Università
Ca'Foscari
Venezia

Master's Degree programme – Second Cycle
(*D.M. 270/2004*)
in Informatica – Computer science

Final Thesis

Ca' Foscari
Dorsoduro 3246
30123 Venezia

Comparing metabolic networks at pathway level

Supervisor
Prof. Marta Simeoni

Assistant supervisor
Prof. Sabrina Manente

Graduand
Alberto Meggiato
817987

Academic Year
2015 / 2016

Acknowledgements

First of all, i would like to thank my thesis supervisors Marta Simeoni and Nicoletta Cocco for the precious and essential support that they gave me, during all the time, and especially for the writing of this thesis. I would also like to thank the experts Sabrina Manente and Martina Bocci who were involved in our project. Their suggestions and explanations of biological concepts for experimentation and evaluation of the results have been very important.

Thanks to my classmates Antonio, Enrico, and Gianluca for all good moments passed together during the last three years.

A heartfelt thanks to Lorenzo for convincing me to continue with the Master's degree course.

Finally, I would express my special thanks to my parents who allowed me to follow this path and my girlfriend for giving me the support and continuous encouragement during these years of study.
All this would not have been possible without them. Thank you.

Abstract

Metabolic pathway comparison between different species is important to discover the differences in a metabolic function developed during the evolutionary process. This kind of analysis may allow the detection of important information useful also in drug engineering and medical science. This thesis has been developed together with [1] with the aim to propose a new comparison method that consider the entire metabolic networks overcoming the computational problems. In particular we propose a method for metabolic pathways comparison based on their representation as sets and multisets of chemical reactions. We define different similarity indices: three indices for the computation of the metabolic pathways similarities and two global indices that consider the entire metabolic network (both structure and functionalities). The information is taken from the KEGG database because it has a standardised representation of each pathway in the different organisms. The pathway comparison technique is then used in the context of metabolic networks comparison in order to solve the problems due to the size of the compared networks. The tool implemented in the proposal has been developed in Java and it allows to compare the metabolic networks of different organisms.

Index

Chapter 1

Introduction

In biology, the comparison of metabolic networks is relevant for studying the evolutionary process, finding similarities or dissimilarities between species, discovering drug targets and more in general for supporting medical science activities.

In the literature many approaches to metabolic pathway comparison have been proposed. They make use of data structures like sets, sequences and graphs (including hypergraphs and Petri Nets). Each of these approaches provides different levels of detail and can be used in specific contexts. At the same time they present computational problems that are related to the complexity of the data structures.

The comparison of entire metabolic networks as well as of metabolic pathways, is challenging from a computational point of view. Graphs are the natural data structure for representing metabolic pathway since they provide the most informative representation. As a consequence, the comparison of metabolic pathways implies the resolution of subgraphs isomorphisms problem, which is NP-hard. This thesis has been developed together with [1] with the aim to propose a new comparison method that consider the entire metabolic networks. The aim of both theses is to propose a new approach that overcomes the computational problems of metabolic networks comparison. Our method provides an abstraction of the metabolic network that is defined by two distinct levels. At the higher level we model the net using graphs in which the nodes represent the metabolic functions and the arcs represent the relations betweem the metabolic pathways themeselves. At the lower level instead, we model metabolic pathways as sets or multisets of reactions. The two levels are independent and they allow us to compute different similarity indices both at lower and higher level, respectively.

In this thesis, we consider the lower level of metabolic functionalities. Namely, a comparison method between metabolic pathways is presented. In particular, we propose different similarity indices in order to evaluate how much similar are the corresponding metabolic pathways in different organisms and how much similar are two organisms from a functional point of view. These measures are then combined with the topological similarity indices defined in [1] in order to compute a global similarity value considering the entire metabolisms (both structures and functionalities).

Our method has been implemented in a Java tool that gives the possibility to the user to perform a comparison between two different organisms. The tool is developed by using the principles and the best practices of software engineering in order to offer a good user experience during the use. Multi-threading techniques are also used in order to parallelize the computation and reduce the computational time, where it is possible. The modularization of the tool allows for extending the tool itself with new comparison methods and new functionalities.

Some experiments have been performed using our tool in order to check the quality of the results. The experiments are performed considering differents aspects in order to test the usage of the similarity indices defined in our method. Then, a clustering analysis is also performed through a hierarchical clustering algorithm which provides a data classification.

The thesis is organized as follows.
In chapter 2 we give a general overview of the metabolism from a biological point of view. We introduce the KEGG knowledge base, we discuss its databases structures and the technical aspects used for data retrieval and we introduce some basic notions used in the subsequent chapters.

In chapter 3 we present the state of art in metabolism representation. Moreover, we present some approaches selected from the literature on metabolic pathways comparison.

In chapter 4 we present the metabolic network construction describing the idea, its implementation and data structures. Then, the similarity indices used for comparison are explained.

Chapter 5 describes the tool we developed to implement the proposed approach. The functional and non-functional requirements, the project architecture and the used technologies are discussed. Furthermore a brief documentation on the usage of the tool is given.

In chapter 6 we show the result of some experiments performed in order to evaluate our similarity indices. The results have been discussed using a hierarchical clustering algorithm.

Finally, chapter 7 draws some concluding remarks and highlights some possible future developments about the comparison methods and the tool.

Chapter 2

Metabolism & KEGG Data Base

In this chapter we want to introduce, briefly, the metabolism to understand what are the elements that interact in this complex process. Moreover, we give an overview of the databases for metabolisms and in particular we examine in depth KEGG database. We describe its structure, the information which is contained, the API used to retrieve them and the methods used to represent them. We focus our attention on the description of the KGML structure to understand how to extract fundamental information used for further analysis.

2.1 Metabolism in different organisms

The ***metabolism****[2, 3] is the network of all the chemical and physical reactions that take place within the cells of the organisms.*

The complex set of chemical transformations is responsible of the growth and survival of the cells and the organisms themselves. In general, the metabolism is composed of two different and fundamental phases:

- **catabolism**[3, 4]: it is composed by all the metabolic tasks that produce simpler substances, producing energy (ATP);
- **anabolism**[3, 4]: it is the opposite of the catabolism. It is composed by all the synthesis tasks that produce more complex organic molecules from simpler

ones, consuming the energy released from catabolism;

The metabolism is composed by many different interacting functions called metabolic pathways.

*A **metabolic pathway**[4] is a sequence of reactions such that the product of a single reaction can be used as reagent for another one.*

In this document we refer to metabolic pathway using equivalently the term pathway. A methabolic pathway has an associated function. As an example you may consider the Glicolysis pathway. An interesting aspect is that each function occur in a specific location in the cells. The synthesis of particular substances in distinct compartments requires mechanisms to transport these substances between compartments. For example, to move ATP generated in the mitochondria, to the cytosol where most of it is consumed, a transport of protein is necessary. Figure 2-1 shows some metabolic functions locations in a eukaryotic cell.

In a pathway we may find biochemical reactions through which, using particular enzymes, there is a catalization. A *catalization* is a chemical phenomenon where the speed of reactions undergoes changes for the intervention of one or more substances said catalysts. A *substratum*, a particular molecule where the enzyme operates, is transformed into a product used as substratum in the next step. The *reactants*, the *intermediates* and the *products*, respectively the substance consumed in the chemical reaction, the molecule formed from the reactants and that reacts further to produce the product, the final element of the reaction, are called *metabolites.* The quantitative connection between these elements is specified by the **stoichiometry**. Through *stoichiometric coefficient* it is possible to mathematically represent the reagents and products quantities involved in a reaction[5]. According to the classification of the metabolic phases, sequences of catabolic and anabolic reactions are called *catabolic pathways* and *anabolic pathways* respectively. Their continuous overlap forms a complex exchange system which is the basis of growth and survival of cells.

[1] Image from: Judith G. Voet, Donald Voet, CHarlotte W. Pratt. In Fundamentals of Biochemistry: Life at the Molecula Level, page 442. John Wiley and Sons, 4 edition, 2012

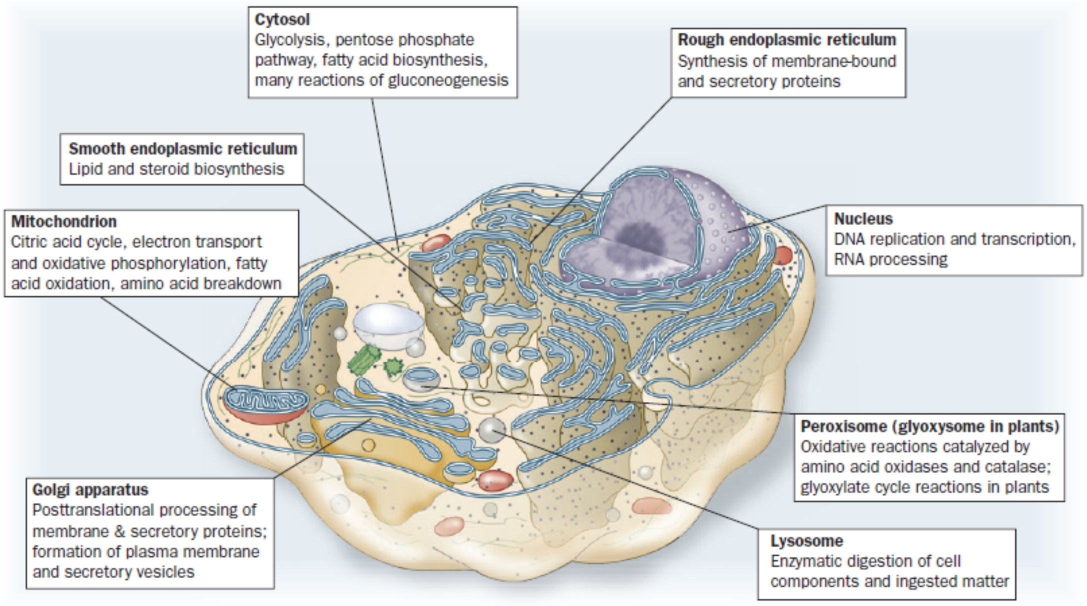

Cytosol
Glycolysis, pentose phosphate pathway, fatty acid biosynthesis, many reactions of gluconeogenesis
Rough endoplasmic reticulum
Synthesis of membrane-bound and secretory proteins
Smooth endoplasmic reticulum
Lipid and steroid biosynthesis
Nucleus
DNA replication and transcription, RNA processing
Mitochondrion
Citric acid cycle, electron transport and oxidative phosphorylation, fatty acid oxidation, amino acid breakdown
Peroxisome (glyoxysome in plants)
Oxidative reactions catalyzed by amino acid oxidases and catalase; glyoxylate cycle reactions in plants
Golgi apparatus
Posttranslational processing of membrane & secretory proteins; formation of plasma membrane and secretory vesicles
Lysosome
Enzymatic digestion of cell components and ingested matter

Figure 2-1: Metabolic pathways[1]

The ***metabolic network*** *represents the complete set of metabolic functions and processes that determine the structure and properties of the cells. These functions are not independent but they interact each one with the others creating a more complex structure, representing the network.*

In Figure 2-2 the entire metabolic network is represented from a general point of view. In the different living species the metabolism is similar in the components and in the organizations into metabolic pathways. The main metabolic pathways, such as Glycolisys or Citric acid cycle, for instance, probably appeared in one universal ancestor and they have been conserved during evolution because of their efficiency (ability to get to the final products with small number of steps). The analysis and comparison of metabolic networks between different organisms may yield important information on their evolution. Moreover, useful applications of such comparison are related to human disease analysis, drug design and metabolic engineering. However, some problems arise in metabolic networks comparison. Using a graph based modelling system, the resulting graph that represent a metabolic network may be composed by hundreds of nodes and thousands of edges. In this case, the graph matching may represent an infeasible computational problem. From graph theory in fact, the exact graph matching problems, like isomorphism and sub-graphs isomorphism, are known to be NP-Complete problems.

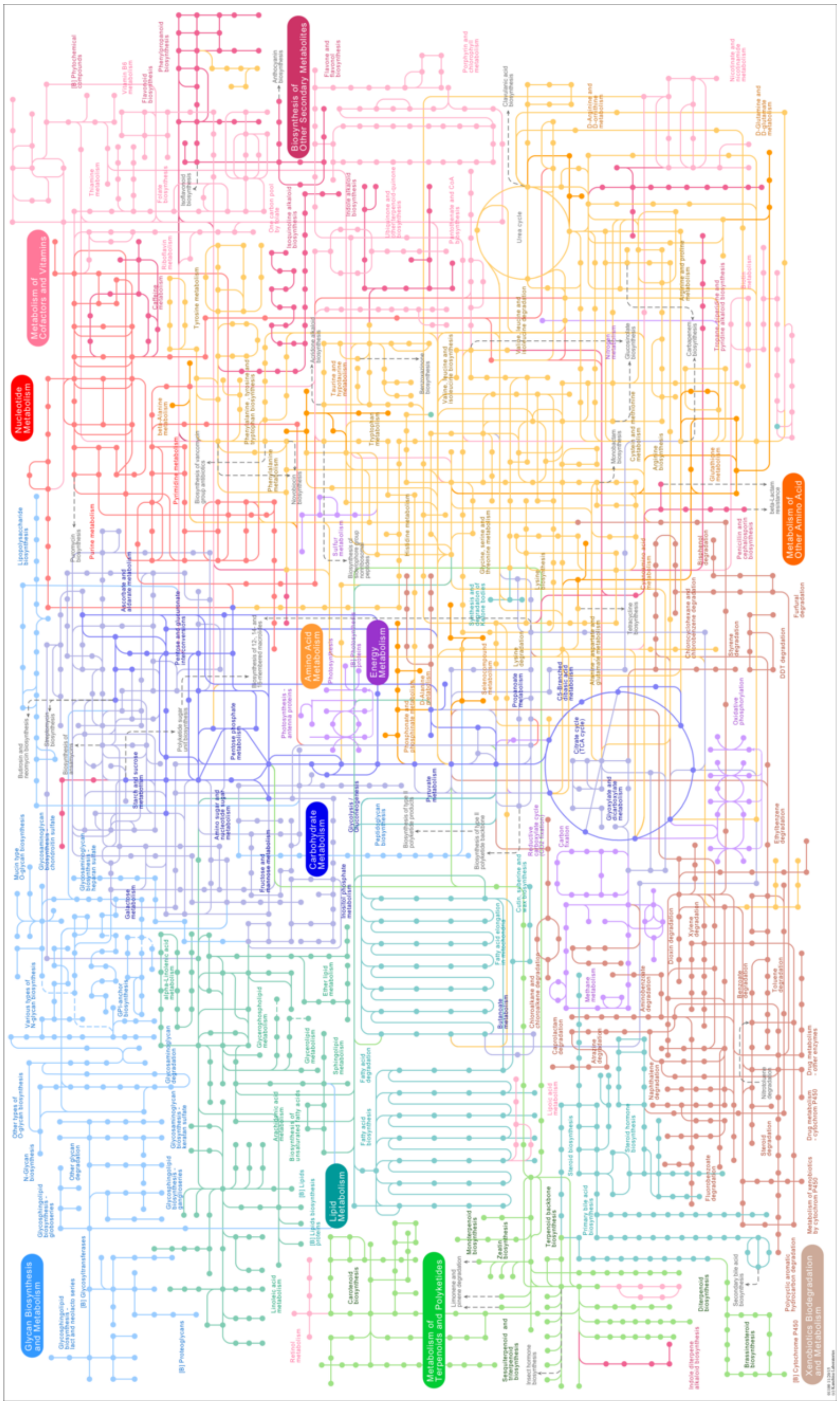

Biosynthesis of Other Secondary Metabolites
Metabolism of Cofactors and Vitamins
Nucleotide Metabolism
Metabolism of Other Amino Acid
Amino Acid Metabolism
Energy Metabolism
Carbohydrate Metabolism
Lipid Metabolism
Glycan Biosynthesis and Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism

Figure 2-2: Metabolic Network[2]

[2] Image taken from: `http://www.kegg.jp/kegg-bin/show_pathway?map01100`

2.2 Databases for metabolisms

Since the early 80's to today, thanks to the development of the technology both in computer science and in biology, the number of databases for biological data is growing.

*A **database** is a collection of related data concerning a same topic that are stored and that can be used from applications.*[6]
In particular, a metabolism database contains data involving metabolities, enzymes, genes and reactions and information about their relations. In general, metabolic databases are incomplete due to the complexity in digitalization of such kind of data. We can cite as the most used:

- KEGG[7]: a good description of this project is given in the next section of this chapter;

- BioCyc[8]: it contains a collection of databases about metabolic pathways and the genome of thousand of organisms integrating information from other databases, such as protein features and Gene Ontology information. BioCyc databases are divided into three categories related to the manual update frequency. Tier 1 is the most frequent updated database, Tier 2 receives a moderate quantity of manual update and Tier 3 receives only computational updates. In particular Tier 1 contains the following databases: EcoCyc, HumanCyc, MetaCyc, AraCyc, LeishCyc, YeastCyc. The first two contains the entire genome of the Escherichia coli and Human organism, MetaCyc contains representative metabolism sample of more than 2600 organisms and the lasts three contains information about Arabidopsis thaliana, Leishmania major and Saccharomyces cerevisiae organism.

- SEED[9]: it is a project developed by the Fellowship for Interpretation of Genomes (FIG) with the aim to develop a comparative genomics enviroment. The curation of genomic data is performed through subsystems by an expert annotator across many genomes. A subsystem in this case is defined by a set of functional related roles. Then, a metabolic pathway is represented by subsystems, as the collection of functional roles, creating complex class of proteins. The result given by the subsystems extracts a set of protein families (FIGfams). The latter in turn, create the core component of the RAST subsystem. RAST, or Rapid Annotation using Subsystem Technology, is a rapid and very accurate annotation technology that makes use of data and procedures provided from SEED framework. Therefore, this technology provides a good level of automation in high quality gene calling and functional annotation.

- BioModels[10]: it contains computational models of biological processes. These are collected from literature and they are integrated with other references and stored in a set of MySQL tables. It is divided into three categories: the curated one that contains the curated models, the non-curate one that contain models that cannot be curated or still not curated and the automatic generated one that contains models generated automatically from other databases.

2.3 KEGG Database

The KEGG[7, 11, 12] (Kyoto Encyclopedia of Genes and Genomes) Database was started in 1995 by Minoru Kanehisa, Professor at the Institute for Chemical Research at Kyoto University with the purpose to collect all the information on sequenced organisms. It is presently one of the most important collection of biological data, containing information on metabolic pathways, Genomic information, Chemical information and Health information of different organisms.

One of the main efforts of the KEGG project is to standardize gene annotations, providing functional information of cellular processes to genomics. In order to do that, all the available knowledge about systemic functions, biochemical pathways and other kinds of molecular interactions have been taken by hands and then reorganized in a computable way, creating a big digital knowledge base. As a result, KEGG becomes one of the reference knowledge bases for data integration and systematic interpretation of sequence data. KEGG project aims to provide and maintain a reliable knowledge base, supporting basic research activities in biology. Moreover, thinking about the benefits that information technology has given and can give through digitalization, large-scale data organization and development of tools for data analysis, a natural evolution of the knowledge bases is expected. KEGG in fact, is being expanded exploiting data extraction coming from the use of applications and tools based on its database. Thus, a new kind of information is collected. The latter, is typically related to health information including human diseases, drugs and other health-related substances.

2.3.1 Informative contents and data organization

As we said before, the KEGG aim is to automatize the interpretation of biological information encoded in sequence data. The prediction of gene function is also a considered problem. In particular, prediction of gene function is treated like a reconstruction process of biological system functioning, starting from genes and their products. Understanding how genes and molecules interoperate defining a biological system is typically a critical task. Therefore, a good organization of data is necessary.

The data are stored in four different macro categories as we can see in table 2.1. Each of them contains some specific databases and in particular:

- *System information category* contains functional information on how molecules and genes interact (KEGG PATHWAY), functional hierarchies of biological entities and functional units for biological interpretation of genomes. Moreover some other kind of information are stored, for instance: cell cycle, membrane transport, and more in general, information about regulatory aspects of cells function.

- *Genomic information category* contains structural information about genomics for all different organisms, gene catalogs, complete genomes and ortholog groups.

- *Chemical information category* contains information about chemical compounds, enzymes, molecules and reactions.

Category	Database	Content	Color
Systems information	KEGG PATHWAY KEGG BRITE KEGG MODULE	KEGG pathway maps BRITE functional hierarchies KEGG modules of functional units	KEGG
Genomic information	KEGG ORTHOLOGY	KEGG Orthology (KO) groups	KEGG
	KEGG GENOME KEGG GENES KEGG SSDB	KEGG organisms with complete genomes Gene catalogs of complete genomes Sequence similarity database for GENES	KEGG
Chemical information KEGG LIGAND	KEGG COMPOUND KEGG GLYCAN KEGG REACTION KEGG RPAIR KEGG RCLASS KEGG ENZYME	Metabolites and other small molecules Glycans Biochemical reactions Reactant pair chemical transformations Reaction class defined by RPAIR Enzyme nomenclature	KEGG
Health information KEGG MEDICUS	KEGG DISEASE KEGG DRUG KEGG DGROUP KEGG ENVIRON	Human diseases Drugs Drug groups Crude drugs and health-related substances	KEGG

Table 2.1: The KEGG database. Table taken from http://www.kegg.jp/kegg/kegg1a.html

- *Health information* contains health information including human diseases, drugs and other health-related substances.

All these information are organized and represented in a big wiring diagrams called *Reference Pathway* that constitutes the core of the resource. In Figure 2-2 is represented the entire metabolic network generated by KEGG.
As we can see, the map identifies specific areas using different colours. In turn, each area corresponds to a specific metabolic pathway/function and integrate all the information available in the knowledge base, like interactions and reactions. Later we will see more details about maps, pathways etc.

In biology organisms are divided into categories that compose a taxonomic hierarchy built on different levels. From the top to the bottom the levels are the following: domain, kingdom, phylum, class, order, family, genus and species. Each level inherits the features from the upper one and it adds others to classify more precisely the organism. KEGG database recalls this organization, in fact it is split into macro-categories divided into more smaller ones. Each organism belongs to a specific category. The mains are:

- *Eukaryotes* divided into animals, plants, fungi and protists;
- *Prokaryotes* divided into bacteria and archaea.

At present, in the database we find 313 Eukaryotes and 3562 Bacteria plus 215 Archea for the Prokaryotes. All these organisms are collected as complete genomes.

It is important to underline that KEGG is a freely accessible resource, that is constantly updated by the staff. So, we may see over the time, changes of data on the basis of new scientific discoveries and integrations.

2.3.2 KEGG Metabolism

As depicted in the figure 2-2, there are different metabolisms, each of them includes several distinct metabolic functions. In particular, we find information about: Carbohydrate metabolism, Energy metabolism, Lipid metabolism, Nucleotide metabolism, Amino acids metabolism, Glycan biosynthesis and metabolism, Metabolism of cofactors and vitamins, Metabolism of terpenoids and polyketides, Biosynthesis of other secondary metabolites, Xenobiotics biodegradation and metabolism, Chemical structure transformation maps.

KEGG subdivides the metabolic network into modules that represent the union of reference pathways. This structure does not constitute a partition over the network since each pathway can share parts of it with another one. It is known that the metabolic pathway are quite preserved among organisms and so, KEGG associates to each function, a unique reference pathway which corresponds to the union of the corresponding pathway in different organisms. It is possible to obtain a specific organism pathway from the corresponding reference one. The same concept is applied to the entire metabolic network: from the *reference network* that represent the union of all the reference pathways, it is possible to obtain a specific metabolic network for a choosen organism. In KEGG, the visual representation of an organism specific network, is given by highlighting the interested parts for the organism and by shading all the rest. Graphically this approach gives to the users a rapid view of the overall functions present in a chosen organism and also the relations between them.

The metabolic network of homo sapiens is represented in Figure 2-3.

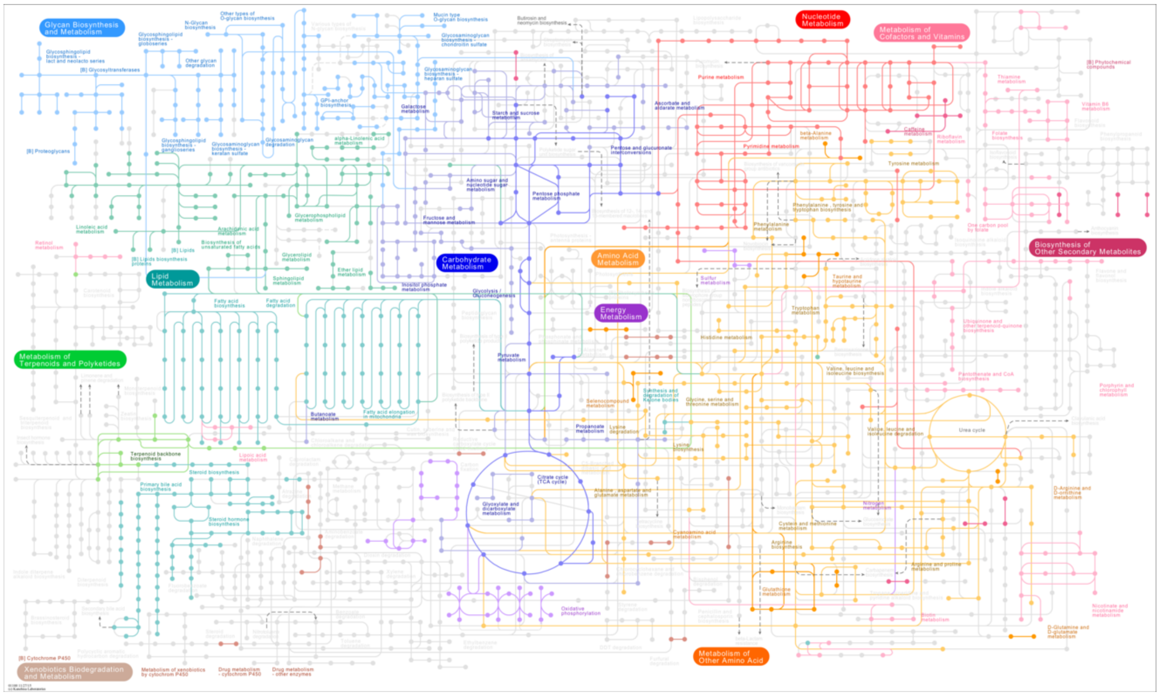

Glycan Biosynthesis and Metabolism
Nucleotide Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Carbohydrate Metabolism
Lipid Metabolism
Amino Acid Metabolism
Energy Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism
Metabolism of Other Amino Acid
Citrate cycle (TCA cycle)
Urea cycle
Oxidative phosphorylation
Purine metabolism
Pyrimidine metabolism
Pentose phosphate metabolism
Starch and sucrose metabolism
Fatty acid biosynthesis
Fatty acid degradation
Steroid biosynthesis
Glycolysis / Gluconeogenesis
Pyruvate metabolism
Tyrosine metabolism
Histidine metabolism

Figure 2-3: Metabolic Pathways of homo sapiens[3]

2.3.3 KEGG Pathways

KEGG PATHWAY is a collection of manually drawn diagrams and related textual informations. For each pathway we find a graphical representation typically called pathway map and a textual one written in XML format called KGML file. The latter representation is described in detail in Section 2.3.3.1. Every map graphically represents all the KEGG knowledge about molecular pathways for metabolism, adding important information about molecular interactions and reaction networks. Reaction networks are obtained through an integration of genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development.

[3] The image is available at the following link: `http://www.kegg.jp/kegg-bin/show_pathway?org_name=hsa&mapno=01100&mapscale=0.35&show_description=hide&show_module_list=`

In the graphical representation, each map is composed by four different objects:

- **boxes** that identify gene products (enzymes);
- **circles** that represent other molecules, typically chemical compounds;
- **rectangles** for other maps representations;
- **lines** for molecular interactions.

Figures 2-4 and 2-5 summarize the entire notation for pathway map representation. They are taken from KEGG documentation[4]. Two kind of maps are provided for

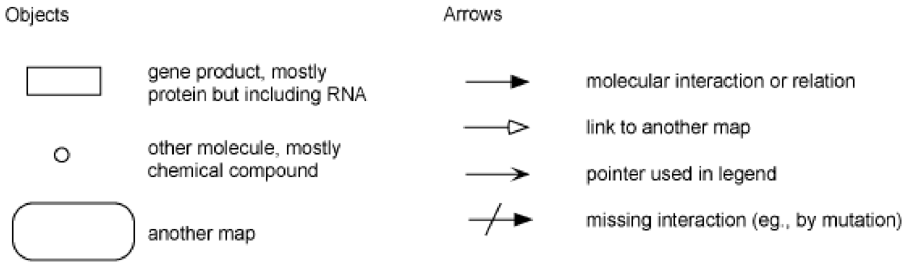

Objects
gene product, mostly protein but including RNA
other molecule, mostly chemical compound
another map
Arrows
molecular interaction or relation
link to another map
pointer used in legend
missing interaction (eg., by mutation)

Figure 2-4: Symbols for map notation

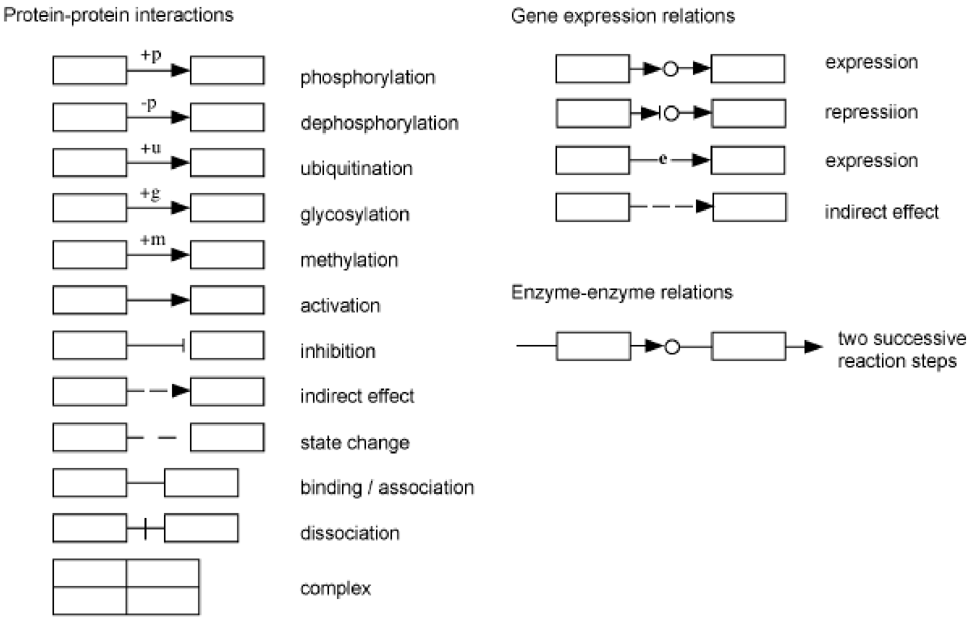


Figure 2-5: Different kind of relations in map Notation

[4] The symbology is available at the address `http://www.genome.jp/kegg/document/help_pathway.html`

each metabolic pathway: the reference pathway and organism-specific pathways. At present, there are 176 metabolic pathways, each one with its reference metabolic map, and for each of them we find more specific maps and KMGL files, one for each organism that includes the metabolic function under examination. A pathway has a unique identifier for reference map composed from the keyword *map* followed by a five digit number to distinguish the metabolic function. The organism-specific pathways instead, are identified by using *org* prefix, expressed with three or four letters that specify an organism. For example, the code *map00010* identifies the glycolysis reference pathway, while the code *hsa00010* identifies a more detailed map, glycolysis in the homo sapiens organism. From a technical point of view, reference pathways are the original manually drawn maps and they do not make use of colours. All the specific-organism maps are instead computationally generated, following some rules, from the corresponding reference pathway. Each component of an organism-specific pathway, in turn, has a unique identifier that satisfies one the following patterns:

- *ko (KEGG Orthology)* identifier, is expressed by *ko:* followed by K and a number of five digits that specifies the Ortholog[5] group for a specific organism. For example *ko:K01568* is the identifier for the pyruvate decarboxylase;
- *rn* identifier, expressed by *rn:* followed by R and a five digit number that identifies the reaction in the relative database. For istance: *rn:R00014* corresonds to the thiamin diphosphate acetaldehydetransferase (decarboxylating) in the pyruvate;
- *ec* identifier, expressed by *ec:* followed by an EC number defined by IUBMB-IUPAC commission that identifies a specific enzyme. As an example we can consider: *ec:4.1.1.1* that corresponds to the pyruvate decarboxylase;
- *cpd* identifier, expressed by *cpd:* followed by C and a five digit number that specifies a chemical compound. For example: *cpd:C00161* corresponds to the formula: $C2HO3R$.

[5] *Orthologs* are collection of genes belonging to different species evolved from a common ancestor. They preserve the same ancestor's function during the evolution process. For that reason, the identification of orthologs is quite critical for reliable prediction of gene function in newly sequenced genomes.

At the beginning, the KEGG project was based on an automatic matching between gene catalogs and enzymes in the reference map using EC[6] numbers. Now, the EC numbers are no longer used as identifiers in KEGG and the system was updated in order to use a different mapping criteria. Taking into account the computation of organism-specific pathway, it uses KEGG Orthology (KO)[7] system as the basis for genome annotation and mapping. Green boxes are linked to genes through a conversion of KO identifiers to gene identifiers in the reference pathway. Figures 2-6 and 2-7 show the reference pathway for the Citrate Cycle and the specific Citrate Cycle for Homo Sapiens organism.

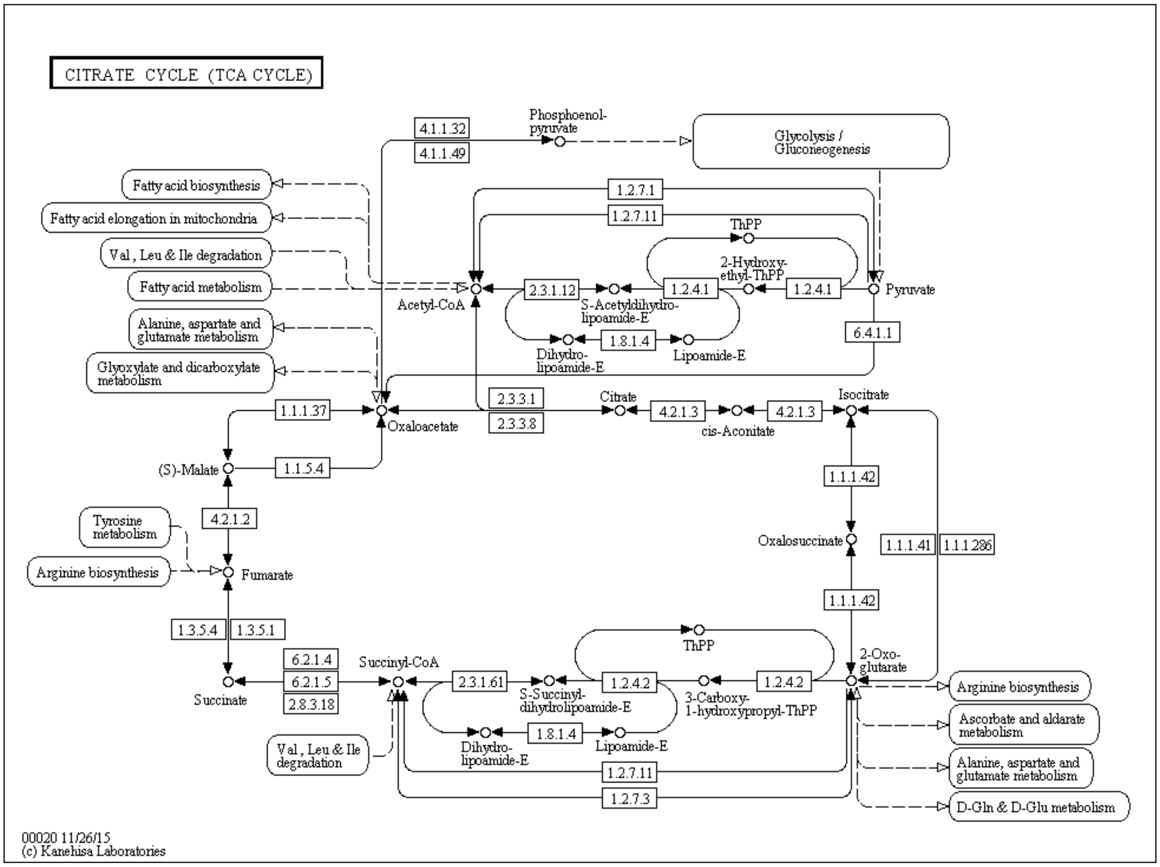


Figure 2-6: Citrate Cycle reference pathway[8]

[6] EC (Enzyme Classification) number: it provides a classification schema for enzymes. The classification is based on their catalizations and on chemical reactions. An EC number is expressed by a set of numbers separated by periods that progressively give a finer classification of the enzyme.

[7] The KEGG Orthology (KO) system is a collection of ortholog groups defined by hand, that capture experimental knowledge from literature and experimental observations. For more info about genome annotation and KO identifiers visit: *http://www.genome.jp/kegg/ko.html*

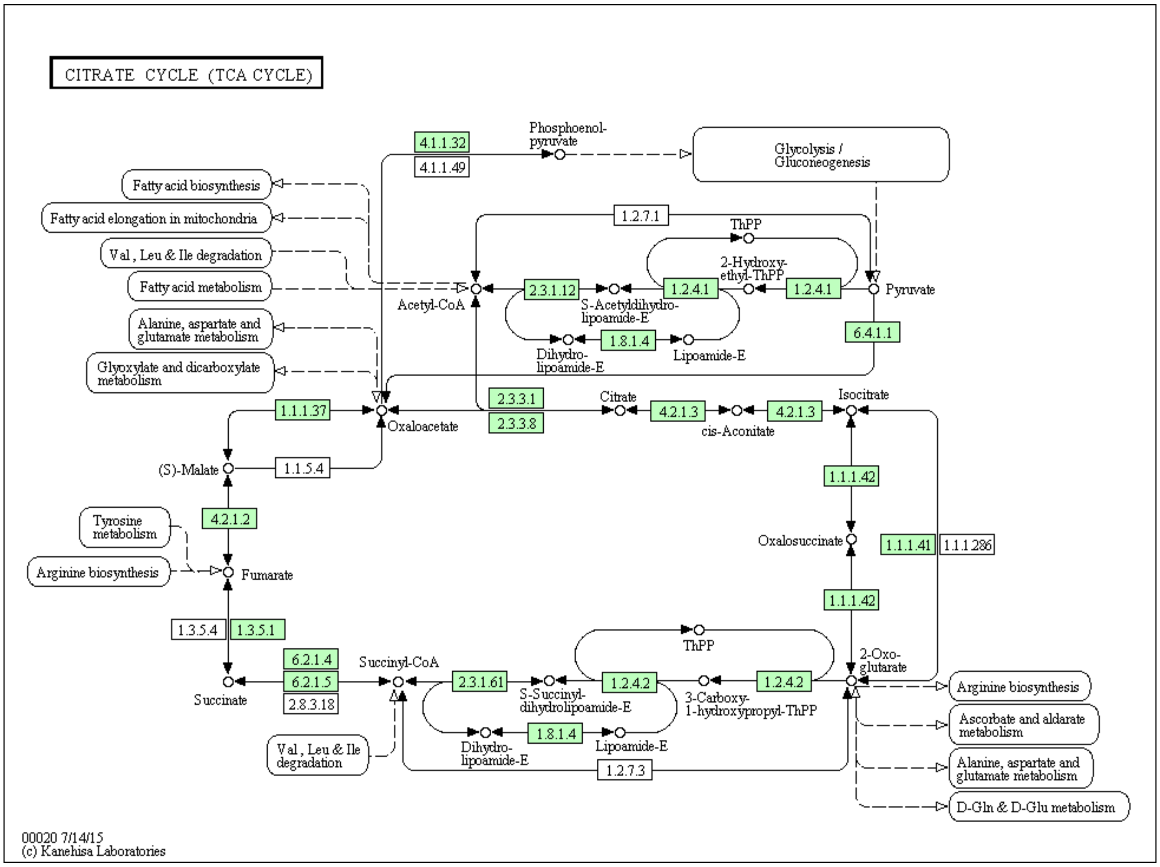


Figure 2-7: Citrate Cycle in the Homo Sapiens organism[9]

Drawing operations are performed using KegSketch software that produces a semi-static image where the map structure is fixed but each element can be coloured using user's preferences. Enzymes, maps and compounds inside a pathway map are all clickable objects and permit one to get more details on molecular structure.

The textual representation and the KGML file structure is described in the next section.

[8] Image taken from: `http://www.kegg.jp/kegg-bin/show_pathway?map=map00020&show_description=show`

[9] Image is available at the address: `http://www.kegg.jp/kegg-bin/show_pathway?org_name=hsa&mapno=00020&mapscale=&show_description=show`

2.3.3.1 KGML

The textual representation contains partial information represented in the respective map. In general, the information are written in the KGML format (KEGG Markup Language) that is based on XML language. By its markup language nature, it allows to define and control the meaning of the elements using customized tags. As we have seen before for the maps, also KGML files have a unique identifier. A specific code corresponds to the relative map. As an example, we can recognize the homo sapiens glycolysis file by the name: *hsa00010.xml*. The first three letters identify the homo sapiens organism, and the code 00010 refers to the glycolysis function.

```
<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM "http://www.kegg.jp/kegg/xml/KGML_v0.7.1_.dtd">
<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="13" name="hsa:226 hsa:229 hsa:230" type="gene" reaction="rn:R01070" link="http://www.kegg.jp/dbget-bin/www_bget?hsa:226+hsa:229+hsa:230">
        <graphics name="ALDOA, ALDA, GSD12, HEL-S-87p..." fgcolor="#000000" bgcolor="#BFFFBF" type="rectangle" x="483" y="407" width="46" height="17"/>
    </entry>
    ...
    <entry id="40" name="cpd:C00033" type="compound" link="http://www.kegg.jp/dbget-bin/www_bget?C00033">
        <graphics name="C00033" fgcolor="#000000" bgcolor="#FFFFFF" type="circle" x="146" y="958" width="8" height="8"/>
    </entry>
    ...
    <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="46" name="ko:K01568" type="ortholog" reaction="rn:R00014" link="http://www.kegg.jp/dbget-bin/www_bget?K01568">
        <graphics name="K01568" fgcolor="#000000" bgcolor="#FFFFFF" type="rectangle" x="431" y="879" width="46" height="17"/>
    </entry>
    ...
    <entry id="140" name="hsa:9562" type="gene" reaction="rn:R09532" link="http://www.kegg.jp/dbget-bin/www_bget?hsa:9562">
        <graphics name="MINPP1, HIPER1, MINPP2, MIPP" fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="571" y="630" width="46" height="17"/>
    </entry>
    ...
    <relation entry1="62" entry2="42" type="maplink">
        <subtype name="compound" value="84"/>
    </relation>
    <relation entry1="133" entry2="61" type="ECrel">
        <subtype name="compound" value="90"/>
    </relation>
    ...
    <reaction id="48" name="rn:R03270" type="irreversible">
        <substrate id="99" name="cpd:C05125"/>
        <substrate id="96" name="cpd:C15972"/>
        <product id="102" name="cpd:C16255"/>
        <product id="136" name="cpd:C00068"/>
    </reaction>
    ...
    <reaction id="13" name="rn:R01070" type="reversible">
        <substrate id="104" name="cpd:C05378"/>
        <product id="130" name="cpd:C00118"/>
        <product id="88" name="cpd:C00111"/>
    </reaction>
</pathway>
```

Listing 2.1: KGML example from hsa00010.xml file

The textual representation is the most useful for applications that manage data, because allows for data extraction. On the contrary extracting data from the visual representation is infeasible even if it is the most readable and understandable representation for the final users. The KGML files are available only for organism-specific pathways and not for the reference pathways. Each file can be downloaded from the KEGG site. However, in order to select and download KGML files, there is an API suite published by the authors. This is the case of our application that makes use of the KEPP API[10] in order to get all the useful files through an automatic process. This aspect will be describe later. In the next section we present the structure of the KGML files.

[10] APIs are available at the address: http://www.kegg.jp/kegg/rest/keggapi.html

2.3.3.2 Pathway structure in the XML file

In this section we describe the KGML content and its relation with the corresponding graphic representation. The Figure 2-8 shows an overview of the file structure.

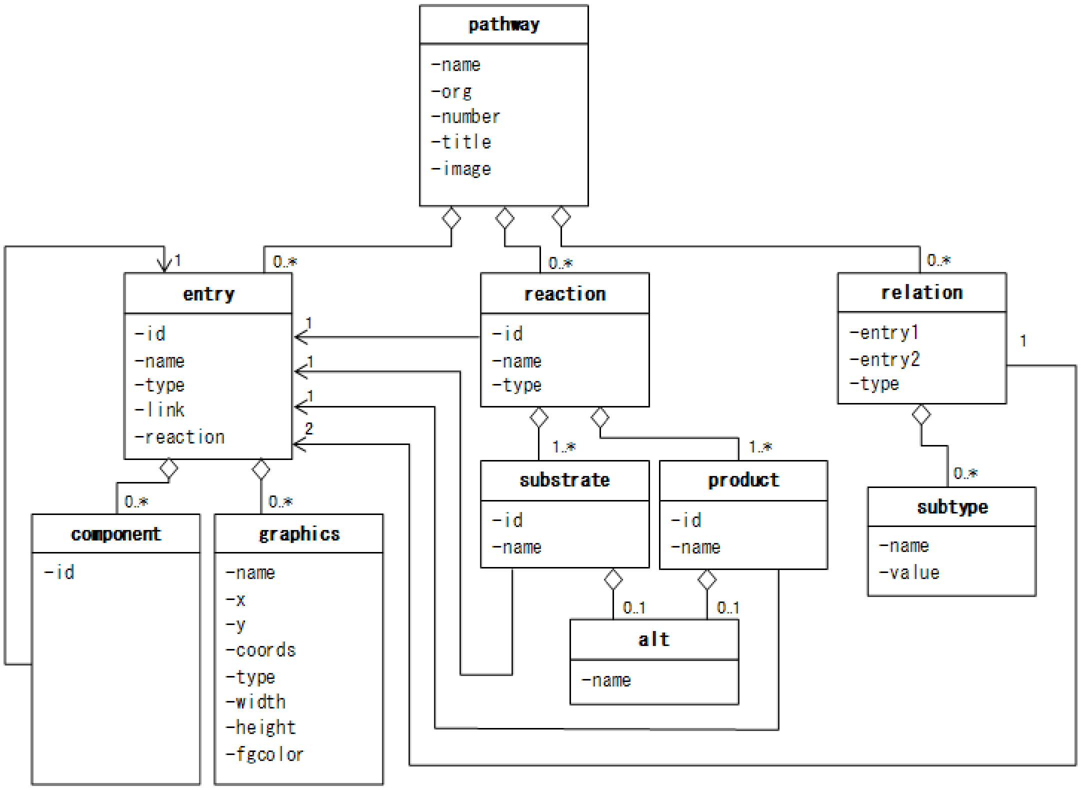


Figure 2-8: KGML Structure[11]

Each node corresponds to an element and the arrows indicate relations between nodes. The numbers associated to the arrows are cardinalities: the minimum and maximum numbers of relation instances. The tags used are:

pathway: This is the root element and it is unique in any file. It has many attributes: the name specifies the id of the pathway, the number specifies the map number and the org represents the classification of the map. The last one can assume a value among the following ones: ko, rn, ec or a three/four letter string representing the organism. All these attributes are required.

[11] Image is available at the address: `http://www.kegg.jp/kegg/xml/docs/`

entry: This element represents a node in the pathway and it contains all the information about it. The id attribute is the identification number of the entry that is unique in the file in which it is located. The same element in different maps can assume different values. The name attribute is the KEGG identifier of the entry and it is expressed in the form *db:accession*, where *accession* is the specific number of that element in the database *db*, and it can be used to perform a request using API to obtain all the information about it. Finally, the type attribute explains the element type (enzyme, reaction, gene, compound, map, group etc). The entry element has two sub-elements:

graphics: It contains all the information needed to draw the object. The name attribute is the label associated to the object, x and y attributes explain the position and the type specifies the object shape. In particular a rectangle represents a gene product, a circle represents others molecules such as compounds, roundrectangle represents linked pathway and lines represent reactions or relations.

component: It is used when the entry element is a complex node. For each component that constitutes the complex node a component element is specified with its own ID.

relation: It defines a relationship between two proteins or between a protein and a compound. Graphically it is represented as a line that connects two nodes. The direction is specified by entry1 (from) and entry2 (to). It contains also a type attribute that specifies the nature of the relation. In particular the type ECrel specifies an enzyme-enzyme relation and the type maplink specifies a relation between a protein and another one belonging to a different map.

subtype: it provides additional information about the nature of the relation such as state transition or molecular events.

reaction: A reaction substrate-product is described by this element and it is represented as an arrow between two circles.

substrate: The substrate of the reaction.

product: The product of the reaction.

alt: The alternative name of the parent.

2.3.3.3 KEGG API

The KEGG API (Application Programming Interface) allows users and applications to perform operation like searching, analizing and retrieving biological information from KEGG database. The general structure of the request is the following one:

http://rest.kegg.jp/$< operation >$/$< argument >$[/$< argument >$][/$< option >$]

where:

$$< operation > \; = \; info \mid list \mid find \mid get \mid conv \mid link$$

$$< argument > \; = \; < database > \mid < dbentries >$$

and the *option* parameter can assume different values wrt the operation:

Data search: $< option > \; = \; formula \mid exact\_mass \mid mol\_weight$

Data retrieval: $< option > \; = \; aaseq \mid ntseq \mid mol \mid kcf \mid image \mid kgml$

In particular the $< database >$ refers to the specific database name we want to use (e.g. KEGG PATHWAY, KEGG GENES ecc). The *dbentries* are in turn defined as:

$$< dbentries > \; = \; < dbentry > [+ < dbentry > ...]$$

where:

$$< dbentry > \; = \; < db : entry > \mid < kid > \mid < org : gene >$$

For our application the KEGG APIs were used to retrieve KGML files for specific organisms. In particular we have used the following type of requests:

$$http : //rest.kegg.jp/get/org : pathway/kgml$$

The *get* operation is used to retrieve data from database, where the *org:pathway* specifies the pathway of the organism and finally *kgml* specifies the format.

2.3.3.4 Problems with data

During the preliminary phases of the projet we bump into some data inconsistency problems. These concern, in particular, the graphical representations and the corresponding KGML files.
We can summarize the problem in the following ways:

- Graphical representations are not always complementary: if in a specific pathway there is a graphic connection with another map, in the second map the corresponding link (the complementary information) can be missing;

- Orientation: graphical representations do not always match the orientation described by the corresponding maplink relation entries in the KGML file;

- *Maplink* relations between different pathway maps: their use sometimes is not so clear and it seems to be incomplete or mismatching.

More in general the graphical representation and the corresponding KGML files are not always equivalent.
In order to clarify the previous issues, we signal the problems to KEGG authors. The reply we got shows that there are reasonable motivations to justify the inconsistencies found. We summarize these below: the relations between different pathway maps may not be clear since there are maps that include other maps or they overlap. Therefore, relations between pathway maps are represented so that the users can refer to other pathway maps in order to get detailed information about the connection with the analized pathway. For these reasons connections between different pathway maps represented by dashed lines are not always represented. Furthermore a metabolic pathway can be connected to another one via compounds or through reactions.

Hence the relations of type maplink are added to each metabolic pathway map only for visual comprehensibility and they are not complete. Moreover, we cannot except to find out all the connections and so the complementary between pathway maps both from a visual depiction as well as in the KGML files.
Everything is due to the fact that data are intrinsecally incomplete and that the dig-

italization of the hand drawn schemas and the translation of biological data requires huge efforts. Accordingly with these reasons, the development of our project is based on the knowledge currently available in KEGG, taking into account such limitations.

Chapter 3

Comparison between metabolism in different organisms

In this chapter, we briefly review the existing methods for the reconstruction and the comparison of metabolic networks and metabolic pathways, which have been developed in the last decade for biochemical applications. These techniques are inter-related each other since metabolic networks comparison implies the reconstruction of the nets themselves from existing data repositories.

3.1 Metabolism representation: State of the Art

In recent years the interest of the scientific community in the development of new methods for metabolic network analysis has grown considerably. This is due to the technological evolution that allows us to deal with complex representations of big data. The methods developed for the metabolic network representation use mainly two mathematical structures:

- **Sets**: with this technique a metabolic network (or a metabolic pathway) can be represented as a set of components that can be enzymes, reactions or compounds. The comparison of such structures is generally based on set operations and it is the simplest one. A variant of this approach may be based on multi-set structures.

- **Graphs**: this representation considers both chemical compounds and their relations. There exist different methods based on simple graphs [13][14], hypergraphs [15][16] and bipartite graphs, such as Petri-nets. Graphs are a more representative structure than the sets but they have a drawback related to the complexity of the comparison between such representations. The computation of graphs or subgraphs isomorphism are known as NP-complete problems.

In this brief review of the literature we take into consideration only proposals based on KEGG information since they are freely available and well structured. We are interested in these particular proposals because also our software is based uniquely on the KEGG database. Moreover we consider only proposal based on graph since our method represents metabolic network as graph.

In [13] the authors propose a method based on graphs in order to provide a rational representation of the metabolic network structure. In order to give the metabolic network representation, they use a directed graph where nodes correspond to compounds, the oriented connections, called *arcs*, represent the irreversible reactions and the non oriented connections, called *edges*, represent reversible reactions. The method uses information taken from KEGG and the network reconstruction is performed using enzymes and reactions information.

Another approach to the representation of the entire metabolic network, has been developed by Markus Rohrschneider [15]. His work intends to provide a visualization of the metabolic network and it is based on the use of hierarchical directed hypergraph with two levels. It makes use of KEGG information. At the first level each node of the hypergraph represents a metabolic pathway of an organism and the hyperedges represent the relations between the pathways themselves (maplink information). Each hypernode is then linked to other nodes that constitutes the second level of the data structure. These ones represent the chemical compounds that specifies enzymes. The compound nodes are in turn connected to each other exploiting the enzymes relations. The structure contains also virtual edges that connect identical compounds in different pathways to allow the user to do interactive operations, like collapse and expansion over the hypergraph.

[16] the authors propose another representation based on direct hypergraphs too, where hypernodes are metabolites and hyperedges are the enzyme-catalized reactions. The aim of this proposal is to demonstrate that metabolic networks contain phylogenetic information analysing the phylogenies obtained from network comparison. An equivalent representation of the hypergraphs is given by bipartite graph where metabolites and reactions represent two different type of vertices.

Zevedei and Schuster [17] propose a solution based on Petri Net where two kinds of nodes are considered: places and transition. Places represent metabolites and transitions represent reactions and enzymes. The static structure of the net is represented by weighted arcs that connect places to transitions and transitions to places. Weights represent the stoichiometric coefficients[1]. From a structural point of view the Petri Net is equivalent to an hypergraph. Beside the structure, PN allows to describe also the dynamic behaviour of a metabolic pathway. Each place (metabolite) can be equipped with tokens, describing the number of molecules of that metabolite, that is, the state of that metabolite. State changing is achieved by the firing of transitions. A transition fires if the number of tokens in the input places is greater or equal to the weights of the corresponding edge. The firing transition produces a new state in the system.

[1] The stoichiometric coefficient indicates how many molecules are needed for the reactions to happen.

3.2 Comparison between metabolic pathways

In the literature we found various approaches to compare metabolic pathways of different species. We found also tools regarding both the reconstruction with a visual representation of the pathways and their comparison. Most of the times, such tools are no longer available because the project or the research work ended. For these reasons, in this section we report and briefly discuss only the available tools that we have found at the date in which we are writing our thesis.

Most of the developed methods give an abstract representation of the information (enzymes, reactions and their relationships) using structures based on Graphs and less frequently with Hypergraphs and Petri Nets. These structures allow for a clear representation of both the components of the metabolic pathway and their relations. Considering graphs, typically we have that nodes represent substrates and edges between two nodes are the enzymes that catalyze the reactions leading to the next node (product). The measure adopted for the comparison of two pathways can be based on the calculation of the dissimilarity between the nodes and on the graph topology, defining in this way a distance measure between the two pathways.

Before discussing the available tools, we report an interesting method for pathways comparison developed by M.Heymans and AK. Singh in 2002. The authors in [18] propose a new analytical technique for metabolic pathways in which, starting from the KEGG data, they reconstruct phylogenetic trees using the structural information contained in the pathways themselves. The aim of their work is to discover new information that may be helpful in understanding evolutionary relationships between different species. In the comparison method we can identify two computational macro-steps:

1. Construction of the enzyme graph: given a pathway P, the enzyme graph $G = (V, E)$ is a directed graph in which the set of nodes V represents the set of enzymes in P, while the set of edges E represents the relationships between the enzymes in P;

2. Pairwise comparison of enzyme graphs: the comparison runs between two graphs

of distinct organisms, producing a distance matrix as a results. The distance matrix is computed starting from the similarity measures between nodes and information on graph structures.

The basic insight of this approach is that two nodes can be considered similar if also their neighbors are similar. Given a pair of nodes, the method considers whether both have the same outgoing and ingoing missing edges and if they have the same mismatches wrt dissimilar nodes. The algorithm for the similarity computation between two graphs $G1$ and $G2$ is divided into four phases. The first phase consists in the similarity computation between each pair of nodes (a, b) where $a \in G1$ and $b \in G2$ are calculated combining both the similarities of the EC numbers and structural information of the graphs. The first similarity measure is computed according to the greatest common prefix of their EC numbers whereas the second similarity measure is calculated by considering both similar and dissimilar edges (ingoing and outgoing). The second phase deals with the construction of a bipartite graph using the similarity scores and looking for the maximum weight matching. Then in the third phase a new similarity score is computed for each matched node found in the previous step. Finally, the similarity between the two graphs is calculated by summing up and normalizing the values of the matches found in the third phase. From this process a distance matrix between the two graphs is obtained which can be used subsequently for the construction of phylogenetic trees using existing tools. The distance is calculated as: $distance = 1 - score$. In this way two identical graphs will have value 0. Unfortunately no tools are available but the intuition behind this method is really interesting.
Now, we introduce a small set of techniques with available tools for comparing metabolic pathways considering the chronological order of publication.

One of the first tool for metabolic pathways comparison was *MetaPathwayHunter* introduced by Pinter [19], in 2005. It makes use of different databases taking information from EcoCyc and SGD. The pathways are modeled by simple graphs where nodes correspond to enzymes that catalyze specific reactions, and the edges connect two nodes if the product of one reaction serves as the substrate of the other. The similarity measure is defined by taking into account both the resemblances between any pair of corresponding nodes and the one between the graphs structures. The topological similarity between graphs is computed by avoiding the problem of subgraph isomorphims and by considering an alternative one. In this case the authors decided to consider multi-source trees, that are directed acyclic graphs whose undirect graphs are trees. The computation of subtree homeomorphism problem is a variant of the problem. In this way, it is possible replace a single enzyme in a pathway with consecutive enzymes into another one, which is a biologicaly reasonable correspondence. The method computes all the optimal and sub-optimal solutions based on a scoring system which measures the distance in the alignments. Moreover, sub-optimal solutions are ranked by statistical significance. Thus, given a query pathway and a collection of pathways, the tool is able find out and report all the approximate matches of the query in the collection. The tool is usable and freely available at the webpage[2] of the author.

Another tool for simple and fast alignment of metabolic pathways is called *MetaPAT* [20]. The tool is based on the Biocyc database and was published in 2007. The metabolic pathways are modeled by using directed graphs in which each vertex represents a metabolite and each edge represents the enzime that catalizes the reaction and leads to the product. Each edge is labeled with an EC number. The algorithm is based on finding a homeomorphism[3] between graphs and, in particular, it tries to solve a combinatorial problem called *Maximum-Score Embedding.* Briefly, given two directed and labeled graphs, $G_P = (V_P, E_P)$ the pattern graph and $G_H = (V_H, E_H)$ the host graph, it finds the maximum-score embedding[4] of a pattern graph G_P into a host graph G_H. Note that this kind of comparison ensures the topological similarity between the graphs but at the same time it is an NP-Hard problem. For this reason

[2] http://www.cs.technion.ac.il/ olegro/metapathwayhunter/

[3] Two graphs G and H are *homeomorfic* if and only if it exists an isomorphism from an edge subdivision of G to an edge subdivision of H.[21]

[4] An *embedding* of a pattern graph G_P into a host graph G_H is a tuple (G'_H, ρ) where G'_H is a subgraph of G_H that is homeomorphic to G_P and ρ is a homeomorphism between G'_H and G_P.[20]

the authors exploit the concept of *local diversity*[5] in order to overcome this issue. The similarity between enzymes is computed by considering the common prefix between their functional EC number. The longer the common prefix, more similar are the enzymes. The tool is freely available at the website[6] of the Friedrich-Schiller-University Jena.

The *SubMAP (Subnetwork Mappings in Aligning of Pathways)* [22] tool, published at the end of 2011, provides a variant of the first method discussed at the beginning of this section and relies on KEGG databases. It is focused on the alignment of metabolic pathways but it uses a method that does not restrict the alignment to one-to-one mappings between the molecules. The idea behind this new approach is that it should be possible to align a single molecule of a pathway with a connected subnetwork into another one. The authors follow the observation that in nature different organisms can perform the same (or similar) functions through different sets of reactions and molecules. Also the number of the molecules and their topology is often different from an organism to another. Thus, given two metabolic pathways P and P' and an upper bound k (a positive integer) on the size of the subnetworks, the aim of this tool is to find the mapping between P and P' with the maximum similarity. The similarity is given by the correspondences of the single molecules in one pathway with the connected subnetworks of size at most k in the other pathway. The SubMAP algorithm performs the following steps:

1. Enumerate the connected subnetwork: create the set of all the connected subnetworks of size at most k for each pathway;

2. Compute the Similarity: it combines both the homological and the topological similarities calculated respectively over reaction sets and subnetworks;

3. Extract subnetwork mappings: since the problem of finding an optimal alignment is an NP-Hard problem, the authors follow an alternative way in order

[5] *Local diversity* property is based on the observation that two paths that have the same starting vertex often carry out very different biological functions [20]. In other words this property allows to characterize a graph exploiting local biochemical diversities.

[6] http://theinf1.informatik.uni-jena.de/metapat/

to avoid this problem. The problem is transformed into a eigenvalue problem. The solution produces alternative mappings in the form of a weighted bipartite graph that is then converted into a weighted graph. The maximum weight of an independent subset of this new graph, corresponds to the maximal alignment ensuring the consistency at the same time.

The tool is usable and freely available at the "The Bioinformatics Lab" website[7].

CoMeta [5] is a prototype tool that implements a method for comparing metabolic pathways of different organisms represented as Petri Nets (PNs). It provides a distance measure considering both homology of reactions and functional aspects of the pathways. Unlike the other methods, the aim of this approach is to take into account also the behavioural aspects (captured by Petri Nets T-invariants) of the pathways by considering the potential dynamic processes information instead of using only static information, like topology or the presence of specific components in the graphs. PNs provide a good representation in modeling metabolic pathways since they are similar wrt the graphical representation used in biology. CoMeta relies on KEGG database for retrieval of metabolic pathways information and the algorithm performs essentially three steps:

1. build the Petri Nets of the corresponding pathways;

2. compute the T-invariants and the similarity measures (the tool offers two different similarity indexes);

3. provide the results as a distance matrix and display the result as a phylogenetic tree with UPMGA or NJ methods.

The structural representation of the pathways is achieved by associating places to metabolites while transitions are associated to chemical reactions. In each place tokens provide information about the number of molecules associated with the metabolite and weights are associated to the input and output arcs of each transition to

[7] http://bioinformatics.cise.ufl.edu/SubMAP.html

represent the stoichiometry of the corresponding reaction. In this way the resulting incidence matrix of Petri Net is equal to the stoichiometric matrix. Then, the T-invariants are computed as a multi-set of transitions. They represent cyclic behaviours in the system and introduce the steady state. The presence of a steady state is natural from a biological point of view because this means that an equilibrium is reached for a set of substances. The distance between corresponding pathways in different organisms is calculated as a weighted combination of two distances. One is called $R - distance$ and it takes into account static information, namely reactions, the other one is called $I - distance$ and it considers behavioural properties expressed by T-invariants. Finally a distance matrix is produced to compare the pathways. CoMeta is freely available at the "Bioinformatics Lab" webpage[8] of the Ca' Foscari university of Venice.

More recently the *MP-Align* tool has been released. It was published in 2014 [23] and it is based on the use of KEGG database information. The proposed method makes use of hypergraphs where the nodes are metabolites, enzymes and compounds while the hyperedges are reactions. The authors decided to make some restrictions. In particular they decided to avoid the representation of ubiquitous substances, they modeled reversible reactions with two corresponding hyperedges, one for the forward reaction and the other for the backward reaction, they distinguished internal and external metabolites by representing external metabolites as input only or output only nodes. The similarity score is based on both compound and enzyme similarities. More precisely, given $R_i = (I_i, E_i, O_i)$ an hyperedge representing a reaction, where I_i is a set of substrates, E_i the enzyme for the catalization of the reaction and O_i the set of products, the similarity for each pair of reactions $R_i = (I_i, E_i, O_i)$ and $R_j = (I_j, E_j, O_j)$ is given by the formula: $SimReact(R_i, R_j) = SimEnz(E_i, E_j) \cdot w_e + SimComp(I_i, I_j) \cdot w_i + SimComp(O_i, O_j) \cdot w_o$. The enzymes similarity (SimEnz) is calculated by comparing the two EC numbers and determining their longest common prefix, the similarity of compounds is computed through an existing tool called *SIMCOMP* [24]. Given two compounds, *SIMCOMP* represents

[8] http://www.dsi.unive.it/ biolab/CoMeta.php

their chemical structure as graphs and provides a measure of their maximal common substructure. The weights w_e, w_i and w_o are used with fixed values in order to give a good balance between enzymes and compounds. The MP-Align algorithm performs essentially six steps:

1. build the hypergraphs of the corresponding pathways;
2. compute the reaction paths: find all the sequences of distinct reactions where the first reaction have a source node;
3. align the reaction paths discovered at the previous step and compute the similarity between reactions through *SimReact* formula;
4. match the reaction paths: find out the most similar paths bethween the two hypergraphs;
5. build the match-frequency matrix M: rows and columns represent hyperedges of the hypergraphs H_1, H_2 and each entry contains the number of times that a specific reaction in H_1 is aligned with a reaction in H_2. After the matrix is built, the best match between reactions is sought;
6. compute the final score and hypergraph alignment: the similarity is calculated by considering both the most similar reactions and the most similar paths. In this way also the topology of the hypergraphs is considered. The alignment of H_1 and H_2 is achieved building a relational graph G that expresses the connections between matched reactions, defining the largest conserved substructure.

The tool is freely available at the website of "Computational Biology and Bioinformatics Research Group"[9] of the Balearic Islands University.

Concluding, we cite *EC2KEGG* [25] a usable command line tool freely available at *sourceforge*[10] website. It relies on KEGG database information and implements a method for comparative analysis and visualization of identified enzymes. The proposed method considers the pathways as sets of enzymes and performs an enrichment

[9] http://bioinfo.uib.es/ recerca/MPAlign/

[10] https://sourceforge.net/projects/ec2kegg/

statistics using the two-tailed Fisher exact test followed by Benjamini and Hochberg correction. Thus, given a specific reference organism and a list of EC numbers, the comparative analysis provide some statistics like shared and unique enzymes, listing the non-mapped enzymes and generating links to visualize pathway maps at KEGG website. The maps are then customized by using different colors in order to underline differences between the analyzed sets of enzymes.

The various approaches presented in this Section are visually summarized by the following table, where we evidentiate three main dimensions. The first one is how pathways are represented, the second one is the source of metabolic data and the last one is the name of the supporting tool, when available.

Publication date	Reference	Representation	Database	Tool
2002	[18]	Graph	KEGG	/
2005	[19]	Graph	EcoCyc and SGD	MPH
2007	[20]	Graph	Biocyc	MetaPAT
2011	[22]	Graph	KEGG	SubMAP
2013	[5]	Petri Net	KEGG	CoMeta
2014	[23]	Hypergraph	KEGG	MP-Align
2014	[25]	Set	KEGG	EC2KEGG

Table 3.1: Summary of the analyzed methods

Chapter 4

Metabolic Networks comparison

4.1 Metabolic network construction

In this chapter we describe how we perform the metabolic network reconstruction starting from the KEGG database information. In particular we explain our algorithm and the data structures used. We propose a comparison method for such metabolic networks which consider both the structure of the network and the similarity between corresponding pathways. The similarity indexes which are computed are also described. Moreover, we discuss some troubles found during the development of the method and the solutions adopted.

4.1.1 Network construction

In the following section we introduce the methodology that allows for reconstructing metabolic networks. Our work relies uniquely on KEGG database information for many reasons. One of these is that the KEGG project has proved to be a reliable knowledge base during the time and it is growing steadily. Another reason is that KEGG provides a digitization of information that are particularly complex. KEGG gives a global metabolic network representation which resumes the metabolisms of all the catalogued organisms. We refer to this data structure as the *reference metabolism*. The net is composed by the union of all the reference pathways thus giving an implicit partition of the the whole metabolism into metabolic pathways which is standardized wrt. all organisms. This choice allows us to analyse and compare specific metabolic functions or entire organisms. However, KEGG data can be incomplete and this leads to issues related to data completeness.

In the literature the majority of the approaches represent the metabolic pathways as graphs of reactions, in order to keep a good level of details, then the metabolic network is obtained by the union of the involved pathway's graphs. The consequence of this approach is that the resulting graph of the metabolism is very huge. The comparison of such large graphs requires to compute some kind of graph isomorphism and it becomes infeasible.

The aim of our thesis is to propose a new method to compare the entire metabolism of different organisms by considering both topology and functionality. We model the metabolic networks using graphs with a certain level of abstraction. This choice simplifies the problem of comparing graphs of big dimensions. We propose the following graph representation.

Let O be a specific organism, then $G_O = (V_O, E_O)$ is the ***metabolic graph*** *of O, where $V_O = \{P_1, \ldots, P_n\}$ is the set of nodes which represent the metabolic pathways of O, namely each P_i, with $i \in [1, n]$, is the $i-th$ pathway represented as a set (or multiset) of reactions, $P_i = \{r_1, \ldots r_m\}$, and E_O is the set of edges that represents the relations between the pathways of O.*

Our representation of metabolism is organized into two levels:

- **Lower level**: it represents a metabolic pathway P_i in terms of set/multiset of chemical reactions;

- **Higher level**: it represents the entire metabolism by a graph G_O, considering the pathways and the relations among them.

This representation fits perfectly the KEGG database organization since each specific pathway P_i, is represented in all the organisms in a standardized way (reference pathway) and the metabolism considers each metabolic function and the interactions defined between them. Our representation guarantees the independence between the two levels, global network level and pathway level. This gives the possibility to use more detailed representations both at pathway level and at network level. At present, we decided to start from the simplest representations leaving more complex representations for future developments.

By adopting a two levels representation, we are able to reduce the size of the graph representing the metabolic network, since nodes represent the pathways rather than the reactions. Hence, comparison between graphs becomes feasible.

4.1.2 Implementation

In order to build the metabolic network of a specific organism, the first step is the data retrieval. We have considered 159 pathways belonging to the following categories in KEGG:

- Carbohydrate metabolism;
- Energy metabolism;
- Lipid metabolism;
- Nucleotide metabolism;
- Amino-acid metabolism;
- Metabolism of other amino-acids;
- Glycan biosynthesis and metabolism;
- Metabolism of cofactors and vitamines;
- Metabolism of Terpenoids and polyketides;
- Biosynthesis of other secondary metabolites;
- Xenobiotics biodegradation and metabolism.

The pathways that are included in the listed categories and that belong to the considered organism O constitute the set of nodes, V_O. KEGG provides KGML files for pathways that include gene/protein network or chemical network, the other pathways are depicted using images. Pathways that don't have a KGML file and use images are not considered in our comparison method. Moreover, we check specific cases in which the KGML files don't contain declaration of chemical reactions. Finally, we consider the set of connections between the pathways themselves and collect such relationship into E_O.

We download through the public KEGG's API, the kgml files for each pathway. The requests are done by using the following URL: `http://rest.kegg.jp/get/org:pathway/kgml` as described in section 2.3.3.3. Then we perform a sequential parsing procedure, reading each KGML file iteratively. In this step the essential information (reactions and maplinks) are extracted for the construction of G_O. In listing 4.1 we can see a fragment of KGML file with some of such data. For convenience we shown only a snippet of code containing the necessary information treated during the parsing phase.

```
<?xml version="1.0"?>
<!DOCTYPE pathway SYSTEM "http://www.kegg.jp/kegg/xml/KGML_v0.7.1_.dtd">
<!-- Creation date: Apr 22, 2016 16:49:05 +0900 (GMT+9) -->
<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>
  <entry id="61" name="hsa:2821" type="gene" reaction="rn:R02740"
       link="http://www.kegg.jp/dbget-bin/www_bget?hsa:2821">
       <graphics name="GPI, AMF, GNPI, NLK, PGI, PHI, SA-36, SA36" fgcolor="#000000" bgcolor="#BFFFBF"
            type="rectangle" x="483" y="265" width="46" height="17"/>
    </entry>
  ...
    <relation entry1="61" entry2="41" type="maplink">
```

```
26         <subtype name="compound" value="90"/>
27     </relation>
28     <relation entry1="41" entry2="56" type="maplink">
29         <subtype name="compound" value="130"/>
30     </relation>
31     ...
32 </pathway>
```

Listing 4.1: Example of maplink information extracted from KGML files

To represent a pathway we extract the set of its reactions. We collect the set of reaction attributes related to the entries of type *gene* (see the lines 14 and 19 in Listing 4.1). In this way we consider only genes belonging to the specific organism without considering orthologs. For the network construction, we extract the tags relation of type *maplink* (rows 25 and 28). The maplink relations have two relevant attributes, *entry1* and *entry2*, that contain the IDs of specific entries (see the lines 25 and 28 in Listing 4.1). Such attributes specify also the orientation of the connection:

- **entry1**: it is the start element of the relation;
- **entry2**: it is the end element of the relation.

By analysing the type attribute of the two entries, we can understand which are the pathways involved. If the entry is of type gene, the corresponding pathway is the pathway tag of the KGML file in analysis, else if the entry is of type *map*, the corresponding pathway is specified by the name attribute of the entry itself.

The information listed in 4.1 is visualized (see the part highlighted in red) in Fig. 4-1.

As an example let us consider the entries with ID 56 and 61 that represent two distinct enzymes which correspond to EC numbers 1.2.1.12 and 5.3.1.9 respectively. The third entry with ID 41, instead, represents the Pentose phosphate pathway. The maplink relation at row 25 in Listing 4.1, connects the enzyme 5.3.1.9 with the Pentose map. The orientation of the connection is given following the order of the entries, as defined before. Since the enzyme 5.3.1.9 constitutes an element of the Glycolysis pathway, an edge from the Glycolysis node to the Pentose phospate node is created.

Our tool allows for representing metabolic network either as directed graphs (i.e. maplink are translated into oriented edges) or as undirected graphs (i.e maplink are represented as undirected arcs). Concerning pathways, the tool offers the possibility to represent the either as set of reactions, or as multiset of reactions (i.e. multiple occurrences of the same reaction are considered).

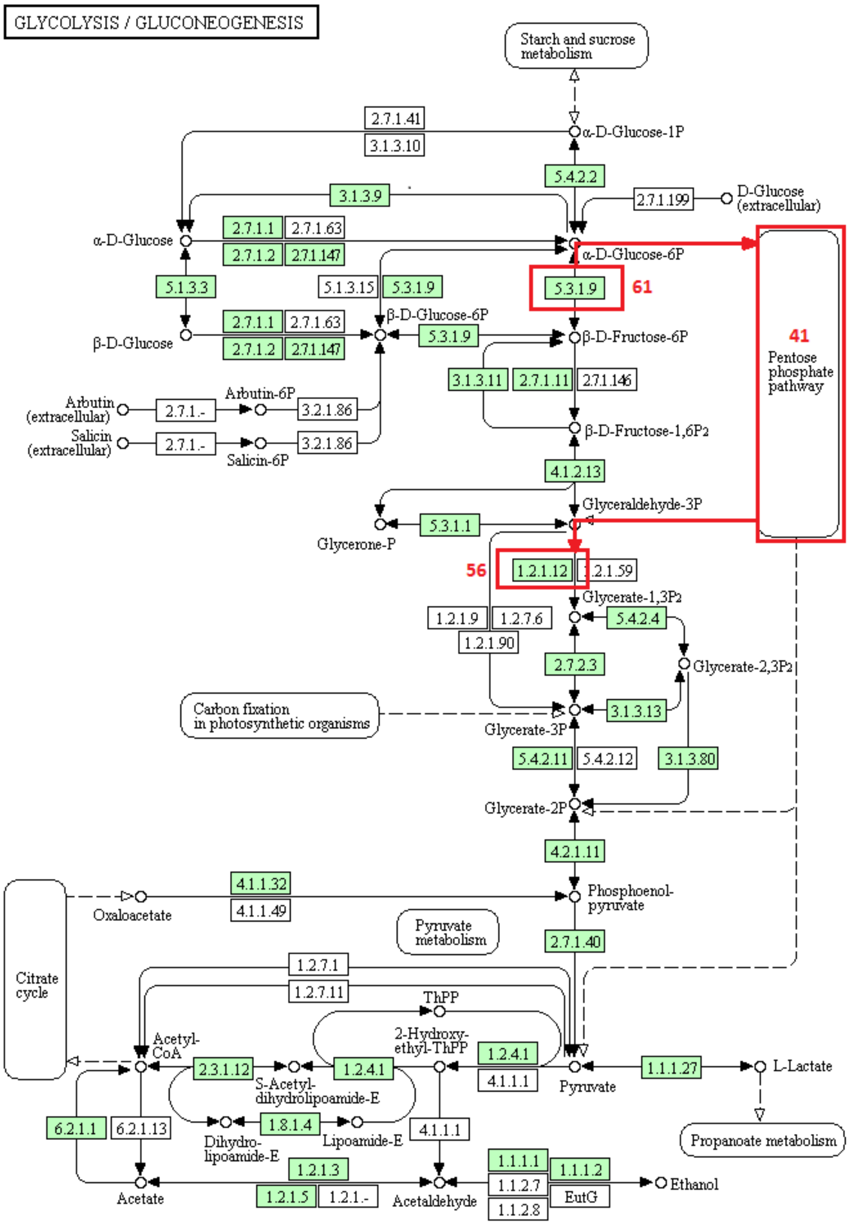


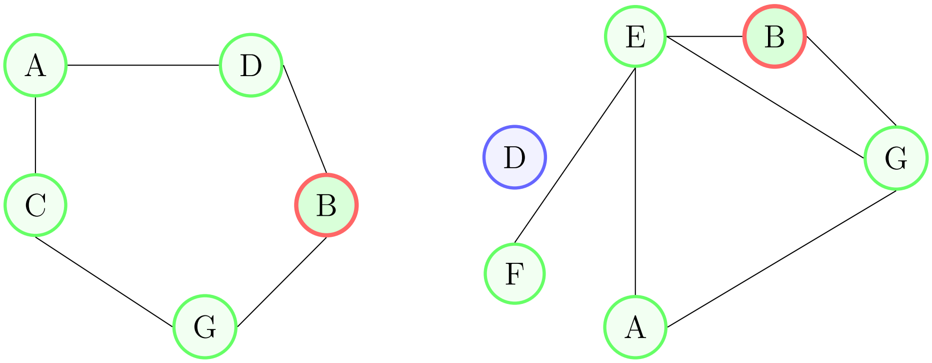
Figure 4-1: Example of maplink detection on Homo Sapiens Glycolysis

4.1.3 Data structures

The implementation of the graph of a metabolic network of an organism is done by using a modified adjacency matrix. We use a square matrix of size n, representing all the pathways listed in Section 4.1.2. This standardized data structure correspond to a mapping between the same metabolic functions (nodes) in different organisms (matrices). A value 1 on the diagonal of an adjacency matrix indicates the presence of a loop in the corresponding graph, but loops are actually never present in our graphs of metabolic networks because edges correspond to KEGG maplinks. Hence a diagonal would be composed only by 0 values. For that reason we exploit the diagonals of the adjacency matrices to represent further information on the nodes of the graphs with the following conventions:

- 1 represents a connected node (pathway);
- 0 represents an isolated node (pathway);
- -1 represents a pathway which is not present in the metabolism of the organism.

This choice allows us to check quickly whether the nodes are connected. 1 values, represent metabolic functions that are connected with at least one other function. 0 values represent metabolic pathways with no connections. The -1 values, indicate that a specific pathway is not present in the metabolism of the organism. The values outside the diagonal represent the edges of the graphs. 0 values represent missing edges and 1 values represent the existing ones. Such matrices represent an abstraction of the reference metabolism given in KEGG. Let us consider a simplified example of metabolic networks of two organisms, O and $O^{'}$. The set of metabolic pathways, in this artificial example is represented by $\{A, B, C, D, E, F, G, H\}$. We show the two graphs and the matrices used in our approach.



(a) Graph of the metabolic network of O

(b) Graph of the metabolic network of O'

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

(a) Matrix of G_O

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

(b) Matrix of $G_{O'}$

We consider the sets of metabolic pathways present in the metabolism of each organism:

$$M(O) = \{A, B, C, D, G\} \quad \text{and} \quad M(O^{'}) = \{A, B, D, E, F, G\}.$$

We use three different colours in the picture: green represents the connected pathways, red represents the metabolic pathways not present in the organism and blue represents the isolated pathways. The correspondence between nodes that represent the same metabolic pathway is given for free since these nodes have the same indexes in the two matrices. This implicit matching is based on KEGG's reference pathways and it allows us to simplify the matrices comparison.

4.2 Comparison of metabolic networks

We present now the technique we propose for comparing metabolic networks of two different organisms. Our comparison method follows our metabolic network representation: it uses a bottom-up approach and it is developed in two distinct levels.

- **Low level**: we perform a comparison between pairs of corresponding pathways in the two organisms. Each metabolic function can be represented either as a set or as a multiset of reactions, depending on the user's choice. We execute a comparison on these sets (multisets) providing a similarity value.

- **High level**: we compare the topologies of the metabolic networks (i.e. their modified adiacency matrices) taking into account also the similarity values computed at the first level. Networks can be modeled as directed or undirected graphs and their comparison produces a similarity value for the entire metabolic networks.

We illustrate now the similarity indexes that allow us to compare two metabolic pathways. After that, we define two different measures on the overall metabolic networks. In order to compute such indexes, we evaluate both the topology and the similarities between corresponding metabolic pathways. Concerning the comparison of the topology of the nets we refer to the definition of the similarity indexes described in [1].

4.2.1 Similarity for metabolic pathway comparison

In this section we describe the similarity indexes between metabolic pathways. The similarity measure depends on the chosen representation. In our case, using a set-based representation (both set or multiset are allowed), the comparison between two pathways consists in finding the number of common elements in terms of reactions. The definition is based on the Jaccard index. Our comparison considers the union of the pathways in the metabolism of the two organisms and it distinguishes different cases. In the first case the $i-th$ pathway is present in only one of the two organisms.

Then, there is the case in which the $i-th$ pathway is present in both the organisms but, no reactions are reported in the corresponding KGML files[1]. Finally, there is the case in which the $i-th$ pathway is present in both the organisms and there are reactions to compare.

Given two different organisms O and $O^{'}$, and a metabolic pathway P_i, we define a measure on how much similar the organisms are wrt. the metabolic function P_i. Let R_i and $R_i^{'}$ be the sets of the reactions of P_i in O and $P_i^{'}$ in $O^{'}$ respectively. We give the definition of the **pathway similarity index** as follows:

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

where P_i represent the $i-th$ pathway in the KEGG order of reference pathways, $|R_i \cap R_i^{'}|$ represents the number of common reactions and $|R_i \cup R_i^{'}|$ represents the number of all reactions belonging to both O and $O^{'}$.

We introduce two distinct global similarity measures based on the definition of $SimP_i$. These measures can be used in order to compute the separated index SI as described in Section 4.2.2. The first global index based on reactions in metabolic pathways called **functional similarity index** is:

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

where $n = |M|$ and M is the union of the metabolic pathways of both O and $O^{'}$. $SimPA$ represents the arithmetic mean of the pathways similarities. The second one, called **weighted functional similarity index**, represents the weighted average of the pathways similarities according to the number of reactions that belong to each

[1] This is the case in wich pathway may involve phisical transormation instead of chemical one.

pathway, and is defined as follows:

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

SimPW permit us to balance the measure with respect to the number of common reactions in the pathways. In particular, a pathway having a few common reactions will have a lower weight. On the contrary, a pathway with a higher number of common reactions will have a higher weight.

4.2.2 Global similarity indexes

We define two different indexes for comparing metabolic networks in different organisms considering both their structure and the similarity of corresponding metabolic functions (pathways). This means to compare the metabolic networks using both $SimP_i$ defined in the previous section and $SimS_i$ defined in [1].

Given P_i and $P_i^{'}$ the $i-th$ pathways corresponding to a specific node in the graphs $G = (V, E)$ and $G^{'} = (V^{'}, E^{'})$, $SimS_i$ provides a similarity value by considering both the numbers of connections and the involved nodes:

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

where $deg(P_i)$ $(deg(P_i^{'}))$ is the degree of the node and E_i, $E_i^{'}$ are the sets of edges of the graphs.

The first global similarity index we define, is called the *Combined Similarity Index* since $SimS_i$ and $SimP_i$ are related to each other. Given two organisms O and $O^{'}$, we define the **combined similarity index** as follows:

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

where $n = |M|$ and M is the union of their metabolic pathways. In order to normalize the index we divide the summation by n, thus the value of CI is in $[0, 1]$.

The second global similarity index is called *Separated Similarity Index* since we introduce an α parameter that allows us to weight the structural similarity index, $SimS$, and the weighted functional similarity index, $SimPW$. $SimS$ represents the similarity measure on the topology of the entire network and it is defined as: $SimS = \frac{\sum_{i=1}^{n} SimS_i}{n}$ where $n = |V \cup V'|$.

We define the **separated similarity index** as follow:

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

where $\alpha \in [0, 1]$. The values assumed by the index SI are in $[0, 1]$. Choosing different values of α allows us to give more relevance either to structural similarity or to pathway similarity. Particular cases are for $\alpha = 0$ or $\alpha = 1$. When $\alpha = 0$ we consider uniquely $SimPW$ and exclude $SimS$, on the contrary, when $\alpha = 1$, we consider uniquely $SimS$ and exclude $SimPW$.

The choice of the global index, either CI or SI, is determined by the context in which the metabolic network comparison is done. If we compare two organisms belonging to the same phylum[2], the topology of their metabolic networks should be almost the same. In this case the use of SI is more suitable, since with $\alpha < 0,5$ we can give more relevance to the comparison of metabolic functions. The use of CI could be more useful when comparing two distant organisms, by considering both the relative topologies and pathways. In Table 4.1 and in Table 4.2 we summarize the local/global similarity indices respectively.

Let us briefly discuss the complexity of the main functions implemented in our

[2] In biology, the phylum is the primary subdivision of a taxonomic kingdom, grouping together all classes of organisms that have the same body plan [26]

Index	Description
$SimP_i = \begin{cases} 0 & \text{if } P_i \text{ is missing in } O \text{ or in } O^{'} \\ 1 & \text{if } P_i \text{ is in } O, O^{'} \text{ but there are no reactions to compare} \\ \frac{\lvert R_i \cap R_i^{'} \rvert}{\lvert R_i \cup R_i^{'} \rvert} & \text{otherwise} \end{cases}$	The **pathway similarity index** considers the union of the metabolic pathways of the organisms, the similarity value of the corresponding pathways is defined in term of reactions.
$SimS_i = \begin{cases} 0 & \text{if } P_i \text{ or } P_i^{'} \text{ is not present} \\ 1 & \text{if } P_i \text{ and } P_i^{'} \text{ are both isolated} \\ \frac{1}{1+deg(P_i)} & \text{if only } P_i^{'} \text{ is isolated} \\ \frac{1}{1+deg(P_i^{'})} & \text{if only } P_i \text{ is isolated} \\ \frac{\lvert E_i \cap E_i^{'} \rvert}{\lvert E_i \cup E_i^{'} \rvert} & \text{if } P_i \text{ and } P_i^{'} \text{ are both connected} \end{cases}$	The **structural similarity index** defines the similarity between two matching nodes in terms of connections

Table 4.1: Summary of the local similarity indexes

tool. The functions used in the comparison procedures are:

- **SetCompare**: this function allows one to compare the reactions of the same metabolic pathway in two different organisms. We store the reactions into HashMap data structures. Usually, basic operations like insertion, deletion and search in such data structure have a constant complexity $O(1)$. In the worst cases they have $O(n)$ complexity. In the simplest cases, the SetCompare returns 0 value if the pathways is missing in one of the two organisms or 1 value if the pathway is present in both organisms but there are no reactions to compare. The more complex case is verified when both pathways contain reactions and thus, when the function returns the ratio between the intersection and the union of the reactions involved in the comparison. The computation of the union is performed using HashSet that correspond to set data structure developed in Java. The complexity of *union* function is $O(m + n)$ where m and n are the number of reactions in the two pathways, since a scan of both HashMaps is required. Each element is added to a set with constant time $O(1)$. The complexity of the *intersection* function is $O(m \cdot n)$ where m and n are the number of reactions in the two pathways respectively. $O(m)$ is the complexity

Index	Description
$SimPA = \frac{\sum_{i=1}^{n} SimP_i}{n}$	The **functional similarity index** is the mean similarity over the union of the pathways of the organisms O and $O^{'}$
$SimPW = \frac{\sum_{i=1}^{n} SimP_i * \lvert R_i \cup R_i^{'} \rvert}{\sum_{i=1}^{n} \lvert R_i \cup R_i^{'} \rvert}$	The **weighted functional similarity index** is the weighted mean similarity over the union of the pathways of O and $O^{'}$ wrt. the number of reactions
$SimS = \frac{\sum_{i=1}^{n} SimS_i}{n}$	The **structural network similarity index** represents the topological similarity of the entire metabolic networks
$CI = \frac{\sum_{i=1}^{n} SimS_i * SimP_i}{n}$	The **combined similarity index** provides a global measure comparing the similarities of both topology and functionalities of the metabolic networks
$SI = \alpha * SimS + (1 - \alpha) * SimPW$	The **separated similarity index** provides a global measure combining with a weight the similarities of both topology and functionalities of the metabolic networks

Table 4.2: Summary of the global similarity indexes

to scan all the elements in the first HashMap and $O(n)$ is the time, in the worst case, for searching the corresponding element in the second HashMap. Considering that the number of reactions for each pathway is on average less that one hundred, the computational complexity becomes reasonable. These operations are repeated for all the pathways of the two organisms. The same considerations can be done for the complexity in using multiset data structures.

- **NetworkCompare**: this function allows one to compare the topology of two metabolic networks. The networks are represented by using square matrices $N * N$, where N is 159, namely the cardinality of the pathways taken from KEGG. The complexity of this function is $O(N^2)$ since all the elements in the matrices need to be inspected. If the metabolisms are represented as undirected graphs, the complexity is $O(\frac{N^2}{2})$ since only half of the matrices need to be considered.

The complexity of the metabolic network comparison, is the sum of the procedures' complexity described above. Thanks to the two-level representation and comparison performed by our approach, the order of all the elements treated is reasonable. This permit us to perform the comparison in a reasonable time respect to the existing method that model the entire metabolism as graph of reactions.

In Chapter 6 we discuss some experiments with the two indexes done in order to validate their application.

Chapter 5

Tool

The goal of our project is to create an application that allows the user to compare whole metabolisms or specific set of metabolic functions of different species. This kind of comparisons are important to find out differences between the metabolic functions of different organisms. The analysis is useful to identify important information that can be used in some branches like drug engineering and medical science. Our application permits the choice of two organisms, performs a fast comparison for which it is possible to select the comparison method on two distinct levels (pathway level and network level) and provides as a result some similarity measures. In this chapter we describe the requirements of the project, the software architecture, the technologies and libraries used and finally we present a brief documentation.

5.1 Requirements analysis

The first step in the development of a software project is the requirement analysis. During this phase we consider the software system requirements as functional ones, which describe the services and the features of the application, and the non functional ones that describe the constrains on the product and the process development.

5.1.1 Functional requirements

Functional requirements permit us to identify functionalities of the software system in terms of services, system reactions under specific inputs and general behaviour of the system. Below we list the functional requirements of our application:

- **Download of the KEGG organisms information**: this functionality should permit the update of the local database with all the information about the organisms;

- **Selection of the comparison of a specific pathway or metabolic network**: the software gives the possibility to compare either the entire metabolism of the organisms or only a subset of the metabolic functions;

- **Selection of the two organisms**: the user should select two organisms from the list of all organisms present in the KEGG database;

- **Download of KGML files**: the application should download automatically the KGML files when they are not already present in the local folders. If the files are already locally present, the user should choose if to update them or to use the existing ones for the comparison;

- **Choice of the comparison methods**: the user must have the possibility to select different methods of comparison, either for the metabolic functions or for the metabolic network;

- **Choice of the α value**: this functionality should allow the user to set a value for the alpha parameter in order to tune the separated similarity index;

- **Automatic exportation of the results as .xls file**: the application must save the computation results in a .xls file for an instant retrieval in a second moment. An .xls file should be saved for each comparison executed by the application;

- **Visualization of data results**: the tool must provide clear and readable results about pathways and networks comparison;

- **Navigation between views**: the user should be able to move back and forward between the windows of the application by using specific buttons;

- **Consecutive comparisons on the same selected organisms**: at the end of an execution, the software must give the possibility to the user to select different comparison methods and execute another run on the same selected organisms.

5.1.2 Non-functional requirements

Non-functional requirements are not directly concerned with the specific services and functionalities defined by functional requirements, but they define constraints on the system or on the development process of the software. They are classified in tree main classes that are:

- **Product Requirements**: they allow us to define constraints on the services offered by the system specifying the usability, efficiency, reliability and portability of the software;

- **Organizational requirements**: they specify process standards, platforms, delivery requirements, etc, to be used;

- **External requirements**: they stem from factors external to the system and its development process (such as the interoperability requirements, legislative, ethical, etc.).

We define a list of non-functional requirements for our application as follows:

- **Fast comparison**: the computation of the similarity indexes must be done in a reasonable time;

- **Parallelized computation**: the software must be developed using threads in order to parallelize the computation as much as possible;

- **Portability**: the application must run on different heterogeneous environments.

5.2 Project architecture

Our project is developed using the MVC (Model-View-Controller) pattern [27]. It is the most used programming pattern to manage software that makes use of GUI (Graphical User Interface). The three main components of the MVC are:

- **Model**: the code collected under this module of the pattern handles data and business logic of the application. In particular, here we find the set of classes that define the context of the application and all the methods that allow the interactions with the databases;

- **View**: the view module collect the set of the GUIs and it is the main responsible for the logic of data presentation. Each view represents the way through which the users interact with the system;

- **Controller**: it reacts to the interactions of the users on the views and it executes the corresponding actions in the model that allow the update of the views.

The three modules interact with each other starting from the main controller that represents the entry point of the application. Then, it initializes the view and it interacts with the model in order to update the view with the data. Every time that a user executes an action through the view, the controller checks the correctness of the inputs blocking bad requests or calling the related methods in the model. In this last case, the execution of the procedure updates the view. The architecture of our application is given in Fig.5-1.

The advantages in using this programming pattern is to achieve a good modularization of the code which gives ease of maintenance, clear separation of tasks during the development process and possibility to work with a certain level of independence on the components to develop. Moreover, the development of additional features and functionalities are possible and made easier thanks to this kind of software architecture.

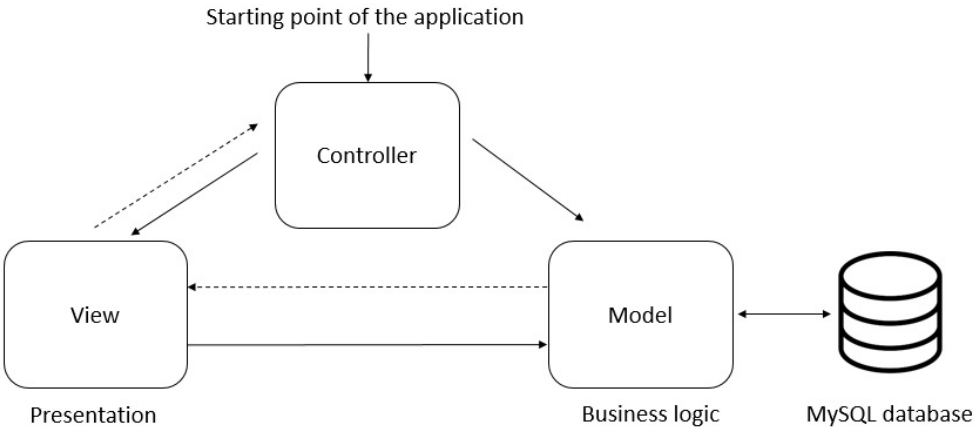


Figure 5-1: Application architecture

The implementation of our tool makes also use of multithreading, a programming technique that allows the execution of distinct threads in a concurrent way in the context of a process[1]. A thread constitutes a part of a process that executes a small part of the code and makes use of shared resources. Thus, processes can be divided in more threads whose execution can be parallelized using the resources of the process itself. Such approach is relevant in our application, since we implement some procedures by dividing the workload in different tasks that can be executed in parallel. Processes like the download of the KGML files and the parsing procedures are implemented as threads. In this way, we increase the performance in the execution of such tasks. Technically, starting from a request made by the user through the GUI of our tool, we create an instance of a thread for each selected organisms whose task is to retrieve all KGML files related to the organism and we store them in an organized structure of folders. In the next step, when the user requires to start the comparison, we create other two threads in order to parallelize the sequential parsing operations for each single file in the corresponding folders.

The tool relies on a MySQL database [28] in which we store all the KEGG organisms information. This choice gives us some advantages. The retrieval of such

[1] A process is an instance of a program that is executed by the CPU. It consists of resources like an image of the code that should be executed, security attributes and the context of a process. Since the CPU handles the processes concurrently, when a process is pre-empted from the CPU, some information must be stored in order to allow a correct resume of the process itself when it comes running again. Such information define the context of the process.

data from our local database allows us to populate dynamically the views in order to support the user during the organisms selection. This task is performed by using dynamic queries instead of multiple connections through the KEGG service. Moreover, assuming the local presence of the KMGL files of the selected organisms, we can use the tool offline. More details about the use of the tool are given in Section 5.4.

5.3 Libraries and technologies

The software was developed using NetBeans IDE [29], a Java-based integrated development environment. It offers an interface to assist developers during coding. The choice to develop the tool in Java is related to the non-functional requirement concerning portability. In this way the application can run in every environment in which the JRE (Java Runtime Environment) is installed, thus ensuring its portability.

For the tool development, we have used external libraries written in Java language:

- **MySQL JDBC Driver** [30]: it allows to create an object for the connection to a MySQL database. In particular, it contains all the methods to perform operations on a specific database like insertions, deletions and data fetching.

- **Guava** [31]: it is an open source library developed by Google company. It contains methods to manage concurrency, I/O operations, string processing and so on. In our case, we use it because it allows to define multi-set structures with all the standard multi-set operations.

- **Poi** [32]: it is a library that belongs to the Apache POI Project and that is developed by the Apache software foundation. It represents the master project for the creation of Microsoft Office documents. This library is used in our application to manage the creation and modification of .xls files.

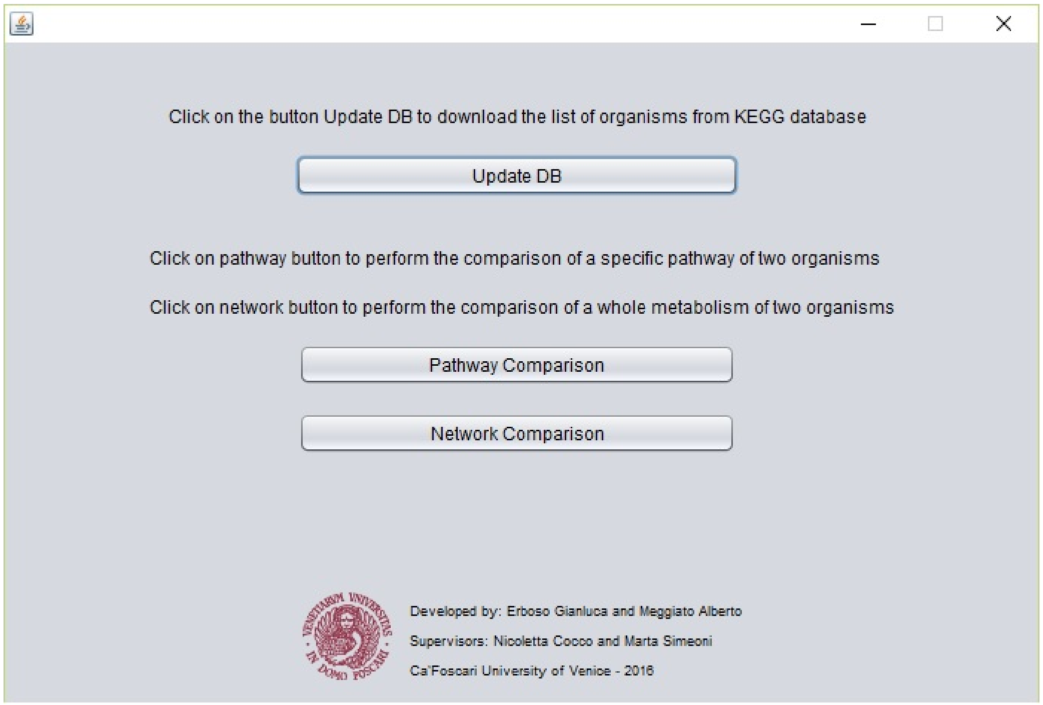
- **SaxParser** [33]: it is a Java library that allows one to perform XML data processing. It is more efficient wrt. a standard DOM parser since it doesn't load the document into memory and it doesn't create a representation of its file. In fact it uses some callback functions to process the XML structure. The main functions are **startDocument()**, **endDocument()**, **startElement()** and **endElement()**. In particular, the lasts two methods are used to inform the client when a specific element tag is open or closed and to fetch all its attributes. The files are scanned sequentially.

- **Seaglass look and feel** [34]: this library is used as an alternative GUI style given by default from Swing Framework. The use of this package gives a better look and feel to the program.

5.4 Documentation

In this section we describe a typical example of use of our tool with the aim to provide a guideline for the users. The tool was thought to guide the user starting from the choice of the comparison to perform, passing through the selection of the organisms to be compared and ending with the selection of the comparison methods.

When we start the application, we see the main view shown in Fig. 5-2.



Click on the button Update DB to download the list of organisms from KEGG database

Click on pathway button to perform the comparison of a specific pathway of two organisms

Click on network button to perform the comparison of a whole metabolism of two organisms

Developed by: Erboso Gianluca and Meggiato Alberto

Supervisors: Nicoletta Cocco and Marta Simeoni

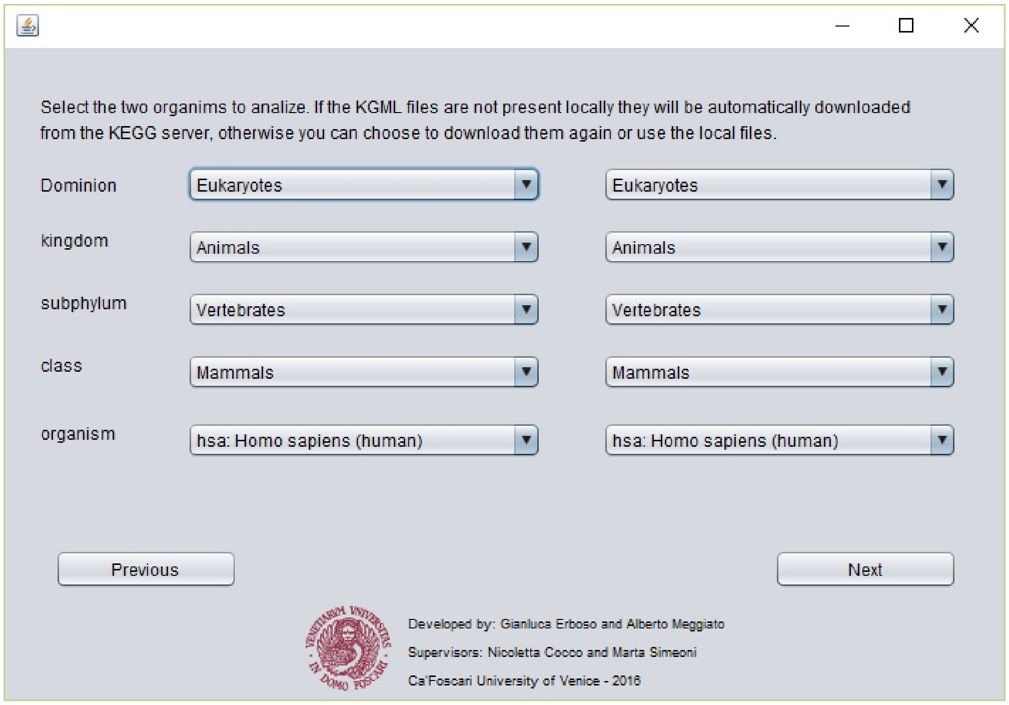
Ca'Foscari University of Venice - 2016

Figure 5-2: Main window of the tool

In this first step the user has three different choices:

- **Update Database**: the main advantage in doing such operation is that it maintains the application up to date, synchronizing the information on catalogued organisms in KEGG. In this way the application is not bound to a specific set of organisms, and it can be used always with the updated information given by KEGG. The updating of the information is a free choice of the user and it is not an automatic procedure;
- **Pathway**: this action allows the comparison of one or more metabolic pathways instead of comparing the entire metabolic networks. The user can select the metabolic function(s) from a predefined list of pathways and then it can select the organisms to compare;
- **Network**: this choice allows the comparison between entire metabolisms of different species.

Below we describe in detail the next steps when the user chooses to compare entire metabolisms.



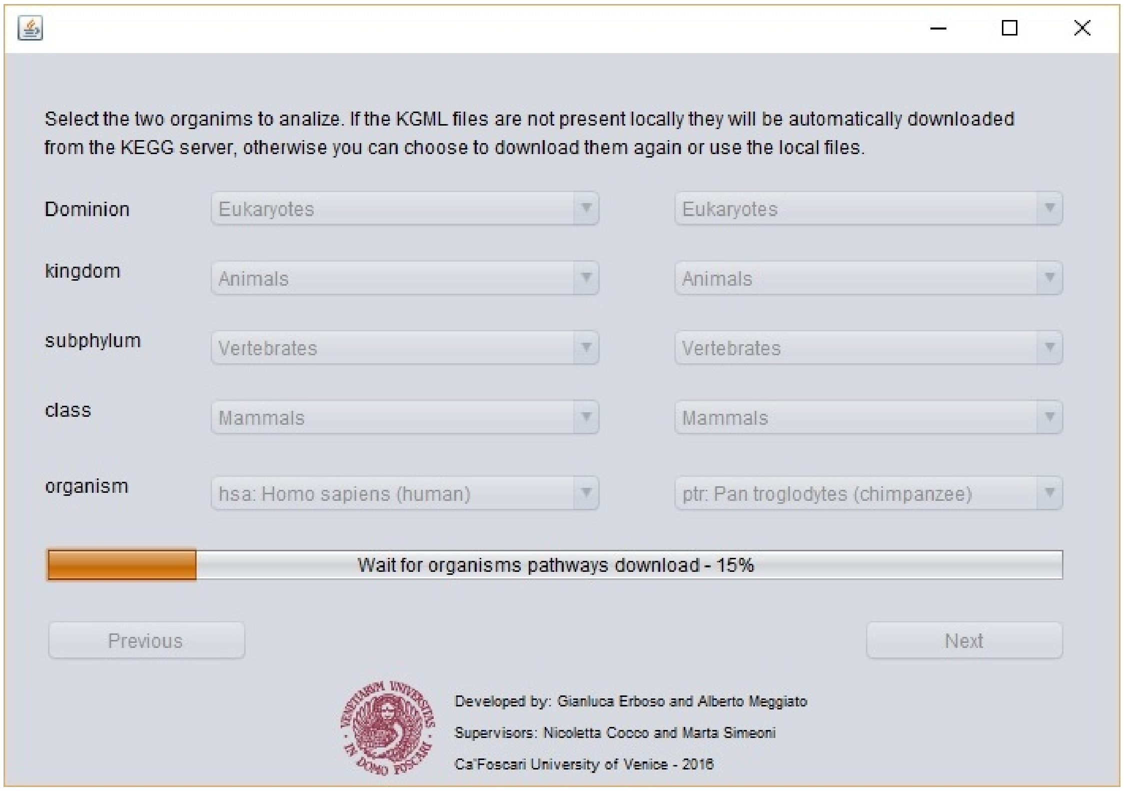
Select the two organims to analize. If the KGML files are not present locally they will be automatically downloaded from the KEGG server, otherwise you can choose to download them again or use the local files.

Dominion	Eukaryotes	Eukaryotes
kingdom	Animals	Animals
subphylum	Vertebrates	Vertebrates
class	Mammals	Mammals
organism	hsa: Homo sapiens (human)	hsa: Homo sapiens (human)

Developed by: Gianluca Erboso and Alberto Meggiato
Supervisors: Nicoletta Cocco and Marta Simeoni
Ca'Foscari University of Venice - 2016

Figure 5-3: Organisms selection

After the choice of the type of comparison, the view shown to the user is given in Fig. 5-3. In this step we give the possibility to select the organisms to compare. Due to the high number of the species catalogued in KEGG, we decided to support the user with a fast selection providing a hierarchical classification of the organisms. The classification is divided by dominion, kingdom, subphylum, class and organism. The population of each pull-down menu is dynamically executed according to the choices performed by the user. Once the organisms selection is done, the user is driven to the next step. Clicking on the next button, the tool checks if the KGML files are already present in the local folders. If the files are present, a new window is shown in order to provide the possibility to download the files again or not. In this way the user can decide either to execute a comparison by keeping the information up to date, wrt. the frequency of the KEGG updates, or to execute the comparison with the existing files. If the selected organisms have never been used in a comparison, the downloading procedure of the KGML files starts automatically before passing to the next window.



Select the two organims to analize. If the KGML files are not present locally they will be automatically downloaded from the KEGG server, otherwise you can choose to download them again or use the local files.

Dominion	Eukaryotes	Eukaryotes
kingdom	Animals	Animals
subphylum	Vertebrates	Vertebrates
class	Mammals	Mammals
organism	hsa: Homo sapiens (human)	ptr: Pan troglodytes (chimpanzee)

Wait for organisms pathways download - 15%

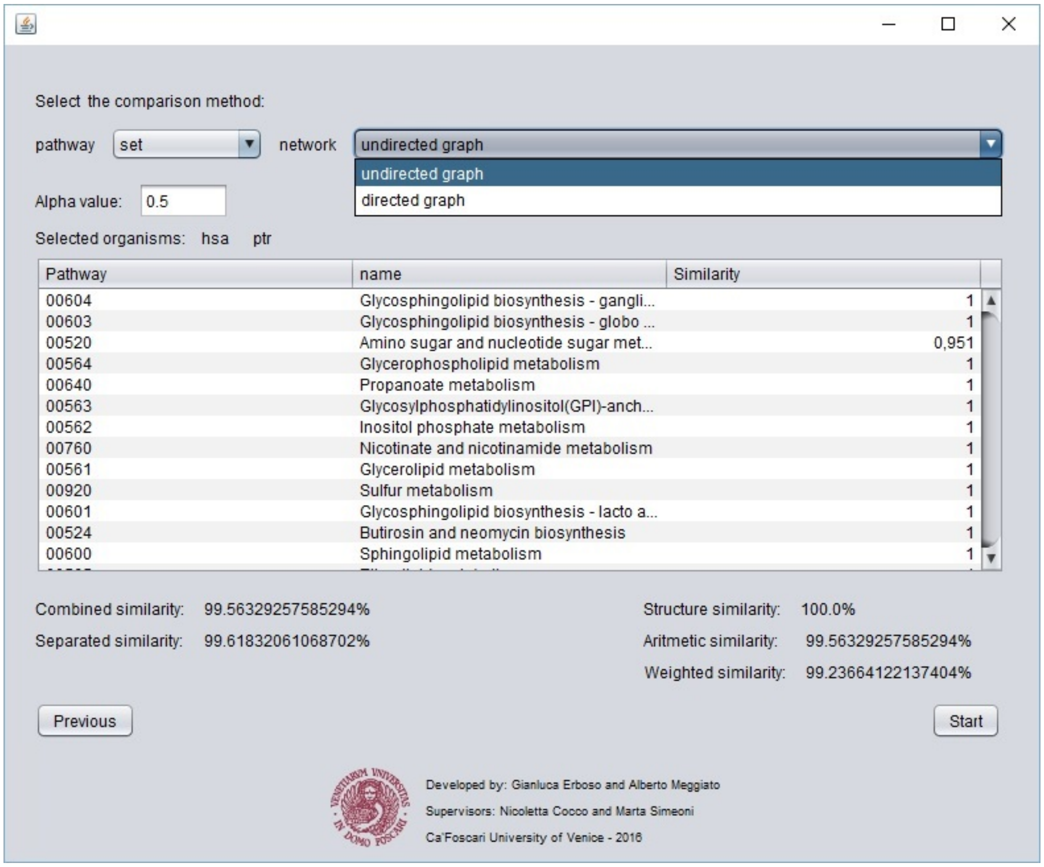
Developed by: Gianluca Erboso and Alberto Meggiato

Supervisors: Nicoletta Cocco and Marta Simeoni

Ca'Foscari University of Venice - 2016

Figure 5-4: Downloading file for selected organisms

The last step is visible in Fig. 5-5. This window represents the core of the application since it permits the choice of the comparison measures both at pathway level and at network level. At pathway level it offers the possibility to compare pathways represented either as sets or as multisets of reactions. At the network level, the comparison methods are based either on directed or on undirected graphs. Moreover the user can set the α parameter in order to associate a weight to the measures involved in the separated similarity index. Setting $\alpha = 0,5$ the same relevance is given to $SimS$ and $SimPW$ indexes. Analogously, setting $\alpha < 0,5$ more significance is given to $SimPW$, while with $\alpha > 0,5$ more significance is given to $SimS$. After the setup of these parameters, the comparison can be launched by clicking on the start button.



Select the comparison method:

pathway: set network: undirected graph

Alpha value: 0.5

Selected organisms: hsa ptr

Pathway	name	Similarity
00604	Glycosphingolipid biosynthesis - gangli...	1
00603	Glycosphingolipid biosynthesis - globo ...	1
00520	Amino sugar and nucleotide sugar met...	0,951
00564	Glycerophospholipid metabolism	1
00640	Propanoate metabolism	1
00563	Glycosylphosphatidylinositol(GPI)-anch...	1
00562	Inositol phosphate metabolism	1
00760	Nicotinate and nicotinamide metabolism	1
00561	Glycerolipid metabolism	1
00920	Sulfur metabolism	1
00601	Glycosphingolipid biosynthesis - lacto a...	1
00524	Butirosin and neomycin biosynthesis	1
00600	Sphingolipid metabolism	1

Combined similarity: 99.56329257585294%

Separated similarity: 99.61832061068702%

Structure similarity: 100.0%

Aritmetic similarity: 99.56329257585294%

Weighted similarity: 99.23664122137404%

Developed by: Gianluca Erboso and Alberto Meggiato

Supervisors: Nicoletta Cocco and Marta Simeoni

Ca'Foscari University of Venice - 2016

Figure 5-5: Results of networks comparison of hsa and ptr organisms with set representation of pathways and the separated index with $\alpha = 0.5$

The computation takes a time that is related to the complexity of the networks and, from our tests, the average execution time is included between 20 and 90 seconds.

Then, the results are displayed on the same window through a summary table of similarities. The table has three columns showing the KEGG pathway number, the relative name and the similarity values computed for the same metabolic function in the two selected organisms. The results are automatically exported as .xsl file in the main tool's folder. After the comparison, further runs can be performed on the same organisms, by selecting a new method and clicking on start again.

Chapter 6

Experimenting with the tool

The comparison of different metabolisms as well as the comparison of metabolic pathways can be useful to discover similarities among organisms. In this chapter we describe the experiments performed with our tool to validate it. This validation is necessary on one hand since it is not possible to compare our results with other proposals in the literature either because their tools are not available or because their network are not based on KEGG's data on the other hand because there is no benchmark on which to perform a data comparison. We use a hierarchical clustering technique in order to provide a results classification and representation.

6.1 Cluster analysis

Clustering analysis is the process of organizing data into groups of observations related to each other. A cluster represents a collection of elements that are similar between them and dissimilar wrt. the elements contained in other clusters. This technique exploits a similarity measure in order to define the concepts of intracluster and intercluster distances. The intracluster measure represents a distance between inner elements of a cluster while the intercluster measure gives the distance wrt. the elements of the other groups. In our case we use a clustering algorithm in order to minimize the intracluster distance (high similarity between cluster's elements) and maximize the intercluster one (low similarity wrt. other cluster's elements). We use a hierarchical clustering, a technique that gives a hierarchical organization of clusters. It produces a set of nested clusters that can be represented by dendrograms. A simple example is given in Fig.6-1.

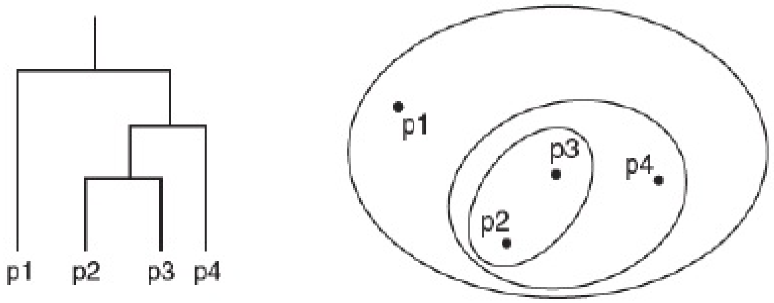


Figure 6-1: Example of nested cluster diagram and the corresponding dendrogram.[35]

There are some advantages in using this kind of algorithms: no assumption on clusters number are needed and moreover they are well suited in taxonomies representation. The hierarchical algorithms may use two different approaches:

- **Agglomerative**: it uses a bottom-up approach in which in the initial phase each element represents a singleton cluster. Then, at each iteration it merges pairs of nearest clusters until a unique cluster is obtained. This implies the use of proximity notion in order to define when two clusters can be merged or not;
- **Divisive**: it uses a top-down approach in which in the initial phase there is a unique cluster containing all the elements. Then, at each iteration it splits a cluster until it reaches the singletons. In this case, the algorithm chooses which cluster to split and how to perform the split [35].

Hierarchical algorithms require a similarity (or distance) matrix as input. Moreover, the key of these procedures is the computation of the proximity measure. Different definitions of proximity provide variants to the algorithms which can be used in specific cases. Examples of measures for proximity are the minimum, maximum or average distances between clusters. Below we give a pseudo-code of a basic agglomerative hierarchical clustering algorithm.

Algorithm 1 Basic agglomerative hierarchical clustering algorithm.

```
Compute the proximity matrix, if necessary.
Repeat:
  Merge the closest two clusters.
  Update the proximity matrix to reflect the proximity between the new cluster and the original cluster.
Until: Only one cluster remains.
```

To measure the distance between clusters we use the *complete linkage* method where the distance between the observations of the two clusters is the maximum one. In order to perform this analysis we exploit an existing implementation of linkage method given in MATLAB software. The *linkage* function and *dendrogram* function are used together to plot the phylogenetic tree. The similarity matrix given as input to the linkage method is created from our tool. Our experiments compare groups of organisms hence clustering techniques are a good way to represent the results. The similarity matrix for applying the clustering, represents the comparison between all possible pairs of such organisms. Our tool is fit to produce such similarity matrix.

6.2 Experiments

We discuss the experiments performed with our tool in order to evaluate the results. We conducted different kinds of experiments considering the entire metabolism of specific sets of organisms, selected by using different criteria. Generally, we use the default configuration of the tool both for pathways and networks representation. Namely we use sets as data structures for metabolic pathways and undirected graphs for metabolic networks and the CI as the similarity index. In the experiments in which we use the SI index, the default value for α is 0.5. Moreover we consider SimPW instead of SimPA since it takes into consideration the number of reactions in the pathways providing a more refined measure. Different configurations are used to perform the experiment 2.

6.2.1 Experiment 1: Metabolic evolution in a group of species

The aim of the first experiment is to verify if the similarities in the metabolism's of a group of organisms find a correspondence in the phylogenesis due to evolution found in the literature [36] [37]. The experiment is executed considering organisms belonging to different taxonomic groups described in the Table 6.1, using the default configuration.

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* (crab-eating macaque)	Animals	Mammals
rno	*Rattus norvegicus* (rat)	Animals	Mammals
fca	*Felis catus* (domestic cat)	Animals	Mammals
gga	*Gallus gallus* (chicken)	Animals	Birds
cmy	*Chelonia mydas* (green sea turtle)	Animals	Reptiles
xla	*Xenopus laevis* (African clawed frog)	Animals	Amphibians
ola	*Oryzias latipes* (Japanese medaka)	Animals	Fishes
crg	*Crassostrea gigas* (Pacific oyster)	Animals	Mollusks
fve	*Fragaria vesca* (woodland strawberry)	Plants	Rose family
pti	*Phaeodactylum tricornutum*	Chromista	Chromalveolata
eco	*Escherichia coli K-12 MG1655*	Bacteria	Proteobacteria

Table 6.1: Group of selected organisms.

What we expect is that our similarity indices produces a classification close to the phylogenetic one. The results of our tool with CI index are shown in Figure 6-2. As we can see, the main groups are clearly separated. There is a clear discrimination between animal's Kingdom and the other ones. Furthermore, in the Animals all the Mammals are grouped together and they are separated from Birds, Reptiles, Fishes and Mollusc. In more details in the Mammals, the distinction between primates and non-primates (*rno*, *fca*) is highlighted. The organisms more distant wrt. Animals (Plants, Protists and Bacteria) are split into another group. Moreover, we note that organisms that perform photosynthesis function are grouped together (*fve* and *pti*).

This experiment shows also some unexpected relations. The *nle* organism should be more similar to the *hsa* wrt. the *mcf* [38]. The same consideration is valid for the *xla* wrt. *ola*. In the last case, from a behavioural point of view, the two organisms have developed the ability to resist at the environmental changes.

We can conclude that our tool with the default setting allows organisms to be grouped together according to main taxonomy.

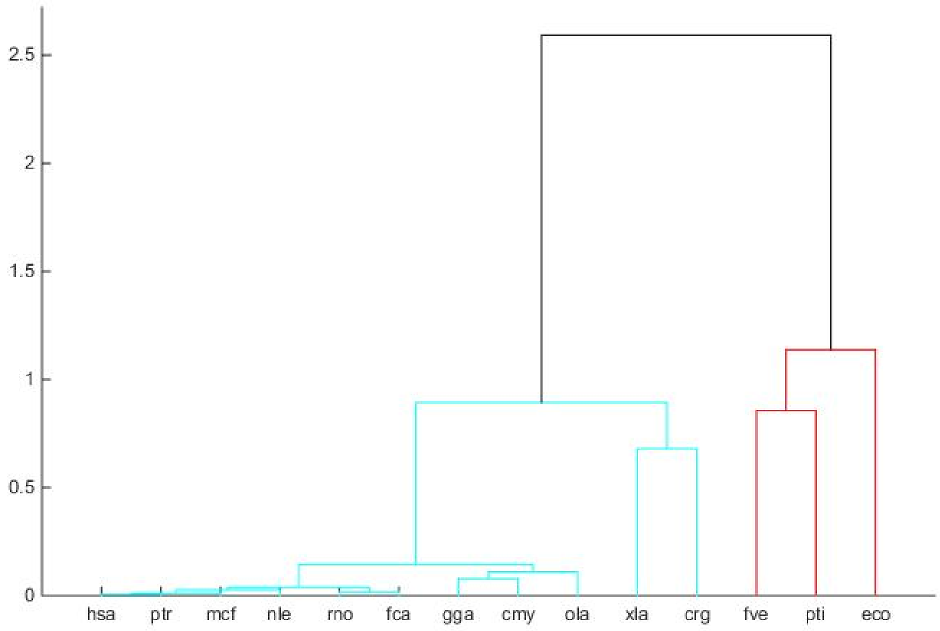


Figure 6-2: Phylogenetic tree produce by clustering with index CI.

6.2.2 Experiment 2: Yeasts and Molds metabolism

The second experiment is more refined since it is meant to test our tool wrt. the classification of a specific group of organisms belonging to the same Kingdom. Differently from the previous one, we select a specific group of organisms of the same Kingdom whose metabolism presents some differences. We select eight organisms among Fungi. The organisms used in the experiment are listed in table 6.2. In particular, we choose four yeasts (*sce, zro, tpf, cal*) and four molds (*fgr, tre, afm, abp*).

Code	**Organism**	**Kingdom**	**Taxonomic group**
sce	*Saccharomyces cerevisiae* (budding yeast)	Fungi	Saccharomycetes
zro	*Zygosaccharomyces rouxii*	Fungi	Saccharomycetes
tpf	*Tetrapisispora phaffii*	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 JB137-S8*	Fungi	Basidiomycetes

Table 6.2: Molds and Yeasts considered in the second experiment.

In this experiment we perform three tests using both the CI index and SI index with different α value, in order to check if differences are detected. In the first experiment we use CI index, in the second one we use SI index with $\alpha = 0.5$ and in the last one SI index with $alpha = 0.2$. The following images Figure 6-3 and Figure 6-4 show the results achieved in the first two cases.

We note that the classification due to the clustering, produces two identical phylogenetic trees. Both the indices produce good results since we have an optimal separation at the top level between Yeasts and Molds, as expected from a phylogenetic point of view. The results with the two indices are different in the distance values as shown by the y axis.

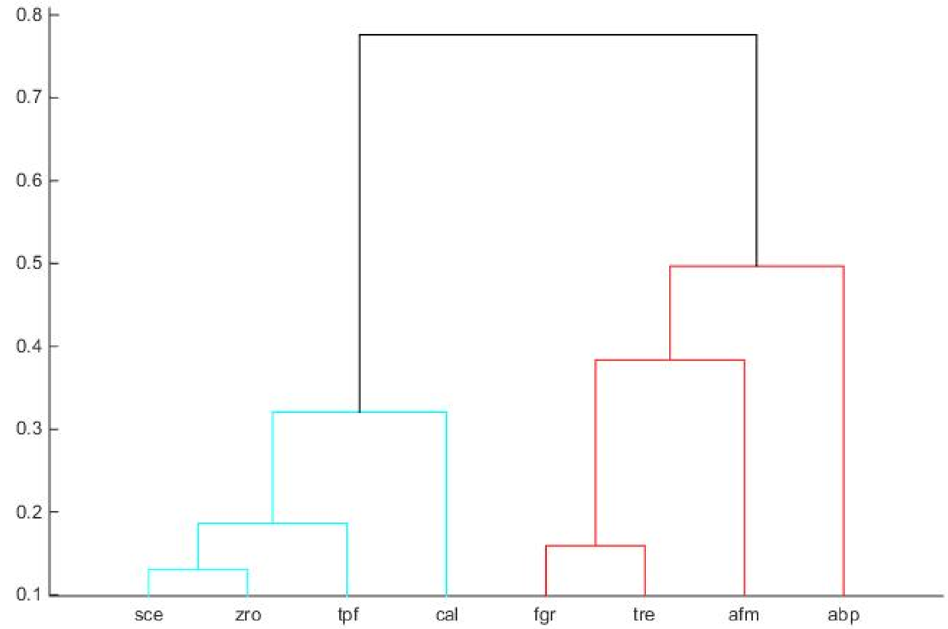


Figure 6-3: Clustering obtained using combined similarity index CI.

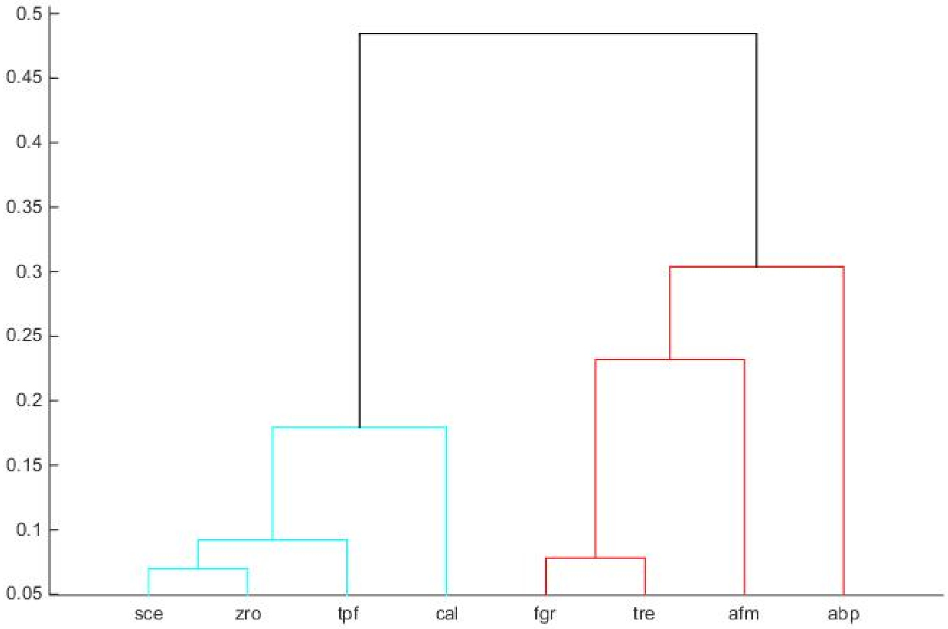


Figure 6-4: Clustering obtained using separated similarity index SI.

On the same group of organisms we consider also the classification due to structural similarity alone (SI with $\alpha = 1$). The results are summarized in Table 6.3. Generally we note that the similarity values are high since they belong to the same Kingdom. For these reasons we perform a further test in which we give different weights to structure and pathways similarities. In particular, we use SI index with $\alpha = 0.2$ in order to give lower weight to the structure (20%). The resulting dendrogram obtained from clustering is shown in Figure 6-5.

	sce	zro	tpf	cal	fgr	tre	afm	abp
sce	1	0,9583	0,9712	0,9365	0,8774	0,8761	0,7925	0,9185
zro	0,9583	1	0,9710	0,9774	0,8777	0,8764	0,8243	0,9183
tpf	0,9712	0,9710	1	0,9491	0,8526	0,8513	0,8007	0,9042
cal	0,9365	0,9774	0,9491	1	0,8991	0,8978	0,8438	0,9326
fgr	0,8774	0,8777	0,8526	0,8991	1	0,9953	0,8832	0,9346
tre	0,8761	0,8764	0,8513	0,8978	0,9953	1	0,8788	0,9335
afm	0,7925	0,8243	0,8007	0,8438	0,8832	0,8788	1	0,8295
abp	0,9185	0,9183	0,9042	0,9326	0,9346	0,9335	0,8295	1

Table 6.3: Structural similarities matrix

From the dendrogram we see that a separation between Yeasts and Molds is performed. However a distortion is introduced in the group of Yeasts. In particular, the *zro* organism is placed distant from *sce*. This is due to the fact that metabolisms structures have lower weights in the comparison. Therefore, we conclude that the structural similarity plays an important role in order to classify correctly the organisms [39].

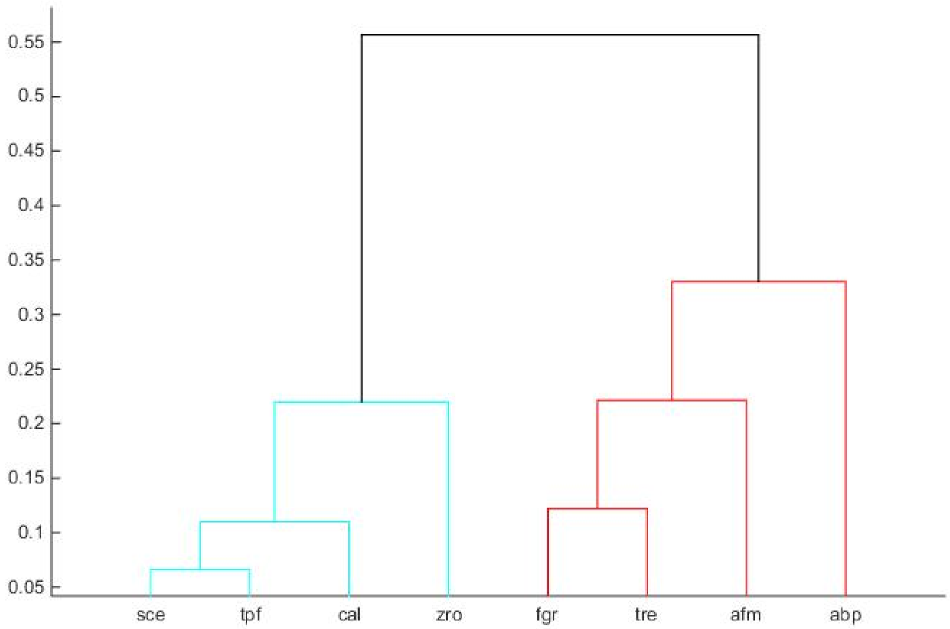


Figure 6-5: Clustering analysis using SI and $\alpha = 0.2$.

6.2.3 Experiment 3: Sulfur metabolism in different Kingdoms

In this experiment we consider specifically the *Sulfur metabolism* pathway (map: 00920 in KEGG). The Sulfur metabolism plays an import role on the amino acid construction, like the cysteine and methionine and other important molecules for the metabolism. Organisms take sulfur in different ways. Plants, Fungi and Bacteria take it and reduce it to sulfide, that is the simplest form of sulfur that can be use for the construction of the amino acids. The other Animals instead, take it indirectly from proteins that they assume through their diet [40].

For this experiment we choose organisms belonging to different Kingdoms considering their behaviour in sulfur reduction. We list the selected organisms in Table 6.4. For this experiment we use $SimP_i$ index in order to compute the similarity of the Sulfur pathway in the two organisms and CI index to analyse their entire metabolism.

The expectations from this test are to obtain high similarity values for organisms belonging to the same Kingdom and low similarity values for organisms of different

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

Table 6.4: Set of considered organisms on *Sulfur metabolism*.

taxonomic groups. The results of the computation are shown in Table 6.5. We obtain expected results, coherent with our previous considerations: higher similarities are reached by the organisms belonging to the same Kingdom while lower similarities are found between organisms of different Kingdoms.

We represent the groups with different colours in the table. As we can see, the Archea group is not well distinguished since the comparison between the *Picrophilus torridus* organism and the other two Archea, produces low similarities. This is due to the fact that Archea considered in our experiment, constitute extreme ecological niches[1]. In particular, *hsl* and *hvo* are associated thanks to the ability to manage/resist to environments with high level of salinity.

For these reasons the metabolism of these Archea can be rather different and the comparison between them can produce low similarities.

We have performed the clustering using the similarity matrix in Table 6.5. The resulting dendrogram in Figure 6-6 shows that the tool provides a good classification

[1] Ecological niches [41] indicate the role, the chemical and the biological properties that permit the existence of an organism within an ecosystem. Extreme niches are organisms that live in extreme environments in which the biological life is constrained by particular conditions. Their survival is given by their adaptability.

	hsa	ecb	gga	tgu	ath	osa	bdi	nfi	ang	cpw	cow	toc	hsl	hvo	pto
hsa	1	1	1	0,7500	0,5385	0,5385	0,5385	0,5833	0,5833	0,5000	0,1000	0,2000	0,0833	0,1667	0,1667
ecb	1	1	1	0,7500	0,5385	0,5385	0,5385	0,5833	0,5833	0,5000	0,1000	0,2000	0,0833	0,1667	0,1667
gga	1	1	1	0,7500	0,5385	0,5385	0,5385	0,5833	0,5833	0,5000	0,1000	0,2000	0,0833	0,1667	0,1667
tgu	0,7500	0,7500	0,7500	1	0,3846	0,3846	0,3846	0,5455	0,5455	0,4545	0,1250	0,2500	0,1000	0,2000	0,2000
ath	0,5385	0,5385	0,5385	0,3846	1	1	1	0,5333	0,5333	0,5714	0,2500	0,3333	0,2143	0,2000	0,2000
osa	0,5385	0,5385	0,5385	0,3846	1	1	1	0,5333	0,5333	0,5714	0,2500	0,3333	0,2143	0,2000	0,2000
bdi	0,5385	0,5385	0,5385	0,3846	1	1	1	0,5333	0,5333	0,5714	0,2500	0,3333	0,2143	0,2000	0,2000
nfi	0,5833	0,5833	0,5833	0,5455	0,5333	0,5333	0,5333	1	1	0,9091	0,1667	0,2500	0,2308	0,3077	0,3077
ang	0,5833	0,5833	0,5833	0,5455	0,5333	0,5333	0,5333	1	1	0,9091	0,1667	0,2500	0,2308	0,3077	0,3077
cpw	0,5000	0,5000	0,5000	0,4545	0,5714	0,5714	0,5714	0,9091	0,9091	1	0,1818	0,2727	0,2500	0,2308	0,2308
cow	0,1000	0,1000	0,1000	0,1250	0,2500	0,2500	0,2500	0,1667	0,1667	0,1818	1	0,7500	0,3333	0,2857	0,1250
toc	0,2000	0,2000	0,2000	0,2500	0,3333	0,3333	0,3333	0,2500	0,2500	0,2727	0,7500	1	0,5000	0,4286	0,2500
hsl	0,0833	0,0833	0,0833	0,1000	0,2143	0,2143	0,2143	0,2308	0,2308	0,2500	0,3333	0,5000	1	0,8333	0,3750
hvo	0,1667	0,1667	0,1667	0,2000	0,2000	0,2000	0,2000	0,3077	0,3077	0,2308	0,2857	0,4286	0,8333	1	0,5000
pto	0,1667	0,1667	0,1667	0,2000	0,2000	0,2000	0,2000	0,3077	0,3077	0,2308	0,1250	0,2500	0,3750	0,5000	1

Table 6.5: Similarity matrix of the sulfur metabolism experiment

of the organisms. As we can see there is a clear distinction between the Kingdoms. Organisms belonging to same Kingdom are grouped together and they are discriminated wrt. the others. At the top level of the tree we find a discrimination between the Bacteria and all the other organisms. At the lower levels instead, Plants and Fungi are separated from Animals.

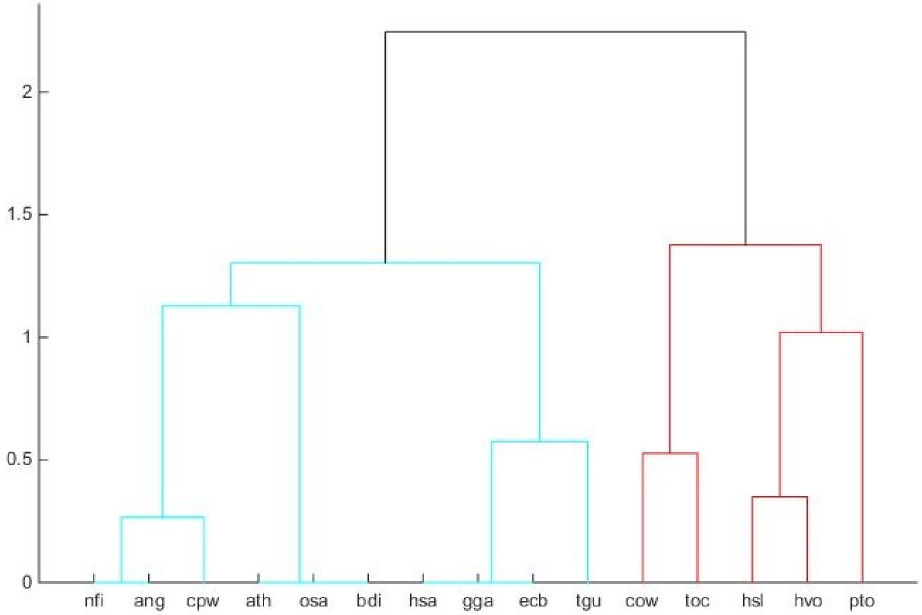


Figure 6-6: Clustering results on Sulfur metabolism.

We perform a further experiment using the same group of organisms and considering the entire metabolisms. The result of this experiment is shown in Figure 6-7. The tree underlines a significant difference in the classification of the organisms. In particular, at the highest level the algorithm provides a discrimination between Animals and all the other organisms. Moreover, Plants and Fungi are separated from Bacteria and Archea.

Considering the *hsa* and *gga* organisms, some differences are present. The analysis of these two organisms, tell us that they are more similar considering only Sulfur metabolism rather than the entire metabolism. Thus, the dimension of the considered dataset is relevant in the comparison.

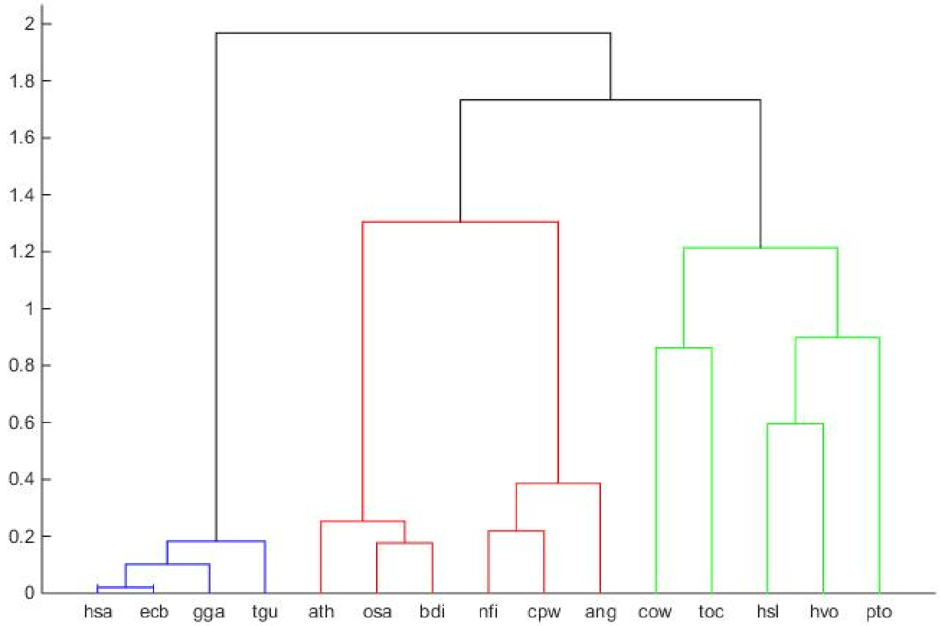


Figure 6-7: Organism classification obtained considering the entire metabolisms in experiments.

6.2.4 Experiment 4: Carbon fixation in photosynthetic organisms

This experiment considers the pathway *Carbon fixation in photosynthetic organisms* (map: 00710 in KEGG). This metabolic function refers to the conversion process of carbon dioxide to organic compound in photosynthetic organisms[2]. Since variants of this metabolic pathway exists due to environmental adaptations, we select a list of organisms that live in different environments. In Table 6.6 we give the organisms selected for the experiments.

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

Table 6.6: Selected organisms for *Carbon fixation* experiment.

The resulting similarity matrix is given in Table 6.7. As we can see, there is a clear separation between organisms that belong to the same Kingdom. Moreover, we note that the *Volvox carteri f. nagariensis* has a low similarity wrt. the other Plants which are coloured in green. The result can be reasonable since we are considering a particular organism, namely a *Green algae.* In general, Green algae should not be considered as Plants due to the fact that they don't have neither roots nor leaves. Furthermore, considering *cre* and *vcn* organisms, differences are related to the multicellular specie (*vcn*) that assume a simplified carbon fixation cycle wrt. to the others.

[2] Organisms that are able to synthesize organic compounds using the sunlight energy.[42]

	gmx	pop	vvi	osa	zma	bdi	cre	vcn	npu	acy	oni	mar
gmx	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
pop	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
vvi	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
osa	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
zma	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
bdi	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
cre	1	1	1	1	1	1	1	0,9524	0,6250	0,5833	0,6250	0,6250
vcn	0,9524	0,9524	0,9524	0,9524	0,9524	0,9524	0,9524	1	0,6522	0,6087	0,6522	0,6522
npu	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	0,6522	1	0,9444	1	1
acy	0,5833	0,5833	0,5833	0,5833	0,5833	0,5833	0,5833	0,6087	0,9444	1	0,9444	0,9444
oni	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	0,6522	1	0,9444	1	1
mar	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	0,6250	0,6522	1	0,9444	1	1

Table 6.7: Resulting similarity matrix in comparing organisms wrt. Carbon fixation.

Clustering produces the classification shown in Figure 6-8, and it is good since it groups correctly Plants and Bacteria at the top level and separates at lower level organisms with differences wrt. the metabolic function in analysis. In Plants we can see that the *Volvox carteri f. nagariensis* presents the problem described before: it is separated from the other Plants according with its simplified carbon fixation cycle.

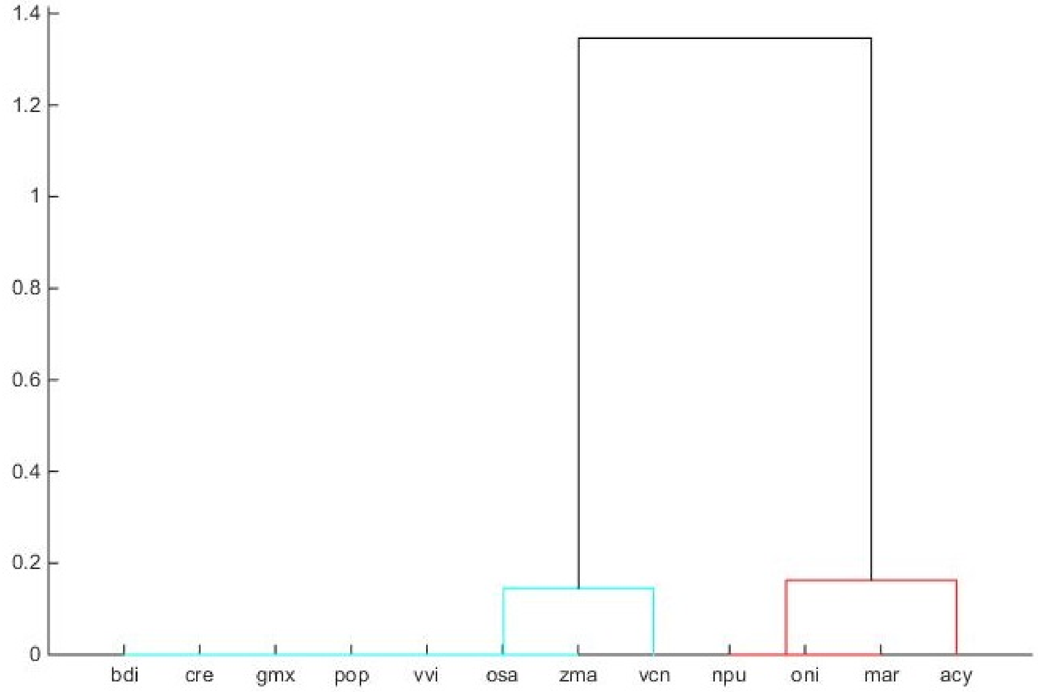


Figure 6-8: Clustering based on Carbon fixation in photosynthetic organisms.

The same group of organisms was compared considering the entire metabolisms. The result given in Figure 6-9 shows a phylogenetic tree in which we have a good separation between Kingdoms. Plants are discriminated from Bacteria at the top level. Then, a separation of Green algae from the other Plants is performed according to the initial considerations. In general, we can conclude that the CI index provides a good classification of the organisms in their Kingdoms. We also note that the *osa* organism is less similar than other plants. This classification can be reasonable because it is the unique plant that lives in highly hydrated environments (paddy field). Other consideration can be done considering Bacteria. *Npu* and *acy* are nitrogen-fixing cyanobacteria, *oni* and *mar*, instead, are cyanobacteria that produce toxins.

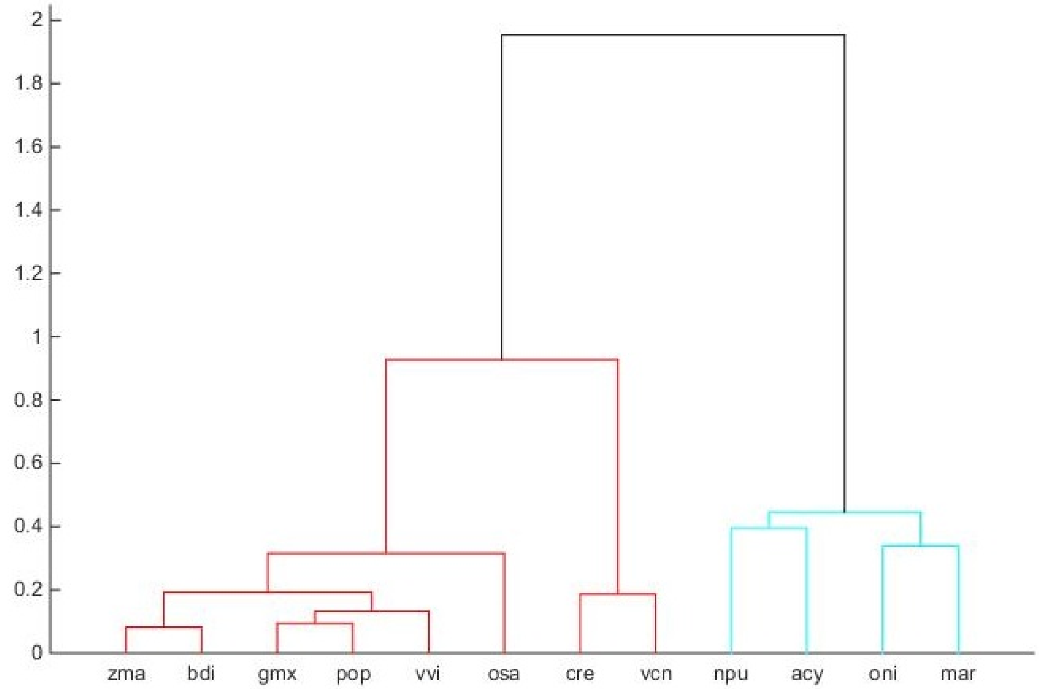


Figure 6-9: Phylogenetic tree from cluster analysis of the entire metabolisms in experiment 6.2.4.

6.2.5 Experiment 5: Glycolysis metabolism

The aim of this experiment is to give a classification of some organisms wrt. the Glycolysis pathway (map:00010 in KEGG). This pathway is responsible to convert the glucose into the pyruvate and during this process it generates energy in the form of ATP. For this experiment we choose a set of organisms which differ wrt. sugar metabolism. We use different configurations in order to perform the experiment. In particular, for the specific pathway analysis we use both set and multiset data structure, and undirected graph for networks. For the global similarity indices we consider both CI and SI with different values of α $(0.25, 0.5, 0.75)$. Below we list the organisms considered for the experiment. They can be divided in four different groups: nitrogen-fixing Bacteria, methanogen Archaea, sulfate-reducing Bacteria, sulfate-reducing Archaea.

Code	**Organism**	**Kingdom**	**Taxonomic group**
dvu	*Desulfovibrio vulgaris Hildenboroug*	Bacteria	Desulfovibrio family
sfu	*Syntrophobacter fumaroxidans*	Bacteria	Syntrophobacter
rsp	*Rhodobacter sphaeroides 2.4.1*	Bacteria	Rhodobacter
cdf	*Peptoclostridium difficile 630*	Bacteria	Peptoclostridium
drm	*Desulfotomaculum reducens*	Bacteria	Desulfotomaculum
ana	*Nostoc sp. PCC 7120*	Bacteria	Nostoc
npu	*Nostoc punctiforme*	Bacteria	Nostoc
tye	*Thermodesulfovibrio yellowstonii*	Bacteria	Thermodesulfovibrio
msi	*Methanobrevibacter smithii*	Archaea	Methanobrevibacter
mel	*Methanobacterium lacus*	Archaea	Methanobacterium
afu	*Archaeoglobus fulgidus DSM 4304*	Archaea	Archaeoglobus
thg	*Thermogladius cellulolyticus*	Archaea	Thermogladius
cma	*Caldivirga maquilingensis*	Archaea	Caldivirga

Table 6.8: Organisms considered for experiment 6.2.5

In particular, *dvu, sfu, drm, tye* are sulfate-reducing Eubacteria, *afu, thg, cma*, are sulfate-reducing Archaeabacteria, *ana, npu, cdf, rsp* are nitrogen-fixing Bacteria and *msi, mel* are methanogen Archaeabacteria. From the test we expect to obtain a good distinction of the above groups.

As we can see, the results in Figure 6-10 give a classification of the organisms with some distortions. In facts, *sfu* and *thg* are placed inside the wrong group. However

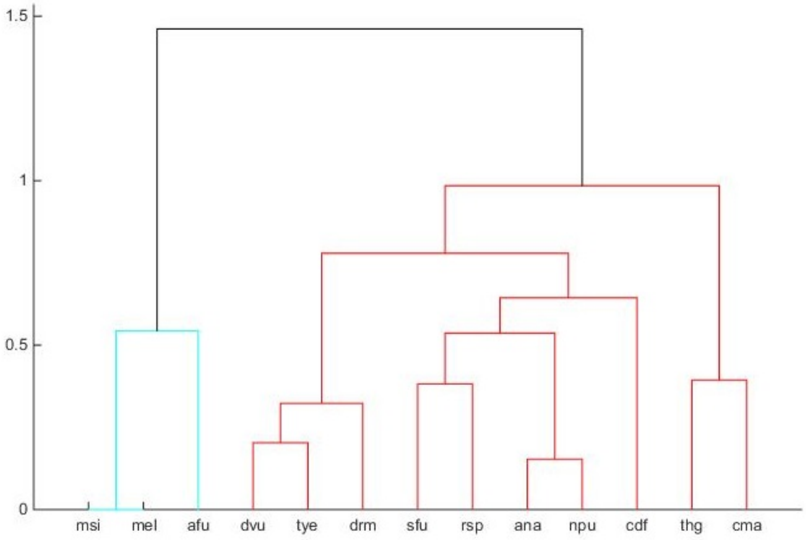


Figure 6-10: Clustering on Glycolysis pathway.

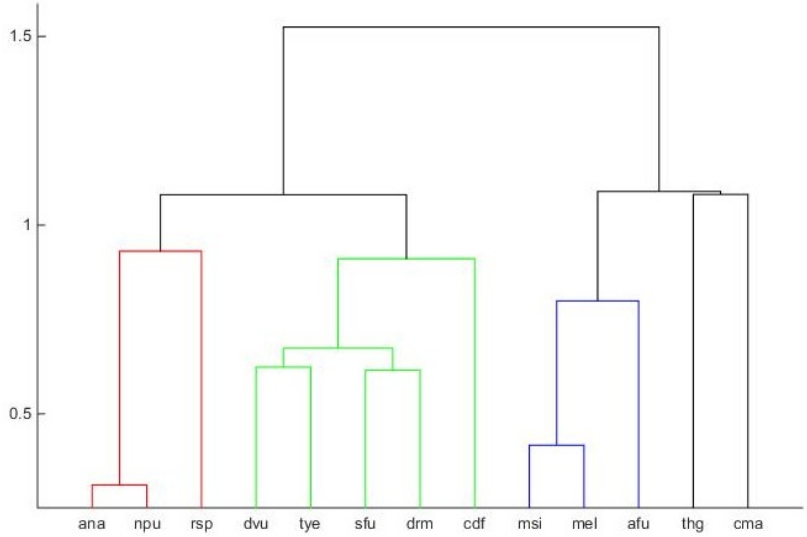


Figure 6-11: Classification based on Bacteria and Archea on the entire metabolism.

these two organisms are grouped correctly at lower level with organisms of the same Kingdom. In particular, in biology, *thg* and *cma* are incline to degrade carbohydrate-based compounds. Considering the results we decided to perform another test on the same group of organisms taking into account the entire metabolism. In this case the result given in Figure 6-11 provides a clear discrimination between Kingdoms.

Analysing the results obtained with the SI index and different α values, the Kingdoms' discrimination is maintained.

6.2.6 Conclusion

Analysing the behaviour of the tool in the experiments, we can make some considerations. First of all, we note that the structure of the metabolic networks is relevant in order to obtain a good classification. In particular, considering only the metabolic networks functionalities, the results present some distortions. This may due since the measure in metabolic pathway comparison assumes a certain level of abstraction. Considering the two global similarity indices, the CI index is in general better than the SI index. However, the SI index permits us to tune the α value in order to weight structure and functionality of the network. Finally, we can conclude that the CI index, in all experiments, provides a good classification between Kingdoms.

Chapter 7

Conclusion

The aim of our thesis is to propose a new approach to compare the entire metabolism between different species considering both topology of the metabolic network and its functionalities. This comparison is useful to discover similarities between organisms providing information about the evolutionary process and supporting medical science activities.

In the literature the proposed techniques try to build and compare the entire metabolic networks in detail. This fact leads to computational problems related to the complexity of the metabolic networks representations. Our proposal is based uniquely on KEGG database information and on the implicit mapping between metabolic pathways represented by the reference pathways in KEGG. Our method is developed on two distinct levels in order to manage the complexity of the networks. In particular we exploits the standardized modularization given in KEGG for representing data.

The proposed comparison method is defined by combining two independent measures. The first one described in [1] evaluates the structural similarity between metabolic networks, the second one instead, is argument of this thesis and defines the similarity between metabolic pathways considering them as sets or multisets of reactions.

We define five similarity indexes: $SimP_i$ that considers the union of the metabolic pathways of the selected organisms and computes the similarity value of the corresponding pathways; $SimPA$ the mean similarity over the union of the pathways of the organisms; $SimPW$ the weighted mean similarity over the union of the pathways wrt. the number of reactions; CI and SI that provide global similarity measures combining the indexes defined in [1] with the above ones.

Our method has been implemented in a Java tool that relies uniquely on KEGG database information. The program allows for comparing the metabolism between pair of organisms selected by the user, and provides different similarity measures. Some experiments have been executed considering both the entire metabolisms and specific metabolic functions on selected sets of organisms. The results are represented in a tree using a hierarchical clustering algorithm. Our analysis of the experiments permit us to conclude that our algorithm is able to classify correctly wrt. the evolution organisms belonging to the same Kingdom. In specific cases, some distortions are verified in comparing organisms of the same Taxonomic group. This is probably due to the level of abstraction of the metabolic pathways.

Further developments of our proposal can be considered. The significance thresholds on the similarities wrt. the Kingdoms or Taxonomic groups, could be determined performing more experiments. Besides, the tool can be extended thanks to its strong modular structure implementing new comparison methods both for networks and pathways. Moreover, new functionalities can be added in order to allow comparison of specific pathways on specific sets of organisms. Again, a clustering algorithm can be integrated in order to provide a cluster analysis and the corresponding phylogenetic tree.

Bibliography

[1] Gianluca Erboso. Comparing metabolic networks at a global level. Master's thesis, Ca'Foscari University of Venice, 6 2016.

[2] Christophe H. Schilling, Stefan Schuster, Bernhard O. Palsson, and Reinhart Heinrich. *Metabolic Pathway Analysis: Basic Concepts and Scientific Applications in the Post-genomic Era.*

[3] Michael Palmer. In *Human Metabolism*, chapter Introduction, pages 1–2. Department of Chemistry, University of Waterloo, 2015.

[4] 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.

[5] Paolo Baldan, Nicoletta Cocco, Andrea Marin, and Marta Simeoni. *Petri nets for modelling metabolic pathways: a survey.* Natural Computing, 9(4):955–989, 2010.

[6] Antonio Albano, Giorgio Ghelli, and Renzo Orsini. Fondamenti di basi di dati. page 8. Zanichelli, 2005.

[7] Kanehisa Laboratories. *Kyoto Encyclopedia of Genes and Genomes.* `http://www.genome.jp/kegg/`, 2015.

[8] SRI International. Biocyc database collection. `http://biocyc.org/`, 2015.

[9] Fellowship for Interpretation of Genomes. Seed. `http://www.theseed.org/`, 2003.

[10] European Molecular Biology Laboratory. The european bioinformatics institute. `https://www.ebi.ac.uk`, 2016.

[11] Hiroyuki Ogata, Susumu Goto, Kazushige Sato, Wataru Fujibuchi, Hidemasa Bono, and Minoru Kanehisa. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Oxford University Press, 28, 2000.

[12] Minoru Kaneisha, Susumu Goto, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanable. *KEGG: Data, information, knowledge and principle: back to metabolism in KEGG*. Nucleic Acids Research, 42, 2014.

[13] Hongwu Ma and An-Ping Zeng. *Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms*. Bioinformatics, 19(2):270–277, 2003.

[14] H. Jeong, B.Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabasi. *The large scale organization of metabolic networks*. Nature, 407:651–654, 2000.

[15] Markus Rohrschneider. *Visualization of Metabolic Networks*. Master's thesis, Universität Leipzig, 2015.

[16] C. V. Forst, C. Flamm, I. L. Hofacker, and P. F. Stadler. *Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation*. BMC Bioinformatics, 7(1):1–11, 2006.

[17] I. Zevedei-Oancea and S. Schuster. *Topological analysis of metabolic networks based on Petri Net theory*. In Silico Biology, 3(3):323–345, 2003.

[18] Maureen Heymans and Ambuj K. Singh. *Deriving phylogenetic trees from the similarity analysis of metabolic pathways*. University of California, 2002. Santa Barbara, CA 93106.

[19] R. Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. *Alignment of Metabolic Pathways*. Bioinformatics, 21(16):3401–3408, 2005.

[20] Sebastian Wernicke and Florian Rasche. *Simple and fast alignment of metabolic pathways by exploiting local diversity.* Bioinformatics, 23(15):1978–1985, 2007.

[21] Jonathan L. Gross and Jay Yellen. Graph theory and its applications (second edition). page 294. CRC Press, 2011.

[22] Ferhat Ay, Manolis Kellis, and Tamer Kahveci. *SubMAP: Aligning Metabolic Pathways with Subnetwork Mappings.* Journal of Computational Biology, 18(3):219–235, 2011.

[23] Ricardo Alberich, Mercè Llabrés, David Sánchez, Marta Simeoni, and Marc Tuduri. *MP-Align: alignment of metabolic pathways.* BMC Systems Biology, 8(1):1–16, 2014.

[24] Kanehisa Laboratories. *SIMCOMP Search.* `http://www.genome.jp/tools/simcomp/`, 2010.

[25] Aleksey Porollo. *EC2KEGG: a command line tool for comparison of metabolic pathways.* Source Code for Biology and Medicine, 9(1):1–4, 2014.

[26] Random House Unabridged Dictionary. Dictionary.com. `http://www.dictionary.com/browse/phylum`, 2016.

[27] Robert Eckstein. Java se application design with mvc. `https://www.oracle.com/technetwork/articles/javase/index-142890.html`, 2007.

[28] Oracle Corporation. Mysql database. `https://www.mysql.it`, 2016.

[29] Sun Microsystems. Netbeans ide. `https://netbeans.org`, 2016.

[30] Oracle Corporation. Mysql jdbc. `https://www.mysql.it/products/connector`, 2016.

[31] Google. Mysql jdbc. `https://github.com/google/guava/wiki`, 2015.

[32] The Apache Software Foundation. Apache poi. `https://poi.apache.org`, 2016.

[33] Oracle Corporation. Sax. http://www.saxproject.org/quickstart.html, 2016.

[34] Kathryn Huxtable. Seaglass look and feel. https://github.com/khuxtable/seaglass/wiki, 2015.

[35] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. 2005.

[36] Mitchell L. G. and Reece J. B. Biology: concepts and connections. chapter 14, 19. Benjamin and Cummings Pub. Comp., 2003.

[37] Arnold M. L. Evolution through genetic exchange. chapter 2, 9. Oxford University Press Inc., 2008.

[38] R. Martin R. Classification of primates. In *The Candbridge Encyclopedia of Human Evolution*. Candbridge University Press, 1994.

[39] Kurtzman C. P. and Robnett C. J. *Phylogenetic relationships among yeasts of the Saccharomyces complex determined from multigene sequence analyses*. FEMS Yeast Research, 4(3):417–432, 2006.

[40] Thomas Leustek, Melinda N. Martin, Julie-Ann Bick, and John P. Davies. Pathways and regulation of sulfur metabolism revealed through molecular and genetic studies. *Annual Review of Plant Physiology and Plant Molecular Biology*, 51(1):141–165, 2000.

[41] Eric R. Pianka. *Evolutionary ecology*. Benjamin Cummings, 6 edition, 1999.

[42] Stefania Azzolini. Treccani.it - enciclopedia della scienza e della tecnica. http://www.treccani.it/enciclopedia/autotrofo_(Enciclopedia-della-Scienza-e-della-Tecnica)/, 2016.