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
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 interactions 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

Metabolism 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 constantly 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
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

Nucleotide m.

Energy m.

Amino acid m.

Carbohydrate m.

Lipid m.

Erboso Gianluca & Meggiato Alberto
Comparison of Metabolic Networks: a two-level approach.

Bioinformatics Day @ DAIS
07/07/2016
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**.

Independent levels allows for computing different similarity indexes (topology and functionality) that can be combined later.
**Metabolic network 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.

During the parsing phase we consider:
- The current metabolic function
- Relation tag of type='maplink'

```xml
<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">
  ...
</pathway>

<entry id="41" name="path:hsa00030" type="map"
  link="http://www.kegg.jp/dbget-bin/www_bget?hsa00030">
  <graphics name="Pentose phosphate pathway" fbgcolor="#000000" bgcolor="#FFFFFF"
    type="roundrectangle" x="856" y="359" width="62" height="257"/>
</entry>

<entry id="56" name="hsa:2597 hsa:26330" type="gene reaction="ra:R01001"
  <graphics name="GAPDH, G6PD, GAPD, HEL-B-1628P..." fbgcolor="#000000" bgcolor="#808080"
    type="rectangle" x="450" y="484" width="46" height="17"/>
</entry>

<relation entry1="41" entry2="56" type="maplink">
  <subtype name="compound" value="130"/>
</relation>
```
Data structure

Graph of metabolism $\rightarrow$ modified adjacency matrix of fixed size $\rightarrow$ 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\}.

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 $O'$.

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

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

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

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

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 \( v \in V \) (\( v' \in V' \)).

The structural similarity index \( SimS_i \) wrt. the i-th pathway, is defined as:

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

The structural network similarity index is defined as:

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

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:

\[
CI = \frac{\sum_{i=1}^{n} SimS_i \ast SimP_i}{n}
\]

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

\[
SI = \alpha \ast SimS + (1 - \alpha) \ast SimPW
\]
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: Simₚᵢ

- 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.

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

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

Organisms that live in different environments present variants of the metabolic pathway due to environmental adaptation.
Experiment 2: results

Considerations:

- Good classification of Plants and Bacteria
- Good discrimination of the green algae \( vcn \) wrt. the other Plants

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

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

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

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Organism</th>
<th>Kingdom</th>
<th>Taxonomic group</th>
</tr>
</thead>
<tbody>
<tr>
<td>sce</td>
<td>Saccharomyces cerevisiae</td>
<td>Fungi</td>
<td>Saccharomycetes</td>
</tr>
<tr>
<td>zro</td>
<td>Zygosaccharomyces rouxii</td>
<td>Fungi</td>
<td>Saccharomycetes</td>
</tr>
<tr>
<td>tpf</td>
<td>Tetrapisispora phai</td>
<td>Fungi</td>
<td>Saccharomycetes</td>
</tr>
<tr>
<td>cal</td>
<td>Candida albicans</td>
<td>Fungi</td>
<td>Saccharomycetes</td>
</tr>
<tr>
<td>fgr</td>
<td>Fusarium graminearum</td>
<td>Fungi</td>
<td>Sordariomycetes</td>
</tr>
<tr>
<td>tre</td>
<td>Trichoderma reesei</td>
<td>Fungi</td>
<td>Sordariomycetes</td>
</tr>
<tr>
<td>afm</td>
<td>Aspergillus fumigatus</td>
<td>Fungi</td>
<td>Eurotiomycetes</td>
</tr>
<tr>
<td>abp</td>
<td>Agaricus bisporus var. burnettii</td>
<td>Fungi</td>
<td>Basidiomycetes</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Code</th>
<th>Organism</th>
<th>Kingdom</th>
<th>Taxonomic group</th>
</tr>
</thead>
<tbody>
<tr>
<td>sce</td>
<td>Saccharomyces cerevisiae</td>
<td>Fungi</td>
<td>Saccharomyces</td>
</tr>
<tr>
<td>zro</td>
<td>Zygosaccharomyces rouxii</td>
<td>Fungi</td>
<td>Saccharomyces</td>
</tr>
<tr>
<td>tpf</td>
<td>Tetrapisispora phai</td>
<td>Fungi</td>
<td>Saccharomyces</td>
</tr>
<tr>
<td>cal</td>
<td>Candida albicans</td>
<td>Fungi</td>
<td>Saccharomyces</td>
</tr>
<tr>
<td>fgr</td>
<td>Fusarium graminearum</td>
<td>Fungi</td>
<td>Sordariomycetes</td>
</tr>
<tr>
<td>tre</td>
<td>Trichoderma reesei</td>
<td>Fungi</td>
<td>Sordariomycetes</td>
</tr>
<tr>
<td>afm</td>
<td>Aspergillus fumigatus</td>
<td>Fungi</td>
<td>Eurotiomycetes</td>
</tr>
<tr>
<td>abp</td>
<td>Agaricus bisporus var. burnettii</td>
<td>Fungi</td>
<td>Basidiomycetes</td>
</tr>
</tbody>
</table>

Clustering obtained with Cl index

Clustering obtained with Si index (α=0.5)

Clustering obtained with Si index (α=0.2)
Conclusions & further dev.

Benefits:

- **Independent 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 specific 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


Thank you for the attention.