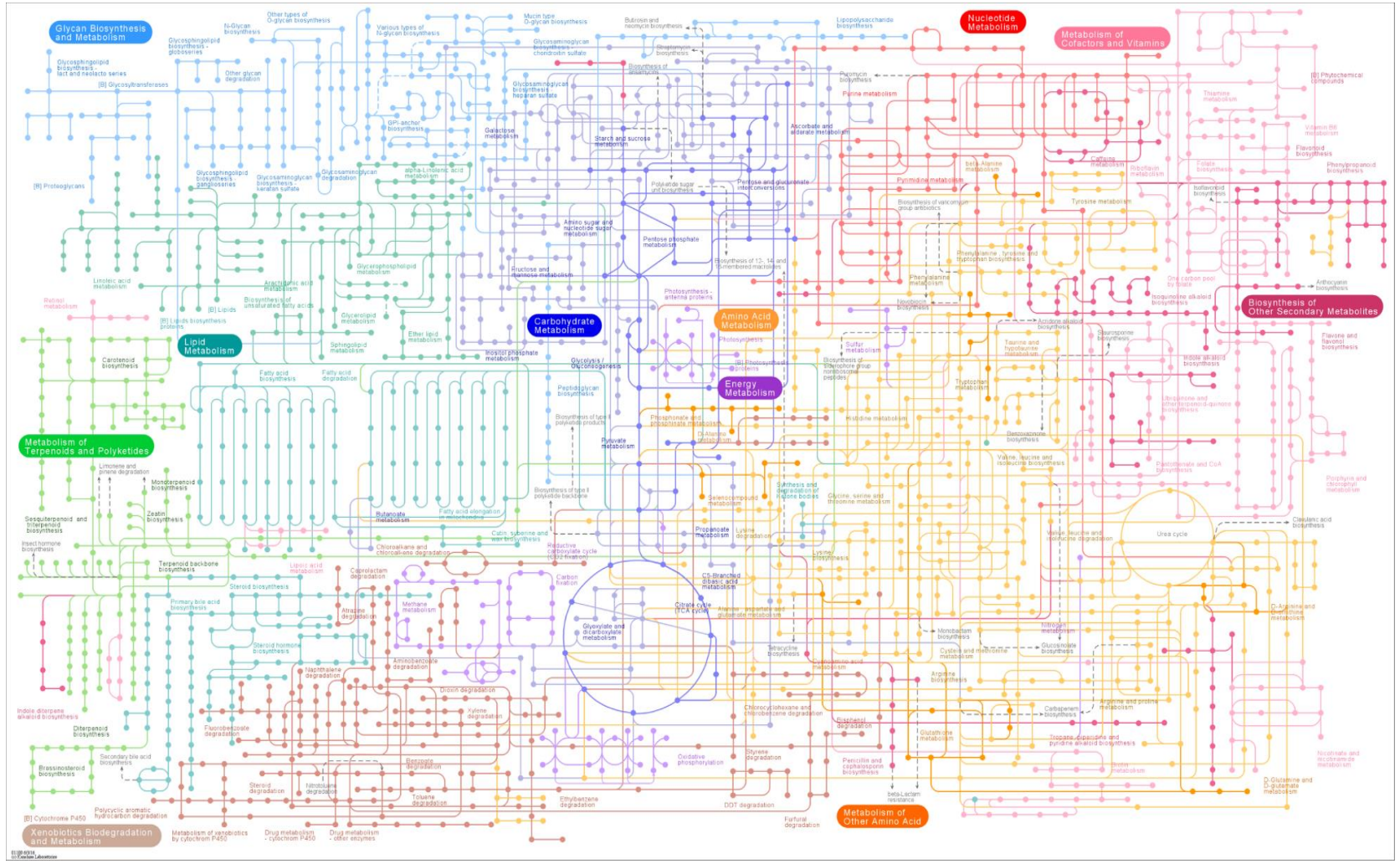


# KEGG: metabolic network

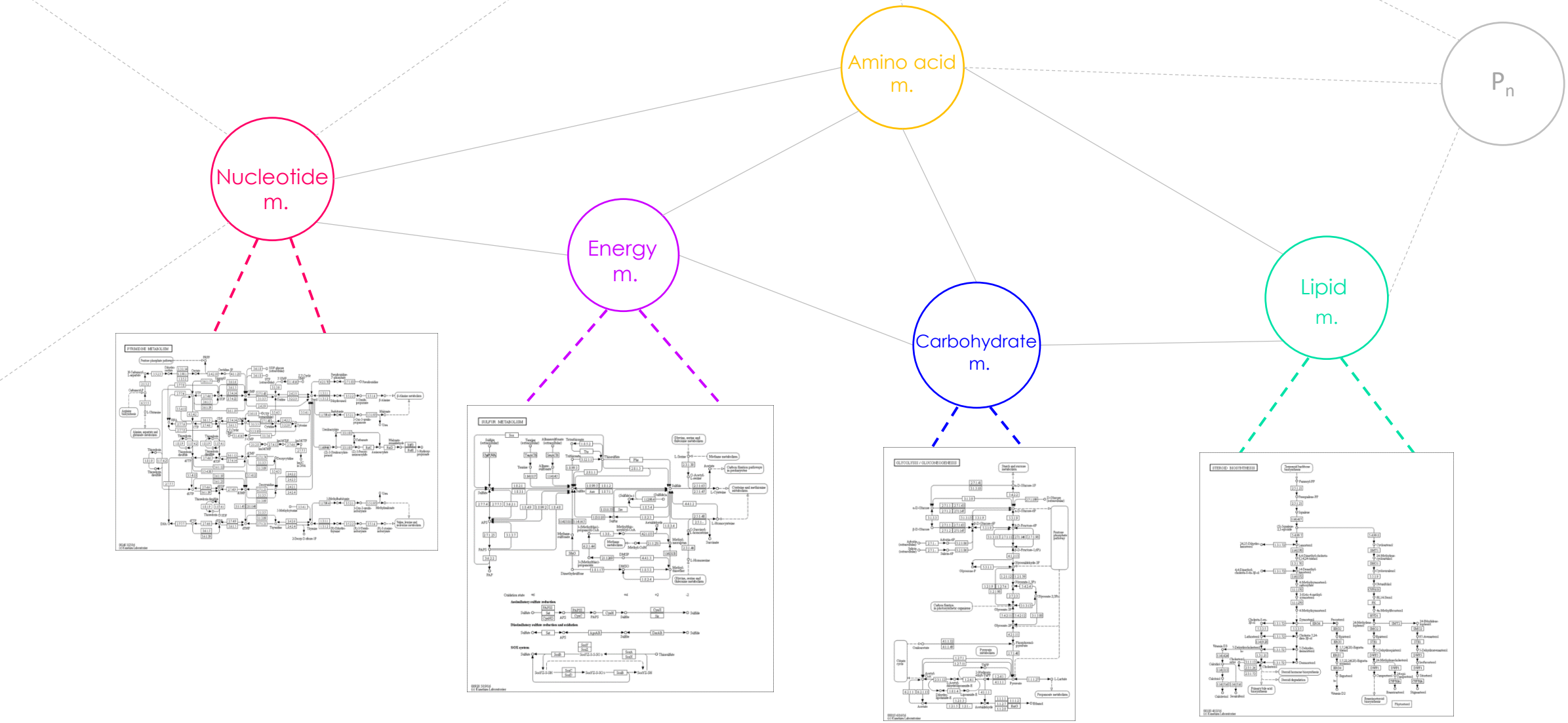
**Reference metabolism:**

- Union of the reference pathways
- Implicit subdivision of the metabolism

» Exploiting the **standardized modularization** of the pathways given by KEGG we are able to reconstruct the metabolic network



# The idea

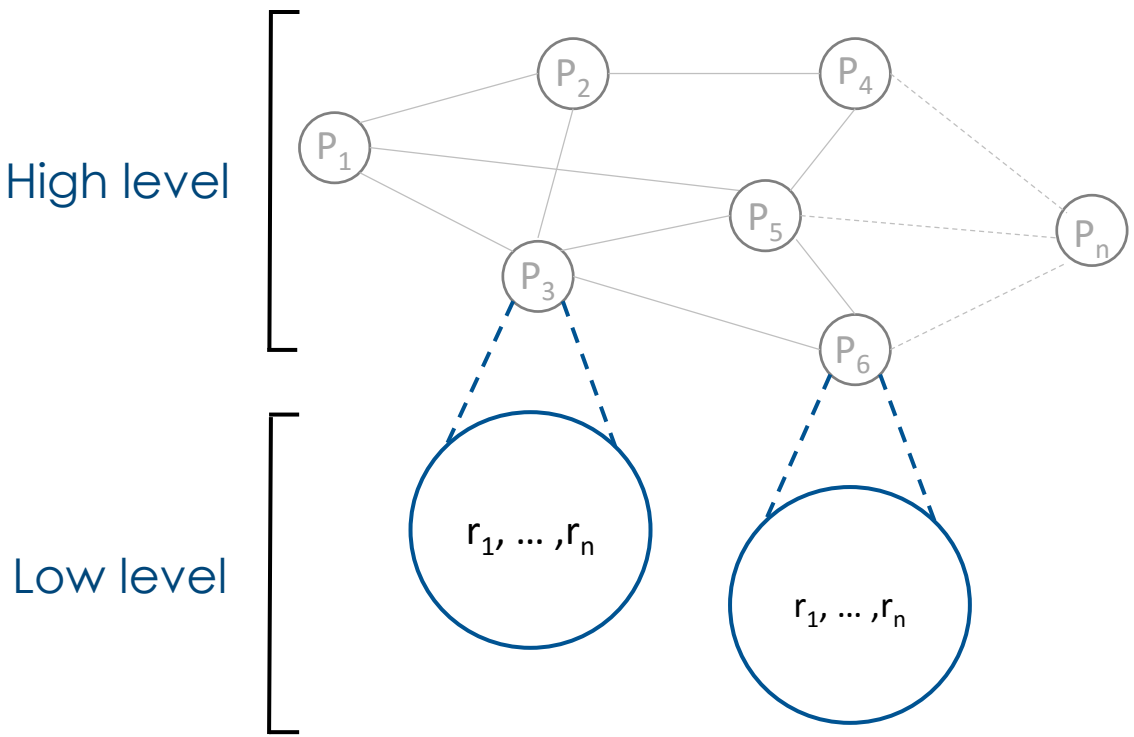


# The proposed method

We propose a new comparison method based on a **two-level approach** providing an abstraction of a metabolic networks**.**

- **High level**: the net is modeled as a ***graph:*** nodes represent metabolic pathways and arcs the relations between pathways themselves;

- **Low level**: each metabolic pathway is modeled as ***set or multiset of chemical reactions***.




Indipendent levels allows for computing different similarity indexes (topology and functionality) that can be combined later.

# Metabolic network reconstruction

Organism's **metabolism reconstruction** →

Consider each pathway (KGML) that belongs to the metabolism of a specific organism. Through a **parsing** of these files we extract the information useful for the metabolism reconstruction

```
<pathway name="path:hsa00010" org="hsa" number="00010"
        title="Glycolysis / Gluconeogenesis"
        image="http://www.kegg.jp/kegg/pathway/hsa/hsa00010.png"
        link="http://www.kegg.jp/kegg-bin/show_pathway?hsa00010">
   ...
   <entry id="41" name="path:hsa00030" type="map"
      link="http://www.kegg.jp/dbget-bin/www_bget?hsa00030">
      <graphics name="Pentose phosphate pathway" fgcolor="#000000" bgcolor="#FFFFFF"
          type="roundrectangle" x="656" y="339" width="62" height="237"/>
   </entry>
   <entry id="56" name="hsa:2597 hsa:26330" type="gene" reaction="rn:R01061"
      link="http://www.kegg.jp/dbget-bin/www_bget?hsa:2597+hsa:26330">
      <graphics name="GAPDH, G3PD, GAPD, HEL-S-162eP..." fgcolor="#000000" bgcolor="#BFFFBF"
          type="rectangle" x="458" y="484" width="46" height="17"/>
   </entry>
   <relation entry1="41" entry2="56" type="maplink">
      <subtype name="compound" value="130"/>
   </relation>
```

During the parsing phase we consider:

- The current metabolic function → **node**
- Relation tag of type='maplink' → **edge**

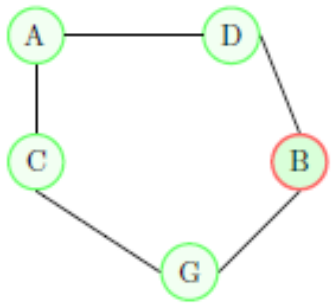
↓

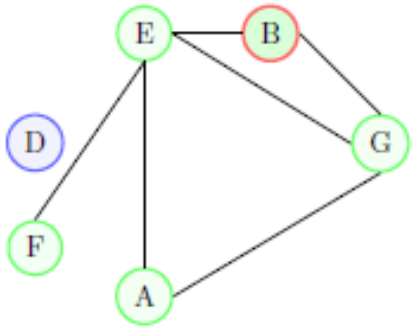
<relation **entry1**="41" **entry2**="56" type="maplink">

# Data structure

Graph of metabolism → **modified adjacecy matrix** of fixed size → Implicit mapping

Let us consider the metabolic networks of two organisms, O and O' . The set of metabolic pathways is represented by {A; B; C; D; E; F; G; H}.




$$\begin{array}{c|cccccccc} & A & B & C & D & E & F & G & H \\ \hline A & 1 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ B & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\ C & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ D & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ E & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ F & 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ G & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{array}$$

$$\begin{array}{c|cccccccc} & A & B & C & D & E & F & G & H \\ \hline A & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ B & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ C & 0 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ E & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ F & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ G & 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ H & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{array}$$

The diagonal represents the state of the node:

- 1 **connected** pathway
- 0 **isolated** pathway
- -1 pathway **not present**

The other values in the matrix represent the edges

# Pathway Similarity Index

It considers the union of the metabolic pathways of the two organisms.

$$SimP_i = \begin{cases} 0 & \text{if } P_i \text{ is missing in } O \text{ or } P_i' \text{ is missing in } O' \\ 1 & \text{if } P_i \text{ is present in } O \text{ and } P_i' \text{ in } O' \text{ but there are no reactions to compare} \\ \frac{|R_i \cap R_i'|}{|R_i \cup R_i'|} & \text{otherwise} \end{cases}$$

- *$O$ and $O'$: the two organisms,*
- *$P_i$ and $P_i'$ : the corresponding metabolic pathway,*
- *$R_i$ and $R_i'$ : the reactions of $P_i$ and $P_i'$ in $O$ and $O'$.*

The similarity measure depends on the metabolic pathway representation. In our case since we use sets, the definitions are based on **Jaccard index**.

# Functional Similarity Indexes

The **functional similarity index** is the mean similarity over the union of the pathways of $O$ and $O'$.

$$SimPA = \frac{\sum_{i=1}^{n} SimP_i}{n}$$

The **weighted functional similarity index** is the weighted mean similarity wrt. the number of reactions of the pathways in $O$ and $O'$.

$$SimPW = \frac{\sum_{i=1}^{n} SimP_i * |R_i \cup R'_i|}{\sum_{i=1}^{n} |R_i \cup R'_i|}$$

where $n = |M|$ and $M$ is the union of the metabolic pathways of both $O$ and $O'$.

The $SimPW$ index provides a refined measure since it balances the values wrt. the number of common reactions.

**» The two indexes can be used in the definition of the Separated Similarity Index.**

# Structural similarity indexes

Let us consider two organisms O and O' and their corresponding graphs of metabolic network, G=(V,E) and G'=(V',E'). Let us consider the i-th pathway, $P_i \in V$ and $P_i' \in V'$. Let $E_i$ and $E_i'$ be the sets of edges that connect $P_i$ and $P_i$', respectively, with other nodes. Let deg(v) (deg(v')) the degree of the vertex v ∈ V (v′ ∈ V′).

The **structural similarity index** wrt. the i-th pathway, $\boldsymbol{SimS_i}$, is defined as:

$$\boldsymbol{SimS_i} = \begin{cases} 0, & if\, P_i\ or\, P_i'\ is\ not\ present \\ 1, & if\, P_i\ and\ P_i'\ are\ both\ isolated \\ \dfrac{1}{1\ + \deg(P_i)}, & if\ only\ P_i'\ is\ isolated \\ \dfrac{1}{1 + \deg(P_i')}, & if\, only\ P_i\ is\ isolated \\ \dfrac{|E_i \cap E_i'|}{|E_i \cup E_i'|}, & if\ P_i\ and\ P_i'\ are\ both\ connected \end{cases}$$

The **structural network similarity index** is define as:

$$SimS = \sum_{i=1}^{n} \frac{SimS_i}{n}$$

where $n = |V \cup V'|$

# Global similarity indexes

The **global similarity indexes** compare two metabolic networks considering both the similarity of their structure and the similarity of the corresponding functions.

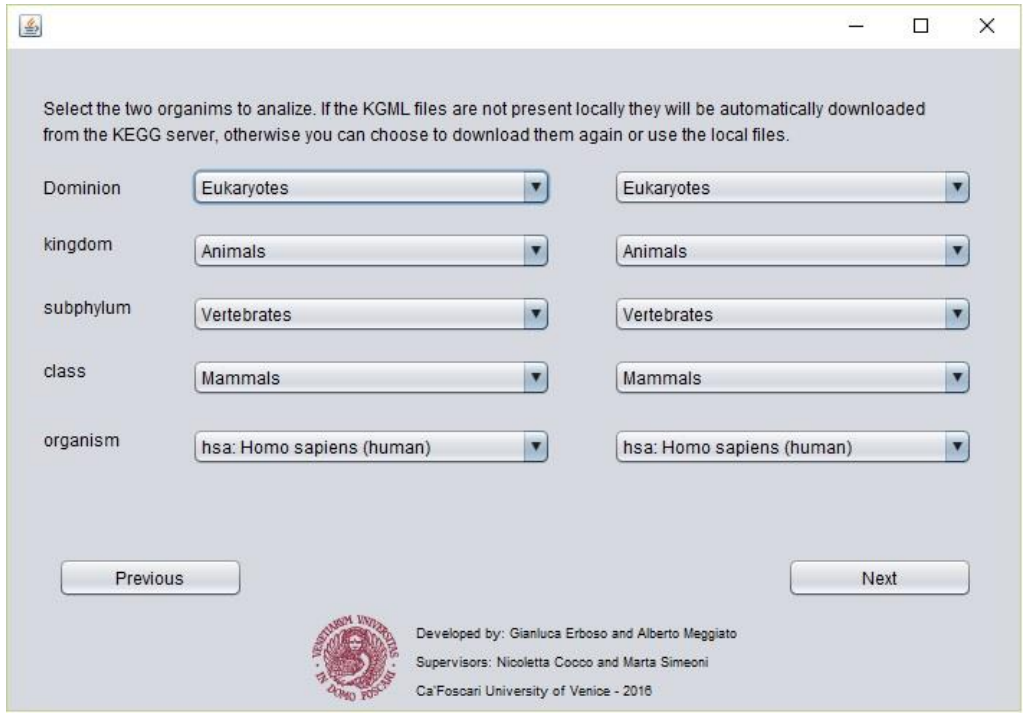
The **combined similarity index** is defined as follows:

$$CI = \frac{\sum_{i=1}^{n} SimS_i * SimP_i}{n}$$

The **separated similarity index** is defined as:

$$SI = \alpha * SimS + (1 - \alpha) * SimPW$$

# The tool



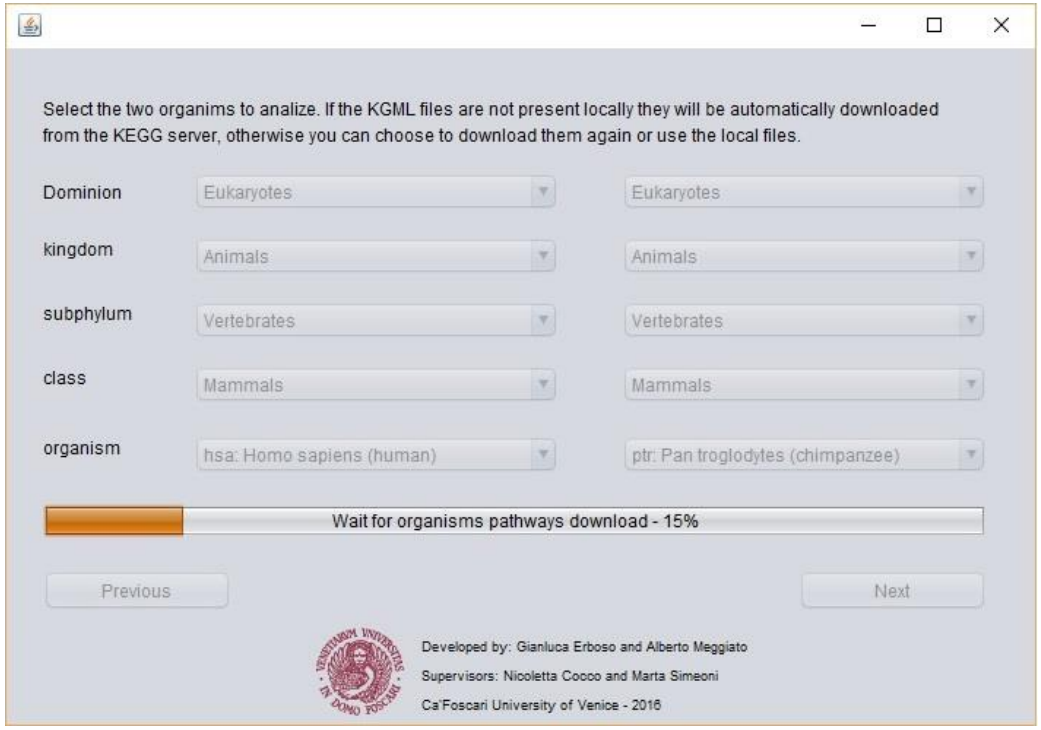

## Functionalities:

- Selection of two different organisms from KEGG database
- Selection of the comparison methods (at high and low level)
- Computation of different similarity indexes
- Management of KGML files
- Automatic exportation of the results as .xls file

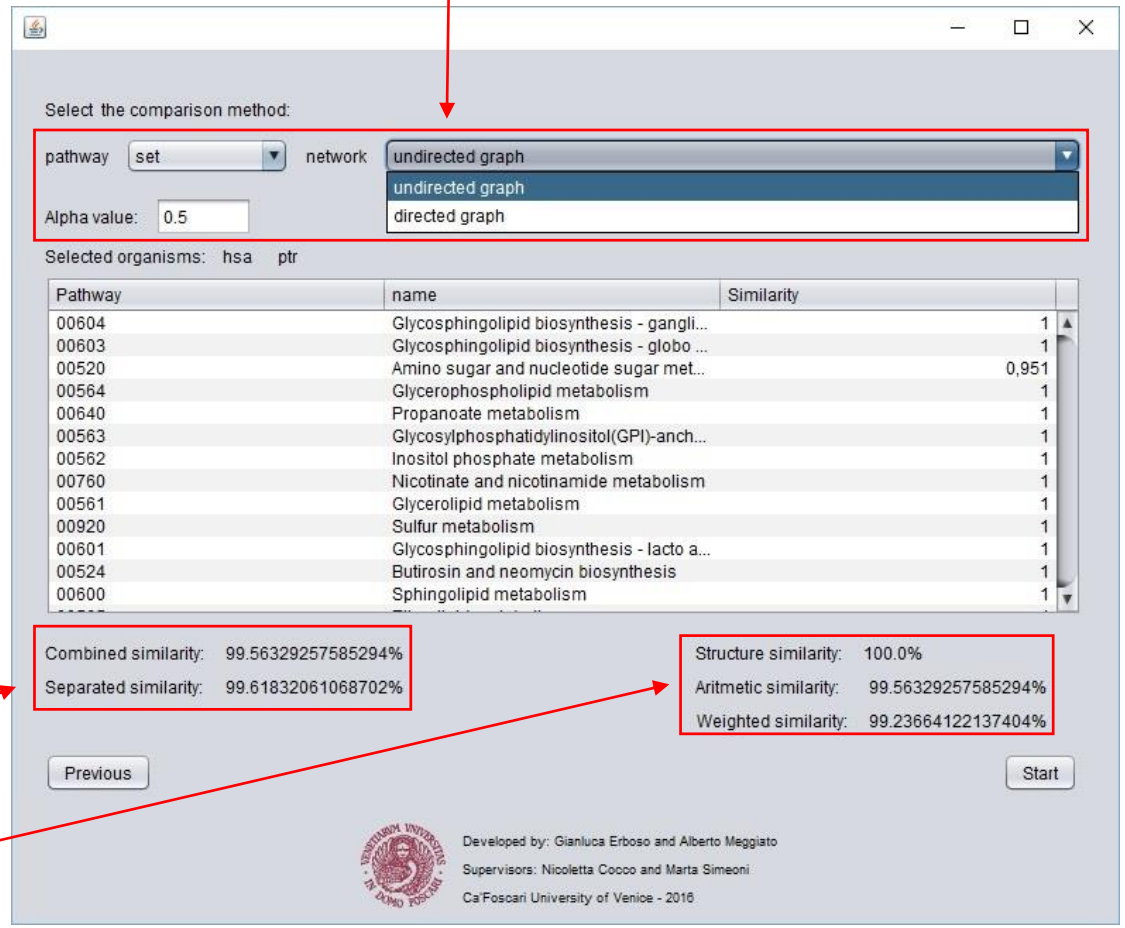
## Strengths:

- ✓ **Portable** across different platforms (Java Technology)
- ✓ Use of **multi-threading** techniques to parallelize the computation
- ✓ **Fast comparison** thank to the abstraction of the metabolic networks (30 ~ 90s)
- ✓ **Offline use** (KGML required)
- ✓ **Good modularization** thanks to MVC pattern
- ✓ Ready for further development

# The tool




Selection of the comparison methods.




Global similarity indexes

High and low level similarities

# Experiment 1: Sulfur metabolism

**Pathway**: Sulfur Metabolism

**Aim**: Test the classification of our method analyzing a set of organisms that take sulfur in different ways.

**Sim. Index**: $SimP_i$

- Animals take sulfur indirectly from proteins that they assume through their diet;
- Plants, Fungi and Bacteria are able to perform sulfur reduction producing sulfide, the simplest form of sulfur useful for amino acids construction.

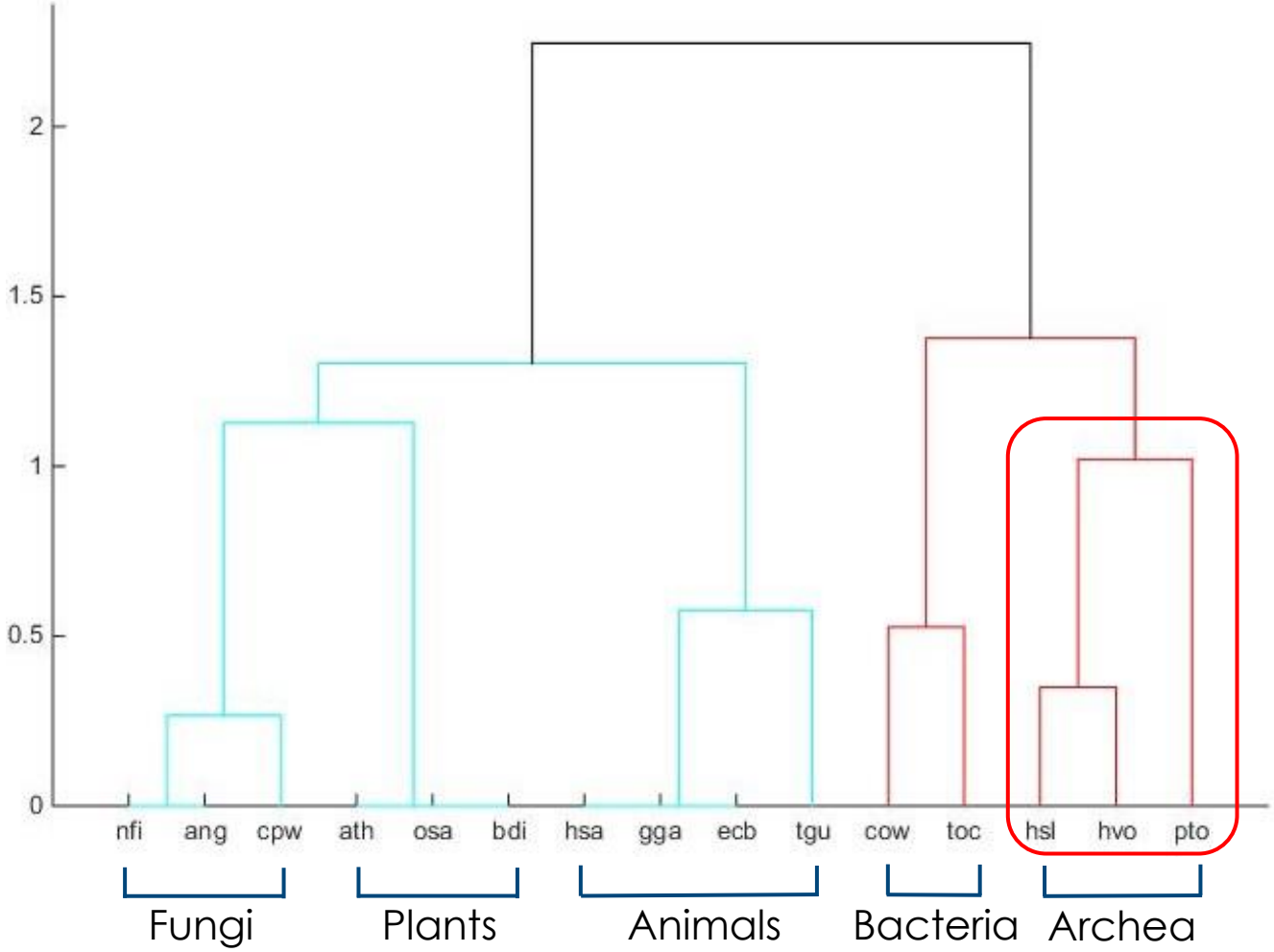
| Code | Organism | Kingdom | Taxonomic group |
|---|---|---|---|
| *hsa* | *Homo sapiens* (human) | Animals | Mammals |
| *ecb* | *Equus caballus* (horse) | Animals | Mammals |
| *gga* | *Gallus gallus* (chicken) | Animals | Birds |
| *tgu* | *Taeniopygia guttata* (zebra finch) | Animals | Birds |
| *ath* | *Arabidopsis thaliana* (thale cress) | Plants | Mustard family |
| *osa* | *Oryza sativa japonica* (Japanese rice) | Plants | Grass family |
| *bdi* | *Brachypodium distachyon* | Plants | Grass family |
| *nfi* | *Aspergillus fischeri* | Fungi | Eurotiomycetes |
| *ang* | *Aspergillus niger* | Fungi | Eurotiomycetes |
| *cpw* | *Coccidioides posadasii* | Fungi | Eurotiomycetes |
| *cow* | *Caldicellulosiruptor owensensis* | Bacteria | Caldicellulosiruptor |
| *toc* | *Thermosediminibacter oceani* | Bacteria | Thermosediminibacter |
| *hsl* | *Halobacterium salinarum* | Archaea | Halobacterium |
| *hvo* | *Haloferax volcanii* | Archaea | Haloferax |
| *pto* | *Picrophilus torridus* | Archaea | Picrophilus |

# Experiment 1: results

**Considerations**

- ✓ Good classification between Kingdoms
- ✓ Good discrimination of the organisms belonging to the extreme ecological niches

- $hsl$ and $hvo$ are more similar thanks to their ability to resist in environment with high level of salinity
- $pto$ survives in torrid environments



# Experiment 2: Carbon fixation

**Pathway**: Carbon fixation in photosynthetic organisms

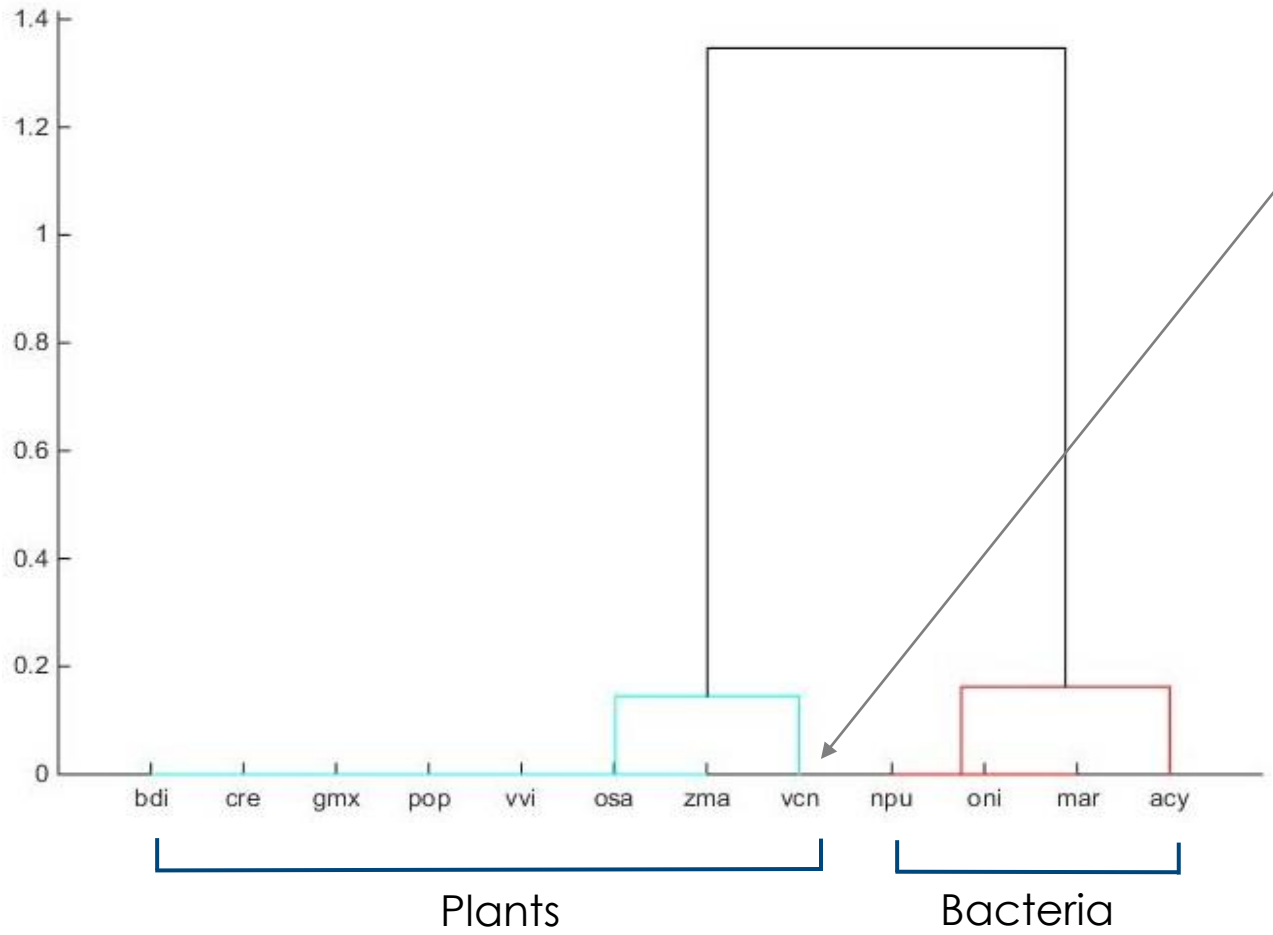
**Aim**: Test the discrimination of our method wrt. a set of organisms that perform variants of the carbon dioxide conversion process.

**Sim. Index**: $SimP_i$

Organisms that lives in different environment present variant of the metabolic pathway due to the environmental adaptation.

| Code | Organism | Kingdom | Taxonomic group |
|---|---|---|---|
| *gmx* | *Glycine max* (soybean) | Plants | Pea family |
| *pop* | *Populus trichocarpa* (black cottonwood) | Plants | Willow family |
| *vvi* | *Vitis vinifera* (wine grape) | Plants | Grape family |
| *osa* | *Oryza sativa japonica* (Japanese rice) | Plants | Grass family |
| *zma* | *Zea mays* (maize) | Plants | Grass family |
| *bdi* | *Brachypodium distachyon* | Plants | Grass family |
| *cre* | *Chlamydomonas reinhardtii* | Plants | Green algae |
| *vcn* | *Volvox carteri f. nagariensis* | Plants | Green algae |
| *npu* | *Nostoc punctiforme* | Bacteria | Nostoc |
| *acy* | *Anabaena cylindrica* | Bacteria | Anabaena |
| *oni* | *Oscillatoria nigro-viridis* | Bacteria | Oscillatoria |
| *mar* | *Microcystis aeruginosa* | Bacteria | Microcystis |

# Experiment 2: results




$vcn$ is a green algae and in particular a pluricellular organisms with a simplified carbon fixation cycle.

**Considerations:**

- ✓ Good classification of Plants and Bacteria
- ✓ Good discrimination of the green algae $vcn$ wrt. the other Plants

# Experiment 3: Metabolic evolution

## Aim

The aim of the experiment is to verify if the similarities in the metabolism of a group of organisms find a correspondence in the phylogenesis found in the literature

## Organisms

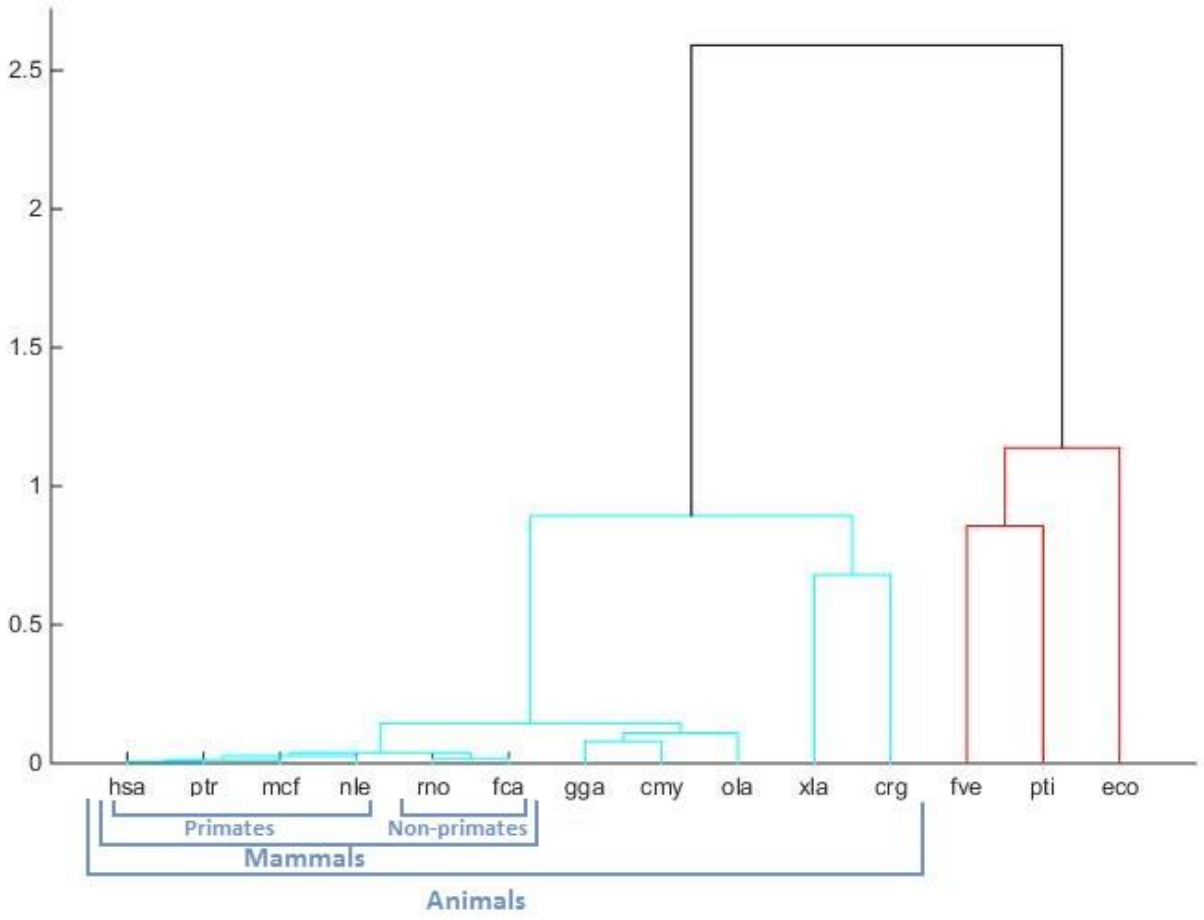
| Code | Organism | Kingdom | Taxonomic group |
|---|---|---|---|
| hsa | Homo sapiens (human) | Animals | Mammals |
| ptr | Pan troglodytes (chimpanzee) | Animals | Mammals |
| nle | Nomascus leucogenys (gibbon) | Animals | Mammals |
| mcf | Macaca fascicularis | Animals | Mammals |
| rno | Rattus norvegicus | Animals | Mammals |
| fca | Felis catus (cat) | Animals | Mammals |
| gga | Gallus gallus (chicken) | Animals | Birds |
| cmy | Chelonia mydas (green turtle) | Animals | Reptiles |
| xla | Xenopus laevis (frog) | Animals | Amphibians |
| ola | Oryzias latipes | Animals | Fishes |
| crg | Crassostrea gigas (Pacic oyster) | Animals | Mollusks |
| fve | Fragaria vesca (strawberry) | Plants | Rose family |
| pti | Phaeodactylum tricornutum | Chromista | Chromalveolata |
| eco | Escherichia coli | Bacteria | Proteobacteria |

## Tool configuration

- Pathway: set
- Network: undirected
- Index: Combined Index

What we expect is that our similarity indexes produces a classification close to the phylogenetic one.

# Experiment 3: Results




| Code | Organism | Kingdom | Taxonomic group |
|---|---|---|---|
| hsa | Homo sapiens (human) | Animals | Mammals |
| ptr | Pan troglodytes (chimpanzee) | Animals | Mammals |
| nle | Nomascus leucogenys (gibbon) | Animals | Mammals |
| mcf | Macaca fascicularis | Animals | Mammals |
| rno | Rattus norvegicus | Animals | Mammals |
| fca | Felis catus (cat) | Animals | Mammals |
| gga | Gallus gallus (chicken) | Animals | Birds |
| cmy | Chelonia mydas (green turtle) | Animals | Reptiles |
| xla | Xenopus laevis (frog) | Animals | Amphibians |
| ola | Oryzias latipes | Animals | Fishes |
| crg | Crassostrea gigas (Pacic oyster) | Animals | Mollusks |
| fve | Fragaria vesca (strawberry) | Plants | Rose family |
| pti | Phaeodactylum tricornutum | Chromista | Chromalveolata |
| eco | Escherichia coli | Bacteria | Proteobacteria |

# Experiment 4: Yeasts and Molds metabolism

## Aim

The aim of the experiment is to test the classification of a group of organisms belonging to the same Kingdom.

## Organisms

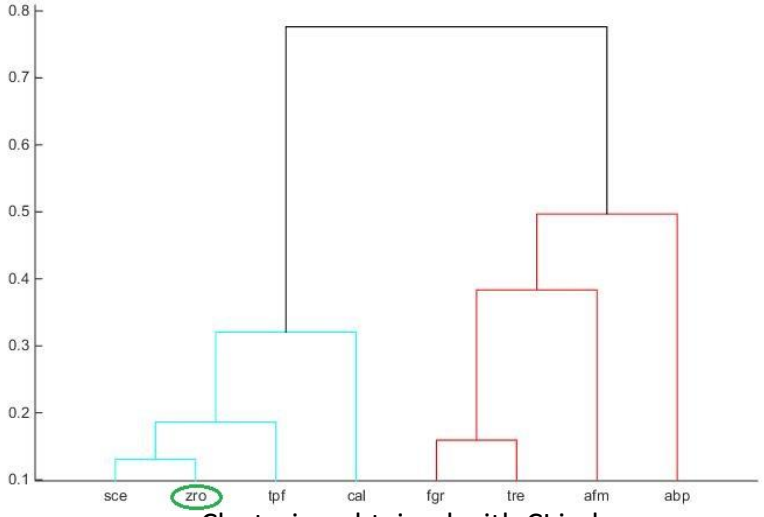
| Code | Organism | Kingdom | Taxonomic group |
|---|---|---|---|
| sce | Saccharomyces cerevisiae | Fungi | Saccharomycetes |
| zro | Zygosaccharomyces rouxii | Fungi | Saccharomycetes |
| tpf | Tetrapisispora phai | Fungi | Saccharomycetes |
| cal | Candida albicans | Fungi | Saccharomycetes |
| fgr | Fusarium graminearum | Fungi | Sordariomycetes |
| tre | Trichoderma reesei | Fungi | Sordariomycetes |
| afm | Aspergillus fumigatus | Fungi | Eurotiomycetes |
| abp | Agaricus bisporus var. burnettii | Fungi | Basidiomycetes |

## Tool configuration

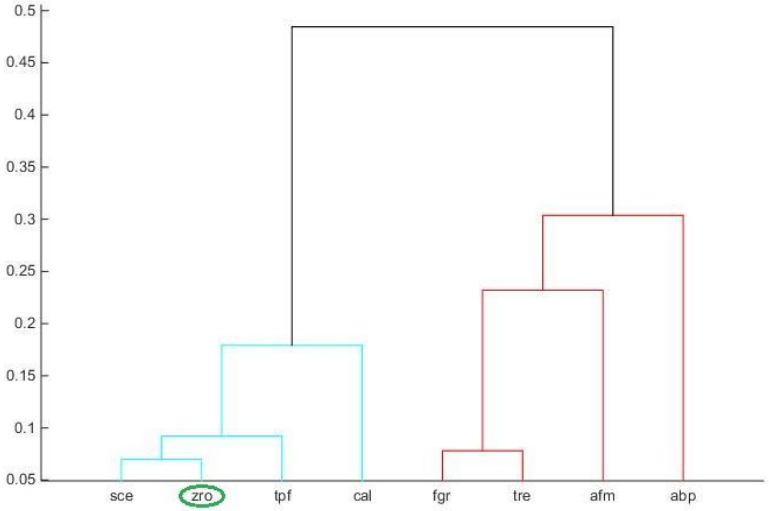
- Pathway: set
- Network: undirected
- Index: Combined Index, Separated Index
- Alpha: 0.2, 0.5

What we expect is a clear separation between Yeasts and Molds

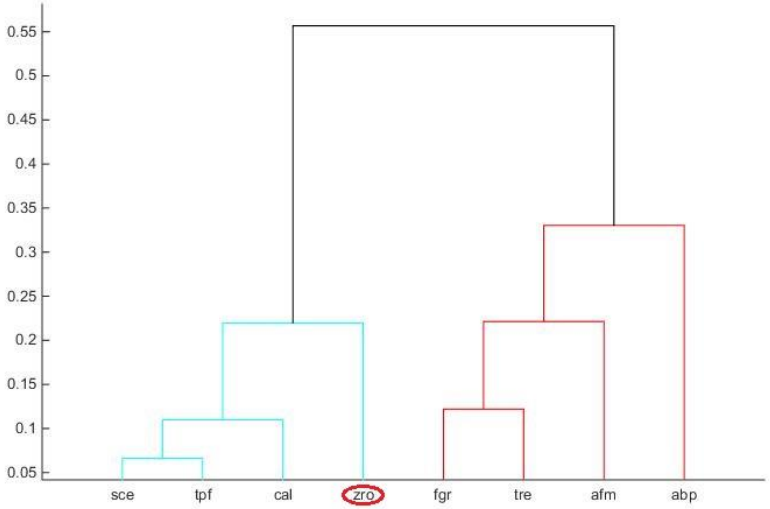
# Experiment 4: Results




Clustering obtained with CI index




Clustering obtained with Si index (α=0,5)




Clustering obtained with Si index (α=0,2)

| Code | Organism | Kingdom | Taxonomic group |
|---|---|---|---|
| sce | Saccharomyces cerevisiae | Fungi | Saccharomycetes |
| zro | Zygosaccharomyces rouxii | Fungi | Saccharomycetes |
| tpf | Tetrapisispora phai | Fungi | Saccharomycetes |
| cal | Candida albicans | Fungi | Saccharomycetes |
| fgr | Fusarium graminearum | Fungi | Sordariomycetes |
| tre | Trichoderma reesei | Fungi | Sordariomycetes |
| afm | Aspergillus fumigatus | Fungi | Eurotiomycetes |
| abp | Agaricus bisporus var. burnettii | Fungi | Basidiomycetes |

# Conclusions & further dev.

**Benefits:**

- ✓ **Indipendent levels** allow for different comparisons between pathways and networks
- ✓ **Avoid the computational problems** reducing the size of the metabolic network graph and exploiting the standardized modularization of KEGG data
- ✓ Allows for **fast comparison** between metabolic pathways
- ✓ Provides a **good classification** of the organism at pathway and global level

**Further developments:**

- New refined methods for comparison of both networks and pathways
- New functionality for the selection of one or more pathways
- New functionality for the selection of specif groups of organisms
- Determine a threshold value on the similarity measure for each Kingdom
- Integration of hierarchical clustering algorithm for cluster analysis and the generation of the corresponding phylogenetic trees

# References


- Christophe H. Schilling, Stefan Schuster, Bernhard O. Palsson and Reinhart Heinrich. *Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-genomic Era*.
- Michael Palmer. *In Human Metabolism*, chapter: Introduction, pages: 1-2. Department of Chemistry, University of Waterloo, 2015.
- Donald Voet, Charlotte W. Pratt and Judith G. Voet. *In Fundamentals of Biochemistry: Life at the Molecular Level*, pages: 436-439, 442. John Wiley and Sons, 4° edition, 2012.
- Bernhard O. P. *Systems Biology: Properties of Reconstructed Networks.* Cambridge University Press,1° edition, 2006.

# Thank you for the attention.