Comparative Analysis of Metabolic Pathways

by

José Carlos Clemente Litrán

submitted to
Japan Advanced Institute of Science and Technology
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Supervisor: Associate Professor Kenji Satou
External Co-supervisor: Associate Professor Gabriel Valiente

School of Knowledge Science
Japan Advanced Institute of Science and Technology

March 2007

Abstract

The recent epidemic growth of diseases related to metabolically abnormal conditions has raised questions about how correct is our understanding of the nature of those diseases, and to what extent can they be explained exclusively in terms of genotypical causes. Additionally, the use of high-throughput methods has produced a significantly larger amount of metabolic data yet to be thoroughly analyzed. In particular, the development of computational methods to understand the metabolic similarities among different species has gained an increased interest as a way to help close the genotype-phenotype gap, i.e., how differences at the genotypical level can produce the greatly diverse phenotypes we observe in nature.

Metabolic pathway alignment is a promising approach to understand the structural similarities of metabolism in various organisms, based on the idea of establishing a correspondance between metabolic reactions in a similar way to how sequence alignment finds a correspondance between nucleotides or amino acids.

In this thesis, we present a new method for metabolic pathway alignment based on the similarity of metabolites, enzymes and reactions present in the pathways. The alignment is constructed by maximizing a similarity score that takes into account both shared and non-shared reactions between the pathways. We also present several applications of our method to problems of biological interest: phylogenetic reconstruction from metabolic similarity, election of model organisms metabolically similar to humans for specific diseases, detection of conserved reactions among a set of organisms and their link to fundamental processes, and identification of possible misannotations of reactions in a metabolic data repository. A web server implementing the phylogenetic reconstruction functionality is also described, together with a standalone distribution of the code.

Our approach has several advantages over previous methods: it relies exclusively on metabolic data, it can assess the relative importance of enzymes and metabolites in the global measure of metabolic similarity, and it is computationally faster. Results presented in this thesis show that our method outperforms previous approaches for phylogenetic reconstruction based on comparison of metabolism. Furthermore, we show how filtering of noisy data, use of complete metabolic information and fuzzy clustering can produce robust, highly accurate phylogenies.

Acknowledgments

I would like to thank my advisers, Professors Kenji Satou (JAIST) and Gabriel Valiente (Technical University of Catalonia, UPC), for their guidance through this thesis. Professor Satou provided the means and the support to achieve the goals I aimed at during my Ph.D. He was kind enough to accept me as a student in the first place, and then patient to let me explore research ideas that probably seemed quite unfeasible when we first discussed them. Professor Valiente has been the adviser any researcher would wish for, and I feel words cannot trully express my gratefulness. Despite sleepless nights working on his requests for results, working with him has been both intelectually exciting and good fun. Their continuous advice and suggestions have been fundamental to achieve the results here presented.

I feel in debt with my former adviser, Professor Akihiko Konagaya (RIKEN), who helped me to come to JAIST as a master student, and provided support during my early years of research.

I am extremely grateful to Professors Xavier Defago and Kentaro Torisawa (JAIST), who taught me a great deal on how to do scientific research. Professor Tu-Bao Ho (JAIST) was always open to hear and discuss new ideas, no matter how busy he was.

Professors Yoshiteru Nakamori and Takashi Hashimoto (JAIST), as members of the tribunal in charge of evaluating this thesis, provided insightful comments on how to improve and expand it further.

Several people helped with their comments to previous oral presentations of this work: Professors Takashi Gojobori and Kazuho Ikeo (National Institute of Genetics), Professor Satoru Miyano (The University of Tokyo), and Oliver Ebenhöh (Humboldt University). I would especially like to thank Professors Susumu Goto and Minoru Kanehisa (Kyoto University) for their extensive suggestions during my presentation at the Kyoto University Bioinformatics Center.

My gratitude goes as well to everybody at our laboratory for their help and collaboration: Professor Tomoyuki Yamamoto, Masanori Higashihara, and Dang Hung Tran. Our secretary, Kyoko Hirukawa, made my own work much easier by managing efficiently all the burocratic aspects of research and conference trips. I am also particularly grateful to my former colleague Tho Hoam Phan for his comments and help early on my Ph.D.

I would like to think that, many years from now on, Stijn de Saeger will still be the great friend he has been for me during this time. All those afternoons by the coffee machines plotting for world domination are among the best memories I have, and we even managed to discuss serious science from time to time.

Carlos and Reme were the closest thing to a family I had in JAIST, their support and help made me feel less homesick despite the distance.

Minoru Hashimoto and Shinichiro Omura made me feel at home in Kanazawa, and I should probably blame them more than anybody else for making it so difficult to leave now. Talking with them all night until sunrise was not only great fun, but they also taught me more about this country than any book I could have read.

Shizuka Uchida has been a great friend and colleague, and I have never thanked him as he deserves for all what he has done for me.

Being so far away from home has probably been tougher for my family than it was for me. Nevertheless, they always encouraged me to follow my career in science, and were supportive in those moments when I felt not so convinced about the path I had chosen.

Finally, I owe to Rie her patience with me despite my frequent moods. Her care and love were at the end what gave me the strength to complete this work.

To my parents, who taught me the value of learning

Contents

List of Figures

List of Tables

Chapter 1

Introduction

> *Make no little plans; they have no magic to stir men's blood and probably will themselves not be realized. Make big plans; aim high in hope and work, remembering that a noble, logical diagram once recorded will not die, but long after we are gone will be a living thing, asserting itself with ever growing insistency. Remember that our sons and grandsons are going to do things that would stagger us. Let your watchword be order and your beacon beauty. Think big*
>
> Daniel Burnham [100]

1.1 Motivation

The study of *metabolism*, the process through which living organisms transform nutrients into energy and biomolecules usable for fundamental cell processes, has gained interest in recent years due to the epidemic growth in diseases related to metabolically abnormal conditions [3]. Metabolic processes are usually organized into *metabolic pathways*[1], a series of chemical reactions catalyzed by enzymes that take certain input metabolites and transform them into a set of output metabolites. The comparative analysis of metabolic pathways in different organisms can yield important clues on their evolution [1], on differences in drug targets [39], on new therapeutic strategies [107], or on identification of alternative enzymes [15]. Indeed, there has been a renewed interest in the study of metabolic pathways and their properties, both structural and dynamical [69, 123, 126, 150].

The comparative study of metabolism can also provide interesting clues about similarities and differences at the phenotypical level, and further close the genotype-phenotype gap [43, 65]. Clearly, one of the great challenges in computational biology is to comprehend how complex phenotype traits and cellular functions can emerge from simple, linear nucleotide sequences. It has been shown that genomic differences among species are smaller than they were expected to be [22], with phenotipically different species such as human and chimpanzee sharing up to 99% of their gene sequences [20, 37]. Nevertheless, as we move up from genotype to phenotype, differences among organisms will gradually increase. For instance, recent results show that most human-chimpanzee differences are due to gene regulation rather than due to gene sequence [50, 71]. Furthermore,

[1] Through this work, the terms *metabolic pathway* and *metabolic network* will be used indistinctly

the number of different proteins between humans and chimpanzees are nearly 80% [51]. By studying levels of higher biological complexity, we expect to gain even more clear insights on what are the relevant differences among organisms.

Sequence data has been extensively researched and a variety of tools developed for its study. Methods for sequence alignment are a well-known example of such tools [5, 63, 89, 90]. As the number of fully sequenced genomes increases and available biochemical information on different organisms grows, we also need to develop new approaches for the analysis of such data. Specifically, the comparison of metabolic processes by alignment algorithms is still an open research question which has not yet been properly answered.

We therefore argue that there is a need to develop methods for non-genomic comparison of organisms to better understand their differences. In order to address this problem, in this thesis we will focus on metabolic similarity, and propose *a method for metabolic pathway alignment based on a new measure of structural similarity among pathways in different species.* We will then apply this method to different biological problems to prove its usefulness.

1.2 Structural analysis of metabolic pathways

Understanding the organization of metabolic pathways has been identified as one of the major challenges in genome research [27]. Structural analysis of networks can be addressed from two different perspectives: either by asking whether a network shows some properties of interest (is it connected? is it dense? what is the average path length?), or by comparing how similar is the network to others under some definition of similarity. The first approach is exemplified by studies on scale-free properties of biological networks [4, 7, 136], while the second could be represented by alignment techniques ([45, 62]), where two metabolic networks are compared by establishing a scoring function over the quality of the alignment.

The pathway alignment method that will be described in this work considers metabolic reactions as the basic elements of the alignment. Our algorithm defines first the similarity of any two given reactions as a function of the similarity of the set of enzymes and metabolites present in the reactions. Metabolites can be either similar or dissimilar, while enzymes can be compared according to their functional similarity. A complete alignment is obtained by maximizing the sum of similarity scores of all reactions in the pathways being aligned. This approach can be therefore thought of as a generalization of sequence alignment, where instead using substitution matrices [59] based on evolutionary rates of changes of the residues, a functional definition of similarity is used.

1.3 Contributions of this thesis

In this thesis, we have achieved three main contributions. First, we introduce a *new method for metabolic pathway alignment* based on a measure of similarity of enzymes, compounds and reactions involved in the pathway. This method has several advantages over previous approaches: there is no need to introduce a penalty score for missing reactions, we do not use sequence information, we can establish the relative importance of enzymes and metabolites in the global metabolic similarity measure, and our approach is computationally faster than graph-based methods.

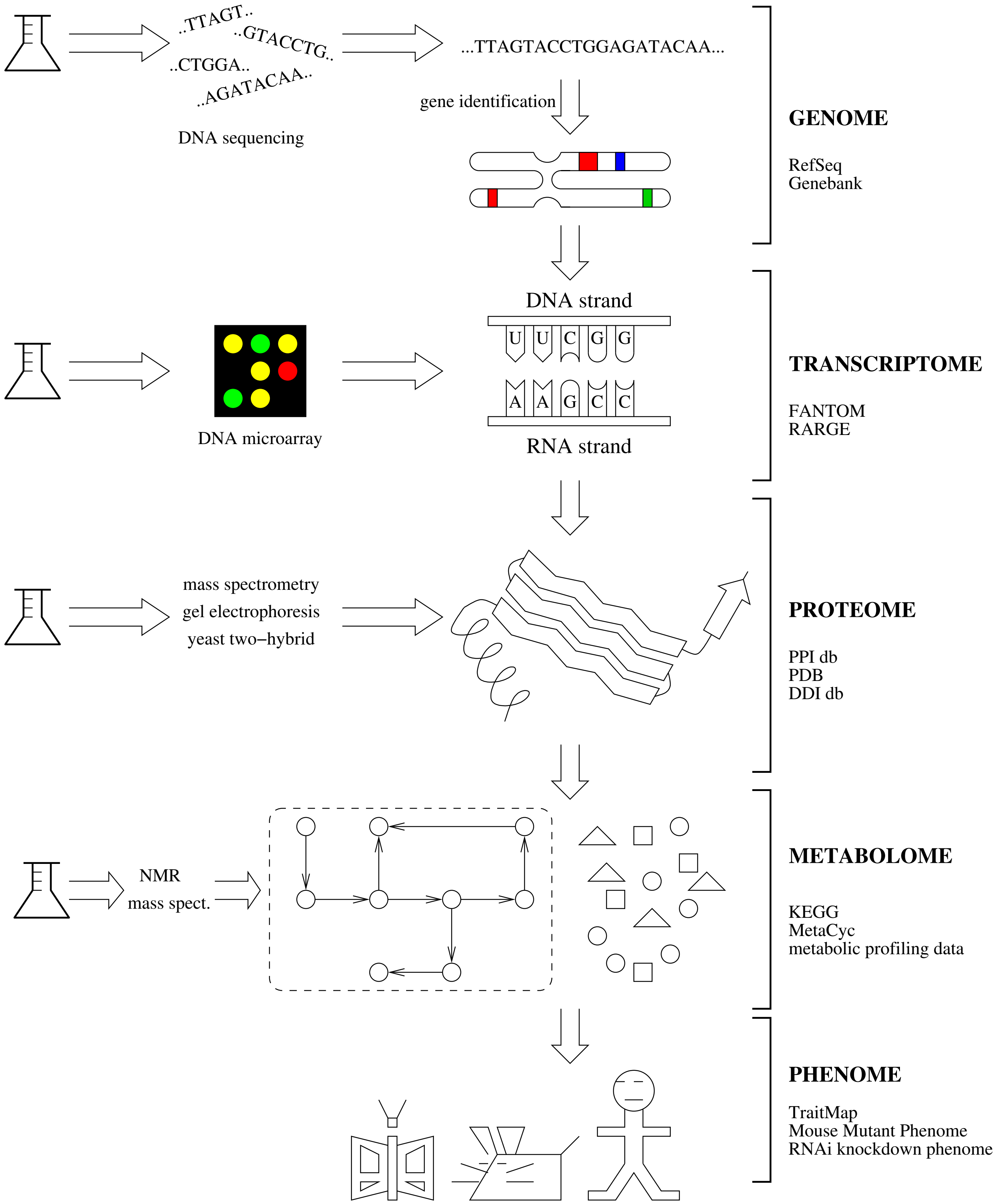


Figure 1.1: From genotype to phenotype

Second, several *applications of this method to problems of biological interest* are described, showing how our method yields better results than previous approaches.

Finally, we present a *web server to reconstruct phylogenetic relationships* among a set of organisms by using their metabolic similarity as calculated by our algorithm. This server can be useful for bioinformaticians interested in phylogenetics and metabolism evolution. A series of Perl modules implementing our algorithm are also discussed.

1.4 Thesis organization

The rest of this thesis is organized as follows:

i. Chapter 2 introduces some basic concepts on metabolism that are considered relevant to understand the biological context in which we will be working This chapter also introduces the Kyoto Encyclopedia of Genes and Genomes (KEGG), the primary resource of metabolic information we utilized in this thesis, as well as some terminology commonly used. Finally, the chapter reviews previous methods for metabolic pathway alignment.

ii. Chapter 3 presents in detail our method for metabolic pathway alignment. We start by describing how will we represent metabolism, and then introduce an algorithm to calculate the alignment between pathways corresponding to a maximum metabolic similarity score. A brief description of the main differences between our method and previous approaches is presented at the end of the chapter.

iii. Chapter 4 presents four applications of this method to biological problems of relevance. First, reconstruction of phylogenetic relationships among a group of organisms is addressed, with the underlying hypothesis that species with similar evolutionary history should have similar metabolic processes. Second, we study the conserved reactions among a set of organisms. In this case, given a set of species belonging to a certain taxonomic unit, we try to find out what reactions are being conserved in a majority of the species and how these reactions are fundamental for the set of organisms under study. Third, we use our algorithm for detecting reactions that might have been incorrectly annotated to certain bacteria in a repository of metabolic data. Finally, we present a possible application of our approach to choose candidate model organisms for clinical trials based on their metabolic similarity with humans under specific conditions.

iv. Chapter 5 presents a web server for phylogenetic reconstruction based on our algorithm for pathway alignment. This chapter also describes a set of Perl modules we developed implementing our metabolic pathway similarity algorithms and providing data structures to facilitate the manipulation of metabolic information.

v. Chapter 6 summarizes this thesis, presents the main contributions and sketches future directions of research.

vi. Appendix A details the computational complexity of our method both for worst- and average-case.

vii. Appendix B contains detailed information on results presented in Chapter 4 related to possible misannotations in the KEGG repository.

viii. Appendix C includes the algorithms implementing our similarity measure method as described in Chapter 3.

ix. Finally, Appendix D presents an additional method for metabolic pathway alignment based on contextual similarity, and compares it to the one presented in Chapter 3, discussing advantages and problems of both approaches.

Chapter 2

Background

> *For he who understands his subject is master of his end; and every workman is king over his work*
>
> Francis Bacon [10]

2.1 An introduction to metabolism

2.1.1 Overview

Metabolism is usually defined as the process through which living organisms acquire and use energy to perform different activities [133]. Metabolism has four main functions [85]:

i. to obtain chemical energy from the degradation of nutrients.

ii. to convert nutrients into the precursors (*building-blocks*) of cell macromolecules.

iii. to assemble these building-blocks into cell components, such as proteins, nucleic acids, lipids and polysaccharides.

iv. to form and degrade biomolecules required in specific functions of cells.

According to their metabolism, we can classify organisms into *autotrophs* or *heterotrophs*. The first ones synthesize all their cellular biomolecules from simple molecules such as H_2O, CO_2, NH_3, and H_2S. These organisms can be further divided into *chemolithotrophs*, which obtain energy through the oxidation of inorganic compounds such as NH_3, H_2S, or Fe^{2+} (for instance, cyanobacteria), and *photoautotrophs*, which obtain their energy through photosynthesis (photosynthetic bacteria and green leaf cells of plants). Heterotrophs, on the other hand, cannot use atmospheric carbon dioxide and obtain free energy through the oxidation of organic compounds obtained from their environment. Ultimately, heterotrophs depend on autotrophs to obtain such substances (see Figure 2.1). Heterotrophs can be classified as *aerobes*, which use molecular oxygen to oxidize their nutrient molecules, *anaerobes*, which can degrade their nutrients without using oxygen, and *facultative*, which can be either aerobes or anaerobes depending on the presence or absence of oxygen.

Living organisms also need a source of nitrogen to synthesize the building blocks of proteins and nucleic acids. Organisms differ in the chemical form of nitrogen they can

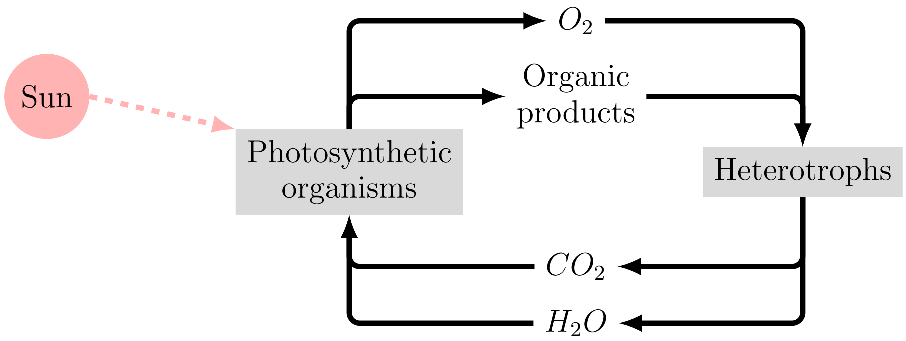


Figure 2.1: The cycle of carbon dioxide and oxygen (adapted from [85])

utilize. Higher animals usually need to obtain nitrogen in the form of amino acids ingested from their diet. Plants can use ammonia as their source of nitrogen, and only a few organisms can obtain their nitrogen directly from the gaseous nitrogen in the atmosphere. Figure 2.2 shows the relation among different organisms according to how they consume nitrogen.

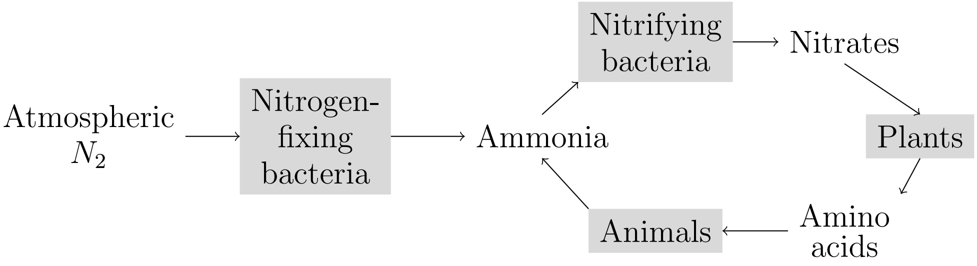


Figure 2.2: The nitrogen cycle (adapted from [85])

Metabolism can be differentiated into *catabolic* (or degradative) pathways and *anabolic* (or biosynthetic) pathways. Catabolism takes organic nutrients and cell constituents (carbohydrates, lipids, and proteins) and breaks them down into simpler products (lactic acid, CO_2, and ammonia) generating energy in the process.This energy is usually conserved in the form of adenosine-triphospate (ATP), or nicotinamide adenine dinucleotide phosphate ($NADPH$). During anabolism precursor molecules are used to build large macromolecules, such as proteins. The energy required in this process is provided by the breakdown of ATP and $NADPH$, as shown in Figure 2.3.

It should be noticed that although catabolic and anabolic pathways represent the two opposite directions of metabolism, there are significant differences among them. Catabolic pathways converge from a large pool of initial macromolecules to a few end products, the energy-poor molecules CO_2, H_2O, and NH_3. Anabolic pathways, on the other hand, diverge into a large number of products from very few precursors. More importantly, the corresponding and oppositely directed pathways (catabolic and anabolic) between a specific precursor and a given product are usually not identical (see Figure 2.4), since some steps in the pathway are catalyzed by two different enzymes (one for the catabolic and one for the anabolic direction). Furthermore, in eukaryotic cells anabolic and catabolic

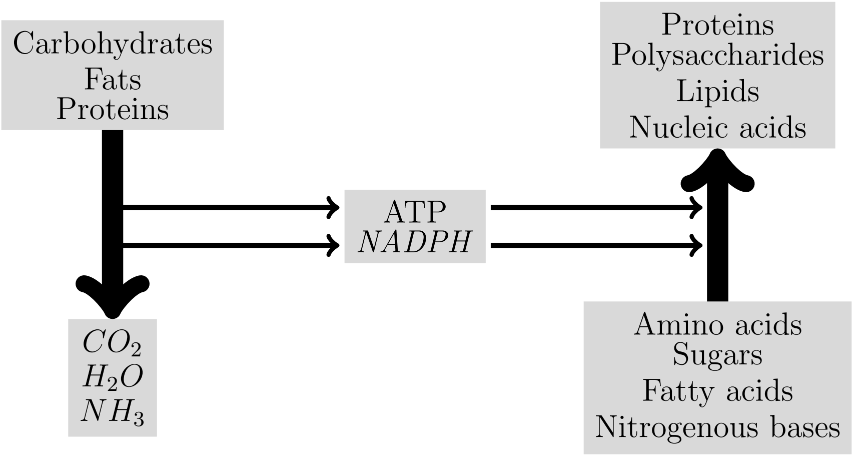


Figure 2.3: Relation between catabolic and anabolic pathways through energy molecules ATP and $NADPH$ (adapted from [85])

reactions involving the same constituents are frequently located in separate locations for simultaneous but independent operation.

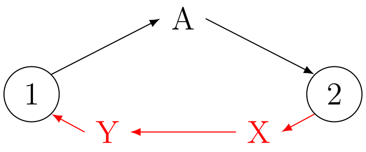


Figure 2.4: Independent catabolic and anabolic routes (adapted from [133])

Metabolism is a process that operates under the principle of maximum economy, with cells consuming only those nutrients necessary to meet the rate of energy utilization. The rates of synthesis of macromolecules is also adjusted to immediate needs, in order to avoid overproduction.

Control of metabolic flux can be regulated through different mechanisms. *Allosteric enzymes*, for instance, can change their catalytic activity in response to effectors that are often substrates, products, or coenzymes of the pathway but not necessarily of the reaction catalyzed by the enzyme itself. Allosteric enzymes are usually near the beginning of a sequence of enzymes, catalyzing its rate-limiting step. For instance, Figure 2.5 presents an example of negative feedback regulation, where the product of a reaction, galactose, competitively inhibits the catalyzing enzyme, beta-galactosidase.

A second mechanism for metabolic control is *hormonal regulation*, where hormones secreted by endocrine glands are carried by the blood to different tissues or organs, where they stimulate or inhibit certain metabolic processes. Finally, the *concentration* of an enzyme in the cell can be used as a way to regulate metabolic activities. The concentration of an enzyme, which depends on the relative synthesis/degradation rate, can be altered by turning on or off the corresponding biosynthesis pathway depending on the needs imposed by environmental conditions. Allosteric control is the fastest responding mechanism for metabolic flux control, while hormonal regulation and enzyme concentration respond more slowly to changing conditions.

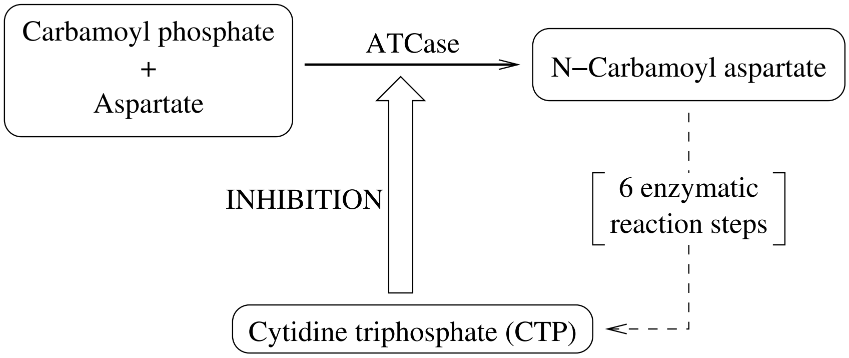


Figure 2.5: Allosteric control (adapted from [133])

Finally, it should be noticed that metabolic pathways occur in different compartments of the cell. Prokaryotic cells do not have internal membranes to separate specific cellular locations, yet there exists a partial separation of certain enzyme systems in bacteria (glycolysis in the cytosol, protein synthesis in the ribosomes, phosphorylation and electron-transport system in the cell membrane, etc). Eukaryotic cells do have a membrane-surrounded nucleus, as well as other membranous internal organelles (mitochondria, endoplasmic reticulum, Golgi complex, chloroplasts in green plant cells). This allows whole metabolic pathways to operate in different locations: glycolysis and fatty acid biosynthesis in the cytosol; citric acid cycle, electron transport and oxidative phosphorylation, fatty acid oxidation and amino acid catabolism in the mitochondria, protein synthesis in the ribosomes; replication of DNA and synthesis of nuclear proteins in the nucleus, etc. Because metabolites are synthesized in different membrane-bounded compartments in eukaryotic cells, a mechanism to transport these substances between compartments is required, making transport proteins essential for many metabolic processes.

2.1.2 ATP and NADPH: high energy compounds

ATP is a high energy intermediate metabolite which occurs in all life forms. It consists of adenosine moiety, linked to three phosphoryl groups via a phosphoester bond followed by two phosphoanhydride bonds. Energy released by degradation of complex nutrient molecules such as glucose is conserved as ATP by the synthesis of adenosine diphosphate (ADP) and inorganic phosphate (Figure 2.6). The energy conserved in ATP can then be used for different cell activities. ATP provides the energy required for the chemical work of biosynthesis, is the energy source for cell motility, can help transport nutrients through membranes against concentration gradients, and is used to ensure accurate genetic information transfer during the biosynthesis of DNA, RNA and proteins.

$NADPH$ carries energy in the form of hydrogen atoms and electrons. In order to reduce double bonds to single bonds, reducing power is needed in the form of hydrogen atoms, as in the formation of glucose from carbon dioxide or when fatty acids are made from acetate. To be effective in the reduction process hydrogen atoms must have free energy, which is obtained from cell fuels by dehydrogenases. Particularly, the hydrogen-carrying form of the coenzyme $NADP^+$, $NADPH$, transports electrons from catabolic reactions to electron-requiring biosynthetic reactions.

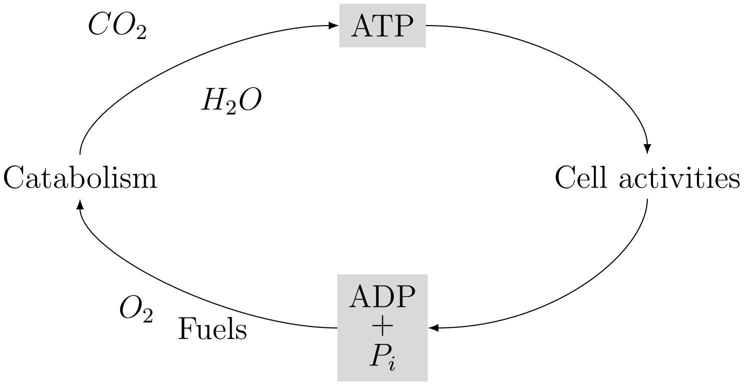


Figure 2.6: Use of energy stored in ATP for cell activity and regeneration from ADP through catabolism (adapted from [85])

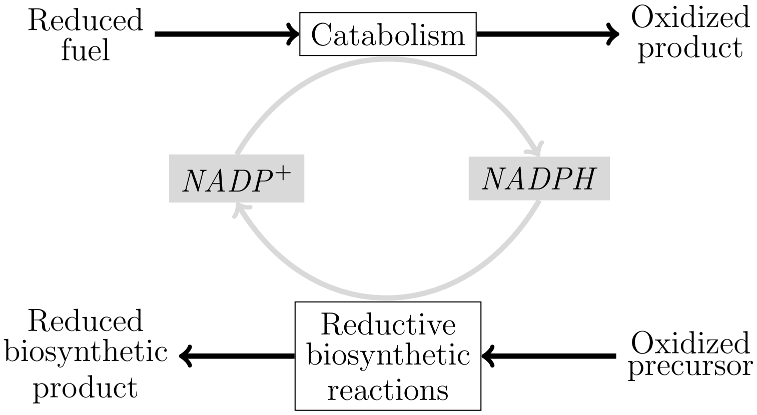


Figure 2.7: Reduction power transfered from catabolic to biosynthetic reactions via the *NADP* cycle (adapted from [85])

2.1.3 Experimental identification of metabolic pathways

Metabolic pathways can be studied from different perspectives:

- how nutrients are converted into end products, and how energy is utilized during the conversion process.
- how each intermediate metabolite is converted to its successor, and how specific enzymes catalyze each reaction.
- which are the mechanisms that regulate the flow of metabolites in the pathway, and how metabolic activity is adjusted to the needs of the entire organism.

Determining metabolic processes on all these levels is a complex problem, and several approaches are used (singly or in combination) to work out the chemical details of metabolic pathways.

A first approach is based on perturbing the system to alter its metabolic activity and study the effects of the perturbation. Adding certain substances to a pathway, such as metabolic inhibitors, can block the pathway at certain points and provoke an accumulation of preceding intermediate metabolites. Addition of substances that block electron transfer can also be used for this purpose.

The study of genetic mutations of organisms in which certain enzymes cannot be synthesized is another important method to elucidate metabolic pathways. When such defects are not lethal, they may result in the accumulation and excretion of the substrate of the defective enzyme. For instance, individuals with inherited alcaptonuria excrete homogentisic acid in their urine on the ingestion of phenylalanine or tyrosine, since these individuals lack the enzyme that catalyzes the breakdown of homogentisic acid. Alcohol dehydrogenase (ADH) and aldehyde dehydrogenase (ALDH) are also well-studied examples of genetic mutations. These enzymes allow the consumption of alcoholic beverages in human. It was found that 85% of Japanese have an atypical ADH and 50% an unusual ALDH [125], which explains the common alcohol sensitivity in individuals of Mongoloid origin.

A similar approach is to directly provoke some genetic modification to inactivate an enzyme in a specific pathway through the use of mutagens (chemical agents that induce genetic changes) or by genetic engineering, making inoperative certain genes (gene "knock-out"). Higher organisms engineered in such way are especially useful, in particular when the modification is not lethal.

Finally, isotropic tracers can be used to label metabolites. For instance, acetic acid can be synthesized so that its carboxyl carbon atom is enriched in the radioactive isotope ^{14}C. This radioactively labeled sample is then fed to an animal, where its metabolic fate can be easily traced. The respiratory CO_2 exhaled will contain ^{14}C, indicating some of the acetate is metabolized so that its carboxyl carbon atom is converted into CO_2. Alternatively, nuclear magnetic resonance (NMR) can be used to detect specific isotopes, such as ^{1}H, ^{13}C, ^{15}N, and ^{31}P, by their characteristic nuclear spin. Recent developments in NMR technology have even allowed to study metabolic pathways noninvasively in animals and humans, localizing the study in specific organs [103, 122].

2.1.4 KEGG: a repository for metabolic pathways

Although there are several pathway databases such as MetaCyc [18], WIT [110], ExPASy [49], PathDB [78] or UM-BBD [38], in this thesis we will work with data obtained from the KEGG repository [70], which has an extensive coverage of several metabolic processes in different species [140]. KEGG provides a knowledge base for linking information from the genomic to the phenotypical level, aiming at understanding high-order functions of biological systems from genomic and molecular data. The database is organized in a series of interconnected components to represent each of the levels, as presented in figure 2.8.

- **KEGG GENES** contains a series of gene catalogs for complete and partial genomes (85 eukaryota, 354 bacteria and 28 archaea as of October 2006) obtained from several public resources, mainly NCBI RefSeq [117]. Draft genomes are stored in the DGENES catalog, and expressed sequence tag consensus contigs in EGENES.

- **KEGG ORTHOLOGS** collects organized knowledge about orthologous and paralogous genes in the form of a pathway-based classification. The KO (KEGG Orthologs) identifier links genomic information in the GENES database to network

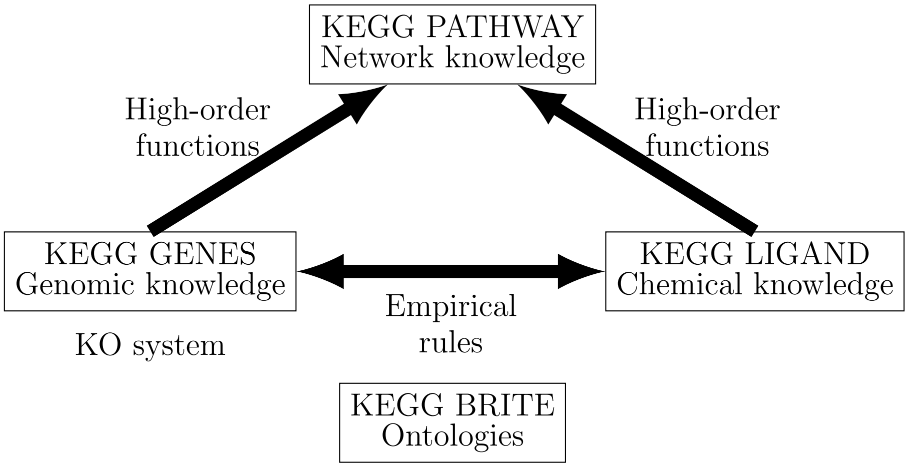


Figure 2.8: KEGG architecture (adapted from [70])

information in the PATHWAY database. Two additional ways to define KO's have been introduced: to use COG [128], to cover a wide range of possible orthologous groups, and to rely on classifications of protein families provided by experts.

- **KEGG LIGAND** stores data on chemical components, both endogenous and exogenous molecules. Enzymes (ENZYME), chemical compound structures (COMPOUND), chemical reaction formulas (REACTION), glycan structures (GLYCAN), reactant pair transformation patterns (RPAIR) and drug information (DRUG) are all stored in this database. The DRUG database, for instance, contains information such as therapeutic categories or target molecules.
- **KEGG PATHWAY** represents the molecular interactions and reaction networks for different processes. Metabolism, genetic information processing, environmental information processing (such as signal transduction), cell processes and human diseases are all modeled in this database as a collection of manually drawn pathway maps.
- **KEGG BRITE** is a set of hierarchical classifications representing knowledge on various aspects of biological systems. This component was designed to help automate functional interpretations associated with the KEGG pathway reconstruction and to assist discovery of empirical rules involving genome-environment interactions.

KEGG was our primary source of data for the experiments described in this thesis, and it is therefore necessary to discuss here the notation utilized to describe the different components in KEGG. Basically, in our work we have extensively used information from the KEGG LIGAND and KEGG PATHWAY databases. Enzymes in KEGG are identified according to the recommendations by the Enzyme Commission [137]: alcohol dehydrogenase, for instance, corresponds to ec:1.1.1.1. Chemical compounds are identified by the three letters "cpd" and an unique sequence of five digits, such as cpd:00001 for water or cpd:00037 for glycine. Reactions follow a similar schema, where R00001 would correspond to the reaction 2.1.

$$Polyphosphate + H_2O \Leftrightarrow Oligophosphate \tag{2.1}$$

Organisms have a three or four letter identifier ("hsa" for *Homo sapiens*). Metabolic pathways are represented by maps, such as "path:00010" for glycolysis. These maps are generic models abstracting the actual pathways occurring in different organisms. The specific pathway for an organism, represented by the organism and the pathway identifier ("hsa:00010" for glycolysis in *Homo sapiens*), would have annotated only those reactions, compounds and enzymes for which there is concrete evidence in the organism.

All information in KEGG can be accessed through their web site or by FTP. The KEGG API service (a SOAP/WSDL interface) provides also an alternative way of access to the databases.

2.2 Pathway alignment: previous approaches

2.2.1 Methods based on graph similarity

The structural similarities between metabolic pathways can be asserted by determining the isomorphisms [74, 131] and/or homeomorphisms [19, 132] of their graph representations. Although in general both problems are NP complete [48], simplifications utilizing properties of metabolic pathways can lead to tractable solutions. A general review of graph-based approaches for biological networks can be found in [2].

Forst and Schulten (1999, 2001)

Forst and Schulten [45, 46] consider a metabolic pathway as a reaction graph with certain topological properties (such as connectivity). Sequences corresponding to each *functional role* (gene product and how this product performs a specific task in a metabolic network) appearing in the pathway are combined into a set of sequences over which a multiple sequence alignment is performed. Pathways with different topologies are then compared using gap penalties for missing enzymes and adjacency matrices to address the graph topology.

The distance between two pathways Γ and Γ' with identical topology is defined by the equation 2.2.

$$\Delta = \sum_{i=1}^{n} \Phi_i \Delta X_i \tag{2.2}$$

where ΔX_i is the distance between the functional roles I_i and I'_i, $\Phi = 1$ for ortholog pair i, $\Phi = f$ for paralog pair i (with f chosen manually to minimize the number of distance triples violating the triangle inequality in the calculated distance matrix), and there are n functional roles $I_i, I'_i (i = 1, \ldots, n)$. Orthologs are genes evolved from a common ancestral gene by speciation, which retain the same function in the course of evolution. Paralogs, on the other hand, are homologous sequences separated by a gene duplication event within a genome. Orthologs usually have the same or similar function, while this is not always true for paralogs (the lack of selective pressure upon one of the duplicated genes allows it to evolve and acquire new functions). It should be noticed that assessment of orthologs and paralogs can only be done for completed genomes. When the networks compared have different graph topology, the common graph that includes both networks is used instead. If a functional role I_k is missing in one of the pathways, the distance ΔX_k is substituted by a gap value $\Delta_{gap} = 0.9$.

Heymans and Singh (2003)

The work by Heymans and Singh [62] takes the enzyme-enzyme relational graph representation for metabolic pathways, as described in [52, 108]. Similarity of two graphs is then calculated in four phases: first, the similarity between every pair of nodes (a, b) is iteratively calculated, where $a \in G_1$ and $b \in G_2$. Second, a bipartite graph is constructed using the similarity scores, and the maximal weight matching of this graph is found. The score between every pair of matched nodes is then recomputed, and finally the similarity score for the two graphs is calculated by normalizing the sum of the similarity of matched nodes.

The similarity of two nodes combines the similarity between the nodes themselves (the enzymes) and the similarity of their neighborhoods. The similarity of any two enzymes is calculated using the hierarchical similarity measure based on the respective Enzyme Commission identifiers [130]. Neighborhood similarity for nodes a and b is computed by summing their similarities and subtracting their dissimilarities. The first four terms correspond to the presence and absence of arcs from and to similar nodes, while the remaining four terms represent the mismatches between these edges.

The total complexity of the method is $O(Kn_1^2n_2^2 + (n_1 + n_2)^3)$, where n_1 and n_2 are the number of nodes of each graph and K is the number of iterations needed to converge to a solution (typical values of K are ≈ 20, depending on the size of the graph).

Pinter *et al.* (2005)

The work by Pinter *et al.* [116] is also based on the enzyme-enzyme relational graph to represent metabolic pathways. The similarity of pathways depends on two factors: the resemblance between any two corresponding nodes in the pathway graph (similarity between matched enzymes based on functional homology and calculated using the information content measure [130]), and the likeness between the pathways' network structure (topological similarity of the networks).

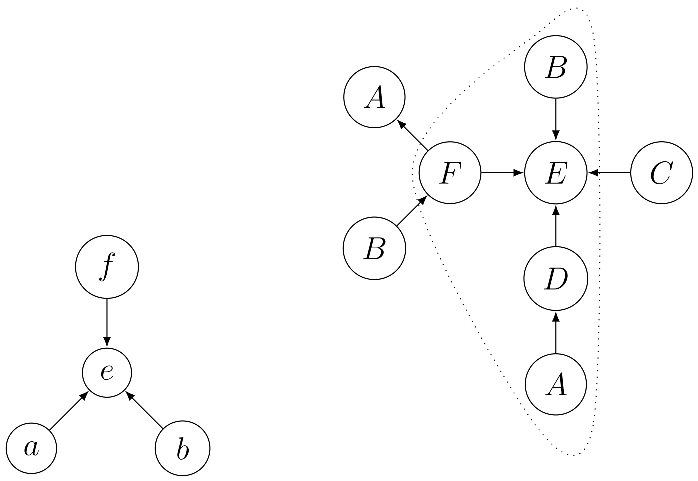
The pathway alignment algorithm is based on subtree homeomorphism, allowing the matching of nodes with distinct labels and scoring the match according to the similarity between the node labels. Extending the work on approximate subtree homeomorphism (see Figure 2.9 and original reference [115]), the authors recursively calculates the alignment scores among all subtrees of two given labeled trees to obtain a final alignment based on the maximum score. The complexity of this method is $O(m^2n/logm + mnlogn)$, where m and n are the number of vertices in the trees.

2.2.2 Methods based on algebra of sets

Algebra of sets describes basic properties of set operations and set relations, such as set union, intersection, equality or inclusion [55, 68]. The use of set theory has certain advantages over graph similarity measures when comparing metabolic pathways, such as simplicity and a lower computational cost.

Tohsato *et al.* (2000)

Tohsato [130] describes the multiple alignment of more than two pathways based on the enzyme hierarchy. It is argued that expressing enzyme similarity by using their



		A	B	C	D	E	F
$\Delta[i, j]$	a	+2	-2	-3	-2	-3	-3
	b	-1	+1	-2	+1	-2	-2
	e	-3	-2	-2	-3	+2	+1
	f	-3	-2	-1	+2	+1	+2

Figure 2.9: Approximate labeled subtree homeomorphism between trees T and T', with node-label similarity specified in Table Δ and deletion score -1. The dotted line encircles a subtree in T' homeomorphic to T, with score +7 (figure adapted from [116])

proximity within the enzyme hierarchy is problematic since there is a large deviation in the distribution of enzymes in the hierarchy (i.e., some subtrees are significantly larger than others). Enzyme similarity for two given enzymes is calculated using a new similarity measure based on the information content of the subtree rooted in their least common ancestor (see section 3.4 for more details).

The pathway alignment is obtained by extending the global alignment algorithm of Needleman and Wunsch on protein alignment [105]. Arranging the enzymes present in each pathway in a two-dimensional array, the algorithms looks for an optimal path starting at the top-left point and reaching the bottom-right point. Choosing a diagonal in the array corresponds to aligning the corresponding enzymes (with a score based on their information content similarity), while left-to-right and top-to-bottom moves in the array are considered gaps (with a score corresponding to the information content of the whole enzyme hierarchy tree, i.e., the smallest similarity). Figure 2.10 presents an example of this method.

The complexity of this method is $O(\ell^2)$, with ℓ being the maximum length of the two pathways. It does not use any information on chemical compounds or sequence. Although this algorithm is not strictly based on set theory, it discards information on graph topology and considers pathways as a set of enzymes, and therefore we have included it in this section.

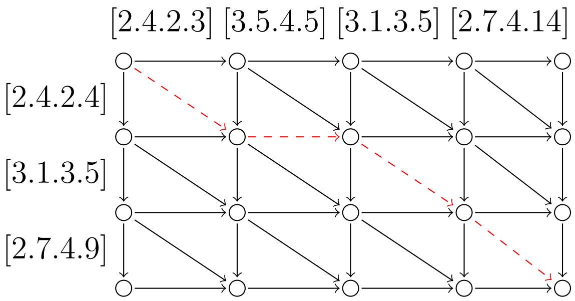


Figure 2.10: Pathway alignment based on global dynamic programming. The red dashed line shows the best path from the top-left to the bottom-right points, indicating the highest scoring match between the pathways

Forst and Schulten (2005)

In a different work by Forst and Schulten [44] to the one previously described, they use a new approach based on set algebra operations (union, intersection, and difference). Metabolic networks are represented as a directed hypergraph, with metabolites as nodes and reactions as directed hyperedges. In general, a metabolic network is defined as a pair (X, ξ), where X is the set of metabolites and ξ is the set of reactions.

The union of two networks M and M' is defined as $M'' = M \cup M' = (X \cup X', \xi \cup \xi')$, the intersection as $M'' = M \cap M' = (X \cap X', \xi \cap \xi')$, the difference as $M'' = M \setminus M' = (X \setminus X', \xi \setminus \xi')$, and the symmetric difference as $M'' = M \Delta M' = (M \cup M') \setminus (M \cap M')$. The number of reactions of a metabolic pathway M is denoted as $||M||$.

Using this notation, a distance measure is calculated as defined by the equation 2.3.

$$d(M, M') = \frac{||M \Delta M'||}{||M|| + ||M'|| - ||M \cap M'||} = \frac{||M \Delta M'||}{||M \cup M'||} \tag{2.3}$$

This method was developed in order to infer phylogenetic trees from the distance among a set of organisms. Since phylogenies based on single genes usually have problems associated with gene transfer, gene duplication, gene deletion, and functional replacement of genes, the authors took instead an approach not based on sequence information, as opposed to their previous work [45, 46].

Chapter 3

A New Method for Metabolic Pathway Alignment

Todo hombre puede ser, si se lo propone, escultor de su propio cerebro

Santiago Ramón y Cajal [144]

3.1 Representation of metabolic pathways

There has been much interest in the structural comparison and alignment of metabolic pathways, and several techniques have been conceived to assess the similarity of such networks for different organisms. In the comparative analysis of metabolism, pathways from different genomes are aligned upon similar enzymes, substrates, and products [29, 116, 130].

The assessment of structural metabolic similarity among different organisms involves both a graph representation of metabolic processes and a similarity measure between individual reactions, enzymes, and compounds present in the pathways. Metabolic pathways are represented as directed hypergraphs, with the compounds and enzymes being the nodes and the reactions activated by the enzymes being hyperarcs [33]. For instance, the directed hypergraph for the Citric Acid Cycle pathway in the bacterium *Escherichia coli* consists of 35 nodes and 18 hyperarcs.

A more abstract representation, called the enzyme-enzyme relational graph, has been used in [62, 109], where nodes represent enzymes and arcs represent compounds shared between reactions catalyzed succesively by specific enzymes/nodes. For instance, the enzyme-enzyme relational graph for the Citric Acid Cycle pathway in *Escherichia coli* consists of only 14 nodes and 23 arcs.

The representation of metabolic pathways as hypergraphs has several advantages over the enzyme-enzyme relational graph, since it allows to use both enzymes and metabolites when calculating the similarity of two pathways. By using this representation we can also establish the relative importance of enzymes and compounds in the global measure of similarity, as will be seen in Section 3.3.

The algorithm for metabolic pathway alignment presented in this thesis will be detailed in the following sections as follows: the similarity of two given pathways (Section 3.2) is constructed as the similarity of the set of reactions involved on them (Section 3.3), which depends on the similarity of the enzymes (Section 3.4) and compounds (Section 3.5). The alignment of pathways is obtained by aligning the reactions in a way such that the total score of the alignment is maximized.

3.2 Similarity of metabolic pathways

The similarity of metabolic pathways presented in this thesis is based on the similarity of reactions, enzymes and compounds as described later in this chapter. This measure is based on two operations, set intersection and set difference, as follows: given two pathways P and P', each reaction belonging exclusively to P ($P \setminus P'$) is aligned to its most similar reaction in P', while each reaction belonging exclusively to P' ($P' \setminus P$) is aligned to its most similar reaction in P. Reactions belonging to the intersection ($P \cap P'$) are aligned to their respective equivalents. Figure 3.1 depicts a graphical representation of this method.

The similarity of two metabolic pathways $\mathbf{P} = (\mathbf{C}, \mathbf{R})$ and $\mathbf{Q} = (\mathbf{D}, \mathbf{S})$, where $\mathbf{C}, \mathbf{D}$ are sets of compounds and $\mathbf{R}, \mathbf{S}$ are sets of enzymatic reactions, is described by Equation 3.1.

$$\begin{aligned} sim(\mathbf{P}, \mathbf{Q}) = \frac{1}{|\mathbf{R} \cup \mathbf{S}|} \Bigg(& \sum_{R \in \mathbf{R} \cap \mathbf{S}} \max_{S \in \mathbf{R} \cap \mathbf{S}} sim(R, S) \\ & + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim(R, S) \\ & + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} sim(R, S) \Bigg) \end{aligned} \tag{3.1}$$

Equation 3.1 can be simplified into Equation 3.2 when the reaction similarity is a normalized metric, as it is the case.

$$sim(\mathbf{P}, \mathbf{Q}) = \frac{1}{|\mathbf{R} \cup \mathbf{S}|} \left(|\mathbf{R} \cap \mathbf{S}| + \sum_{R \in \mathbf{R} \setminus \mathbf{S}} \max_{S \in \mathbf{S}} sim(R, S) + \sum_{S \in \mathbf{S} \setminus \mathbf{R}} \max_{R \in \mathbf{R}} sim(R, S) \right) \tag{3.2}$$

Equation 3.2 has three terms. The first term corresponds to reactions shared by both pathways, so their alignment is immediate and does not have a direction. The second and third term, on the other hand, take into account both directions of the alignment (from the first pathway to the second, and vice versa), resulting in a final measure of similarity that is symmetric. This equation can be implemented using algorithm 1 in Appendix C.

The metabolic pathway similarity measure is a normalized metric, and it can be computed in time quadratic in the number of compounds, enzymes, and reactions in the metabolic pathways. Assuming the comparison of two compounds, enzymes, or reactions can be done in $O(1)$ time, the similarity of two pathways with respectively n and m reactions can be calculated in $O(mn)$ time. Although the actual comparison of two enzymes might take more than $O(1)$ (in particular when using the *information content* or *gene ontology* enzyme similarity measures that will be discussed in Section 3.4), for practical purposes we can precompute and store them in order to achieve $O(1)$ time bound in subsequent queries. Chapter 5 discusses the implementation details of our method, and Appendix A further describes the order of the algorithm for the worst and average cases.

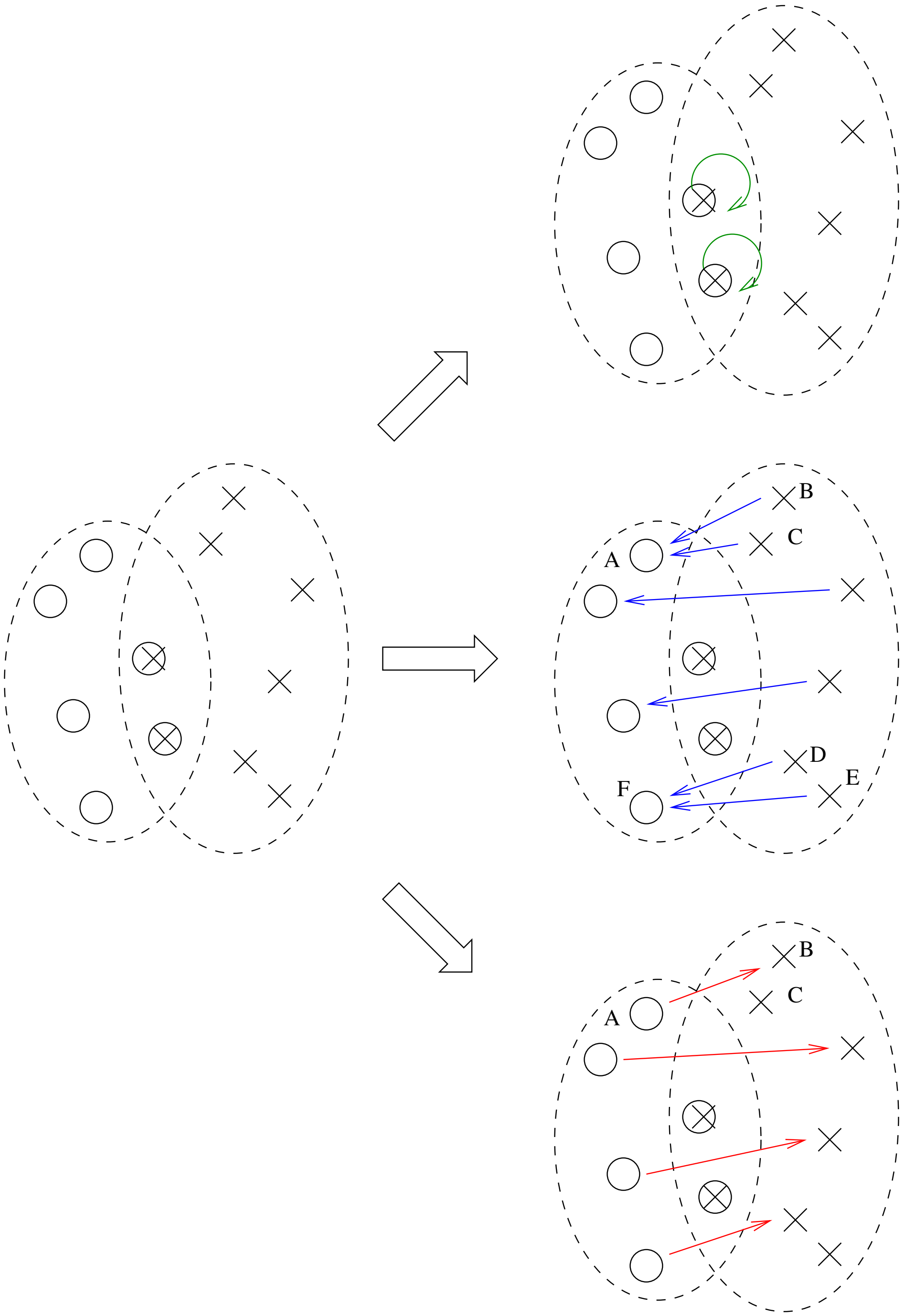


Figure 3.1: Metabolic pathway alignment based on set operations. Given two pathways represented by their set of reactions (left), reactions in their intersection get mutually aligned (right, top), reactions belonging exclusively to the first pathway are aligned to reactions in the second pathway (right, middle), and reactions only in the second pathway are aligned to reactions in the first pathway (right, bottom). Notice that the alignment is directed (reaction C aligns to A, but A aligns instead to B) and many-to-one (reactions D and E get aligned to F)

3.3 Similarity of metabolic reactions

Similarity of reactions can be assessed by measuring the similarity of compounds and enzymes involved in them. Let $\mathbf{R}$ and $\mathbf{R}'$ denote two reactions, with $\mathbf{C_R}$ and $\mathbf{C_{R'}}$ their respective compound sets, and $\mathbf{E_R}$ and $\mathbf{E_{R'}}$ their enzyme sets. Equation 3.3 defines the similarity of $\mathbf{R}$ and $\mathbf{R}'$.

$$\begin{aligned} sim(\mathbf{R},\mathbf{R}') = & \frac{1-\alpha}{|\mathbf{C_R}\cup\mathbf{C_{R'}}|}\left(\sum_{\mathbf{c}\in\mathbf{C_R}\cap\mathbf{C_{R'}}}\max_{\mathbf{d}\in\mathbf{C_R}\cap\mathbf{C_{R'}}} sim(\mathbf{c},\mathbf{d}) + \sum_{\mathbf{c}\in\mathbf{C_R}\setminus\mathbf{C_{R'}}}\max_{\mathbf{d}\in\mathbf{C_{R'}}} sim(\mathbf{c},\mathbf{d})\right. \\ & \left. + \sum_{\mathbf{d}\in\mathbf{C_{R'}}\setminus\mathbf{C_R}}\max_{\mathbf{c}\in\mathbf{C_R}} sim(\mathbf{c},\mathbf{d})\right) + \frac{\alpha}{|\mathbf{E_R}\cup\mathbf{E_{R'}}|}\left(\sum_{\mathbf{e}\in\mathbf{E_R}\cap\mathbf{E_{R'}}}\max_{\mathbf{f}\in\mathbf{E_R}\cap\mathbf{E_{R'}}} sim(\mathbf{e},\mathbf{f})\right. \\ & \left. + \sum_{\mathbf{e}\in\mathbf{E_R}\setminus\mathbf{E_{R'}}}\max_{\mathbf{f}\in\mathbf{E_{R'}}} sim(\mathbf{e},\mathbf{f}) + \sum_{\mathbf{f}\in\mathbf{E_{R'}}\setminus\mathbf{E_R}}\max_{\mathbf{e}\in\mathbf{E_R}} sim(\mathbf{e},\mathbf{f})\right) \end{aligned} \tag{3.3}$$

In Equation 3.3, α represents a weight parameter ($0 \leqslant \alpha \leqslant 1$) to establish the relative weight of compound similarity to enzyme similarity in the assessment of reaction similarity: a value of $\alpha = 0$ would mean no relevance is given to enzymes, while $\alpha = 1$ would give no relevance to compounds.

Equation (3.3) can be simplified when compound and enzyme similarity are normalized metrics, as in the case of the similarity measures used in this work, as follows:

$$\begin{aligned} & sim(\mathbf{R},\mathbf{R}') = \\ & \quad \frac{1-\alpha}{|\mathbf{C_R}\cup\mathbf{C_{R'}}|}\left(|\mathbf{C_R}\cap\mathbf{C_{R'}}| + \sum_{\mathbf{c}\in\mathbf{C_R}\setminus\mathbf{C_{R'}}}\max_{\mathbf{d}\in\mathbf{C_{R'}}} sim(\mathbf{c},\mathbf{d}) + \sum_{\mathbf{d}\in\mathbf{C_{R'}}\setminus\mathbf{C_R}}\max_{\mathbf{c}\in\mathbf{C_R}} sim(\mathbf{c},\mathbf{d})\right) + \\ & \quad \frac{\alpha}{|\mathbf{E_R}\cup\mathbf{E_{R'}}|}\left(|\mathbf{E_R}\cap\mathbf{E_{R'}}| + \sum_{\mathbf{e}\in\mathbf{E_R}\setminus\mathbf{E_{R'}}}\max_{\mathbf{f}\in\mathbf{E_{R'}}} sim(\mathbf{e},\mathbf{f}) + \sum_{\mathbf{f}\in\mathbf{E_{R'}}\setminus\mathbf{E_R}}\max_{\mathbf{e}\in\mathbf{E_R}} sim(\mathbf{e},\mathbf{f})\right) \end{aligned} \tag{3.4}$$

The value of parameter α can have great relevance in deciding which reaction is the most similar to others. For instance, consider reaction R00351 in *Archaeoglobus fulgidus*:

$$R00351 \quad \textit{Citrate} + \textit{CoA} \Leftrightarrow \textit{Acetyl-CoA} + H_2O + \textit{Oxaloacetate} \quad [2.3.1.1]$$

and reactions R00362 and R01323 in *Clostridium perfringens*:

$$\begin{array}{lcr} R00362 & \textit{Citrate} \Leftrightarrow \textit{Acetate} + \textit{Oxaloacetate} & [4.1.3.6] \\ R01323 & \textit{Acetyl-CoA} + \textit{Citrate} \Leftrightarrow \textit{Acetate} + (3S)\textit{-Citryl-CoA} & [2.8.3.10] \end{array}$$

By applying equation (3.4), we obtain:

$$\begin{aligned} sim(\text{R00351}, \text{R00362}) &= (1-\alpha)/2 \\ sim(\text{R00351}, \text{R01323}) &= (4-\alpha)/20 \end{aligned}$$

The relative weight α can then select whether we want to emphasize more on enzyme similarity or compound similarity of the reactions. In this case values of $\alpha < 2/3$ would mean reaction R00362 is more similar to R00351, while values of $\alpha > 2/3$ would mean reaction R01323 is closer. The above equations can be implemented using the algorithms 2, 3 and 4 in Appendix C.

Figure 3.2 presents an example of reactions alignment for the TCA cycle in three different organisms. As it can be noticed the alignment produced is many-to-one (reactions R00342, R00344 and R00351 in *Archaeoglobus fulgidus* get all aligned with R00362 in *Clostridium perfringens*) and directed (reaction R01324 in *Listeria innocua* gets aligned to R00351 in *Archaeoglobus fulgidus*, but the opposite does not hold). The similarity score for a specific alignment is nevertheless symmetric (not directed): the similarity of R01324 in *Listeria innocua* and R00351 in *Archaeoglobus fulgidus* is 0.5 in both directions.

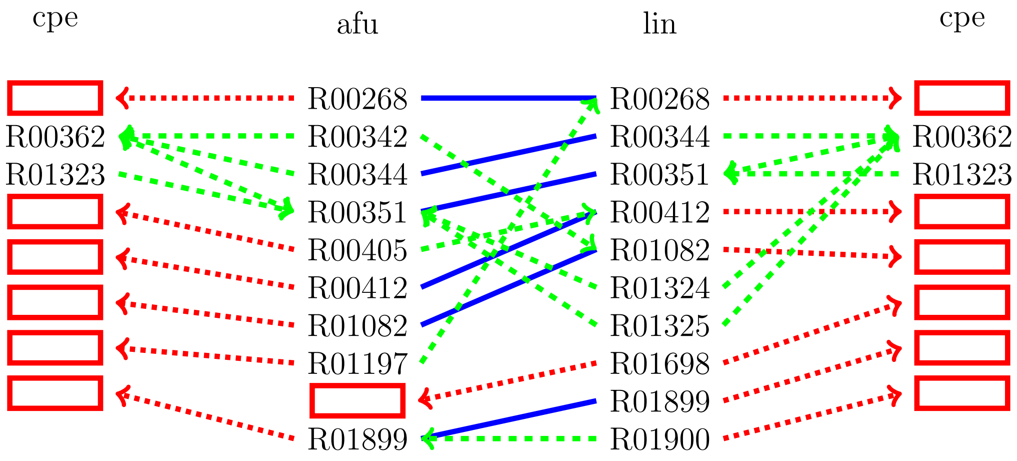


Figure 3.2: Maximum similarity alignment of the TCA cycle (KEGG pathway number 00020) for the organisms *Archaeoglobus fulgidus* (afu), *Clostridium perfringens* (cpe), and *Listeria innocua* (lin). Each reaction in the metabolic pathway of one organism is pseudo-aligned with the most similar reaction in the metabolic pathway of another organism, in both directions, and the mapping is shown in blue/solid in the case of reactions common to the pathway in the two organisms (equivalent reactions), in green/dashed for similarity between reactions greater than 0 but less than 1, and in red/dotted for reaction similarity equal to zero. Notice that pairs of reactions not connected by arrows could be aligned, although with lower similarity scores. Since this graph shows only the best possible alignment those connections are not represented

3.4 Similarity of enzymes

In order to assess the similarity of enzymes, we have utilized three measures: hierarchical similarity [130], information content similarity [130], and gene ontology similarity [23].

The first two measures are based on the enzyme hierarchy, an accepted system for naming and classification of enzymes developed by the Enzyme Commission [137] of the International Union of Biochemistry and Molecular Biology (IUBMB). The enzyme hierarchy classifies enzymes into six main classes on the basis of the reaction activated by the enzyme. Each enzyme is assigned a code, the EC number, which is a string of four digits separated by dots. The first digit shows the main class to which the enzyme belongs. The second and third digits in the EC number further describe the kind of reaction being activated, and their meanings are defined separately for each of the main classes. The fourth digit distinguishes between enzymes activating very similar but non-identical reactions, by defining the actual substrate.

Consider, for instance, EC code 3.2.1.108. This code corresponds to the lactase enzyme, which activates the hydrolysis of the disaccharide lactose to its component monosaccharides glucose and galactose. The first digit corresponds to class 3.-.-.-, the hydrolases. For the hydrolases, the second digit identifies the type of bond hydrolyzed and the third digit further describes the type of bond hydrolyzed. In the case of lactase, 3.2.-.- corresponds to the glycosylases, which have a glycosidic bond (linking carbohydrate units), while 3.2.1.- represents the glycosidases (enzymes hydrolysing O-glycosyl and S-glycosyl compounds). Lactase actually activates hydrolysis of the O-glycosyl bond. The fourth and last digit identifies the particular reaction. In the case of lactase, 3.2.1.108 identifies the actual lactose being hydrolyzed.

The third measure of enzyme similarity used in this thesis is based on the molecular function subtree of the Gene Ontology (GO) [8, 91], which describes cell activities at the molecular level using a directed acyclic graph and includes annotations to enzymatic activity. For example, EC code 3.2.1.108 is annotated in GO to the term/node GO:0000016 (lactase activity), which is child of GO:0004553 (hydrolase activity, hydrolyzing O-glycosyl compounds). This node is in turn child of GO:0016798 (hydrolase activity, acting on glycosyl bonds), which is child of GO:0016787 (hydrolase activity). Figures 3.3 and 3.4 present both the Enzyme Hierarchy and Gene Ontology representation of this enzyme.

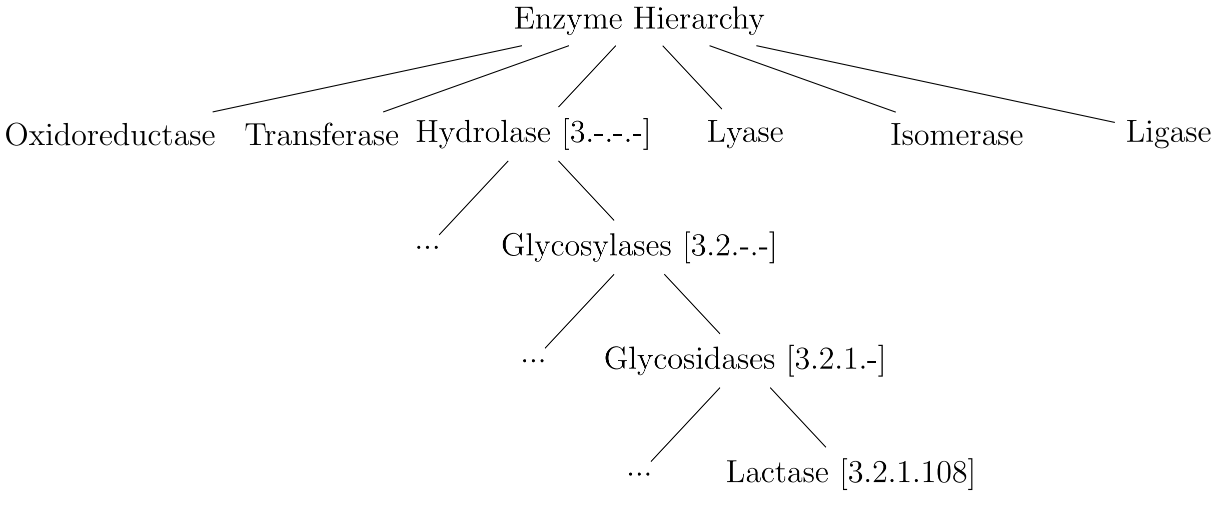


Figure 3.3: Lactase represented in the Enzyme Hierarchy

3.4.1 Hierarchical enzyme similarity

The *hierarchical similarity* of two enzymes [130] is the number of common most significant EC digits of the enzymes over 4. The five possible values of hierarchical similarity are

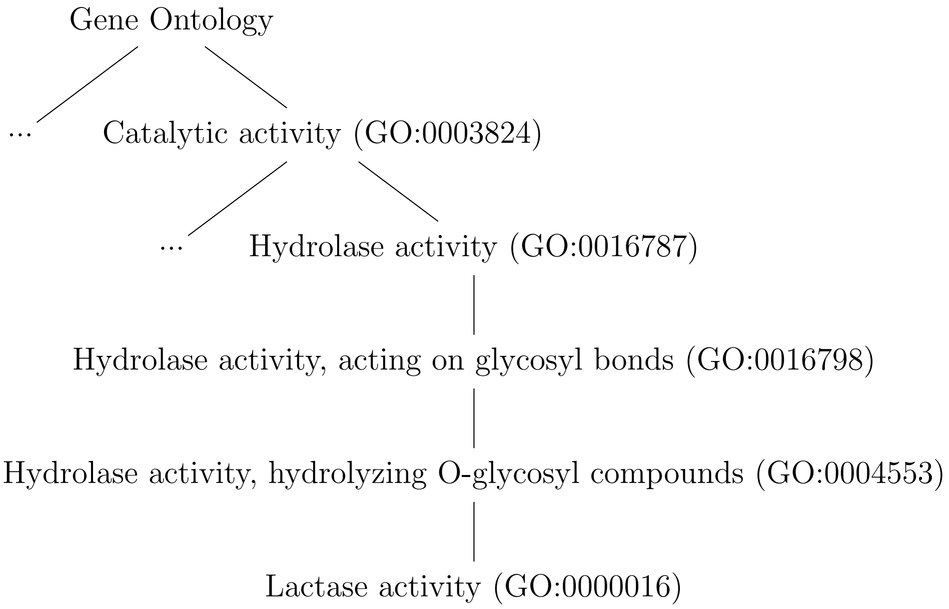

Gene Ontology
···
Catalytic activity (GO:0003824)
···
Hydrolase activity (GO:0016787)
Hydrolase activity, acting on glycosyl bonds (GO:0016798)
Hydrolase activity, hydrolyzing O-glycosyl compounds (GO:0004553)
Lactase activity (GO:0000016)

Figure 3.4: Lactase represented in the Gene Ontology

thus: 0, for two dissimilar enzymes (with their first digit different); 0.25, if the first digit is identical and the second digit is different; 0.5, if the first two digits are identical but the third digit is different; 0.75, if the first three digits are identical but the last digit is different; and 1, for two identical enzymes (with all four digits identical). More formally, consider enzymes e and e', and their string representations $e_a.e_b.e_c.e_d$ and $e'_a.e'_b.e'_c.e'_d$ corresponding to their respective EC identifiers. If α represents the value (0 or 1) of the logical expression $e_a = e'_a$, β is $e_b = e'_b$, γ is $e_c = e'_c$, and δ is $e_d = e'_d$, then the similarity of e and e' can be calculated as shown in Equation 3.5.

$$sim(e, e') = \frac{\alpha(1 + \beta(1 + \gamma(1 + \delta)))}{4} \tag{3.5}$$

The intuition behind hierarchical similarity is to measure how close two enzymes are to each other in the enzyme hierarchy, with higher similarity values for closer enzymes. The hierarchical similarity between two enzymes is inversely related to the shortest path distance between the enzymes in the enzyme hierarchy as presented in Equation 3.6.

$$sim(e, e') = max(0, 1 - \frac{dist(e, e')}{8}) \qquad \text{with } dist(e, e') = 0, 2, 4 \ldots \tag{3.6}$$

For instance, the hierarchical similarity between the enzymes lactase (3.2.1.108) and glycosylceramidase (3.2.1.62) is 0.75 (path length 2), because they share the first three digits, while the hierarchical similarity between lactase and adenosine nucleosidase (3.2.2.7) is 0.5 (path length 4), and the hierarchical similarity between lactase and phloretin hydrolase (3.7.1.4) is 0.25 (path length 6).

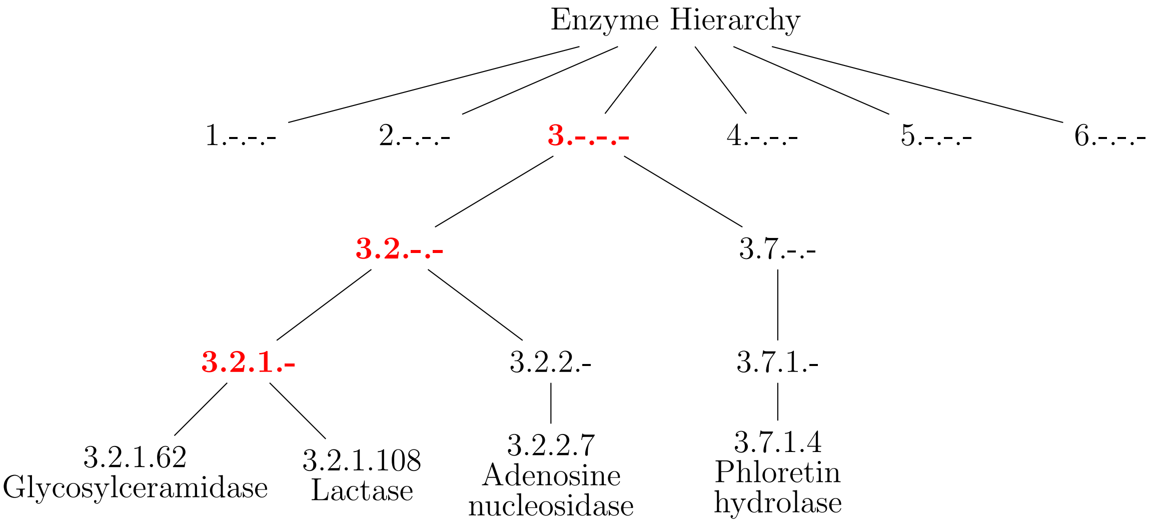


Figure 3.5: Enzyme hierarchical similarity between lactase and glycosylceramidase, adenosine nucleosidase and phloretin hydrolase

3.4.2 Information content enzyme similarity

The similarity of two enzymes can also be taken to be the information content of their least common ancestor in the enzyme hierarchy. The *information content similarity* of two enzymes [130] is minus the logarithm of the size (E) of the enzyme hierarchy subtree rooted at the least common ancestor (lca) of the enzymes. Similarity values based on information content range from a smallest value of 0, for two identical enzymes, to a largest negative value of about -12, for two dissimilar enzymes, and they can be normalized by dividing over the size of the whole enzyme hierarchy (k) and subtracting the obtained value from 1 (Equation 3.7). The intuition behind information content similarity is also to measure how close two enzymes are to each other in the enzyme hierarchy, with higher similarity values for closer enzymes. Unlike hierarchical similarity, though, the basis of the similarity measure is not the shortest path between the enzymes in the enzyme hierarchy but the whole subtree rooted at their least common ancestor, thus avoiding problems related to unequal distribution of enzymes among the hierarchy.

$$sim(e, e') = 1 - \frac{log_2 E(lca(e, e'))}{k} \tag{3.7}$$

For instance, the normalized information content similarity between lactase (3.2.1.108) and glycosylceramidase (3.2.1.62) is 0.404, because class 3.2.1.- has 151 enzymes, between lactase and adenosine nucleosidase (3.2.2.7) is 0.386, because class 3.2.-.- has 176 enzymes, and between lactase and phloretin hydrolase (3.7.1.4) is 0.153, because class 3.-.-.- has 1,252 enzymes.

3.4.3 Gene ontology enzyme similarity

The third method used to assess the similarity of two enzymes is based on the gene ontology. The Gene Ontology (GO) is a widely accepted standard for describing genes and gene products [8]. GO is composed of *concepts*, each of them described by an unique

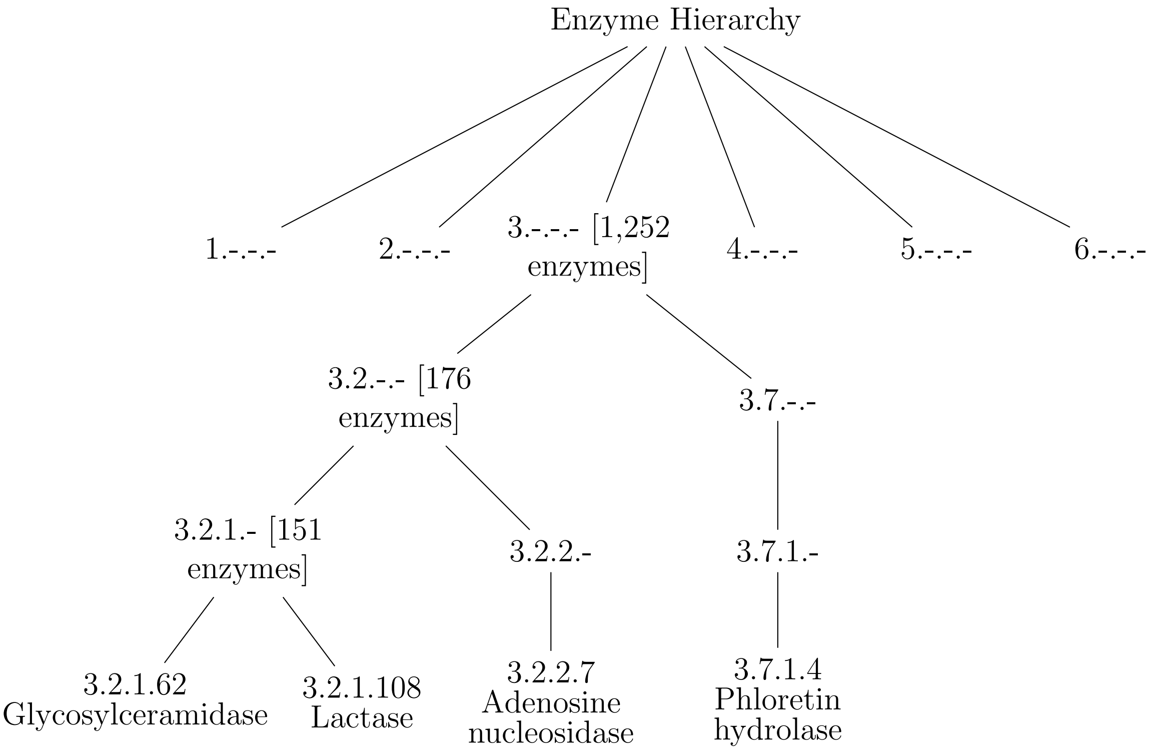


Figure 3.6: Information content similarity between lactase and glycosylceramidase, adenosine nucleosidase and phloretin hydrolase

identifier and one or more strings to name the concept. GO concepts are related to each other by *is-a* and/or *part-of* relations, arranged as a directed acyclic graph.

GO includes three different ontologies: *molecular function*, to describe activities at the molecular level; *biological process*, which deals with series of events accomplished by molecular functions; and *cellular component*, describing different parts of the cell. The molecular function ontology contains concepts representing most of the enzymes present in the Enzyme Commission (EC) database. In a previous work [23] we introduced a new enzyme similarity measure based on the shortest distance in the GO hierarchy (not considering direction or type of relation) between the concepts representing any pair of enzymes. Enzymes that have no associated GO entry are substituted by the concept in the GO corresponding to the closest sibling enzyme in the EC tree. Dijkstra's algorithm was used to calculate the minimum distance between GO concepts.

The gene ontology similarity measure is conceptually similar to the hierarchical one, since both are based on the shortest path between enzymes, but using a different representation of the enzyme taxonomy, namely, the corresponding associated subgraph of the Gene Ontology.

The gene ontology distance between lactase (mapped to GO:0000016, "lactase activity") and glycosylceramidase (mapped to GO:0017042, "glycosylceramidase activity") is 2, since the shortest path is [GO:0000016, GO:0004553, GO:0017042]. Lactase and adenosine nucleosidase (GO:0047622, "adenosine nucleosidase activity") are at distance 4 through the path [GO:0000016, GO:0004553, GO:0016798, GO:0016799, GO:0047622]. The most dissimilar examples under hierarchical similarity are also the most distant ones under gene ontology similarity: lactase and phloretin hydrolase (GO:0050180, “phloretin hydrolase activity”) are at distance 6 in the path [GO:0000016, GO:0004553, GO:0016798, GO:0016787,GO:0016822, GO:0016823, GO:0050180]. It should be noticed that the gene ontology similarity is more fine-grained than those previously presented: while similarity between lactase and alcohol dehydrogenase (1.1.1.1, GO:008016) is positive for this measure (path distance 9), both hierarchical and information content measures score the similarity as 0. The normalized similarity values for the gene ontology measure were obtained by dividing the obtained path distances over the maximum distance among all pairs of enzymes in the metabolic pathway, and subtracting the resulting values from 1.

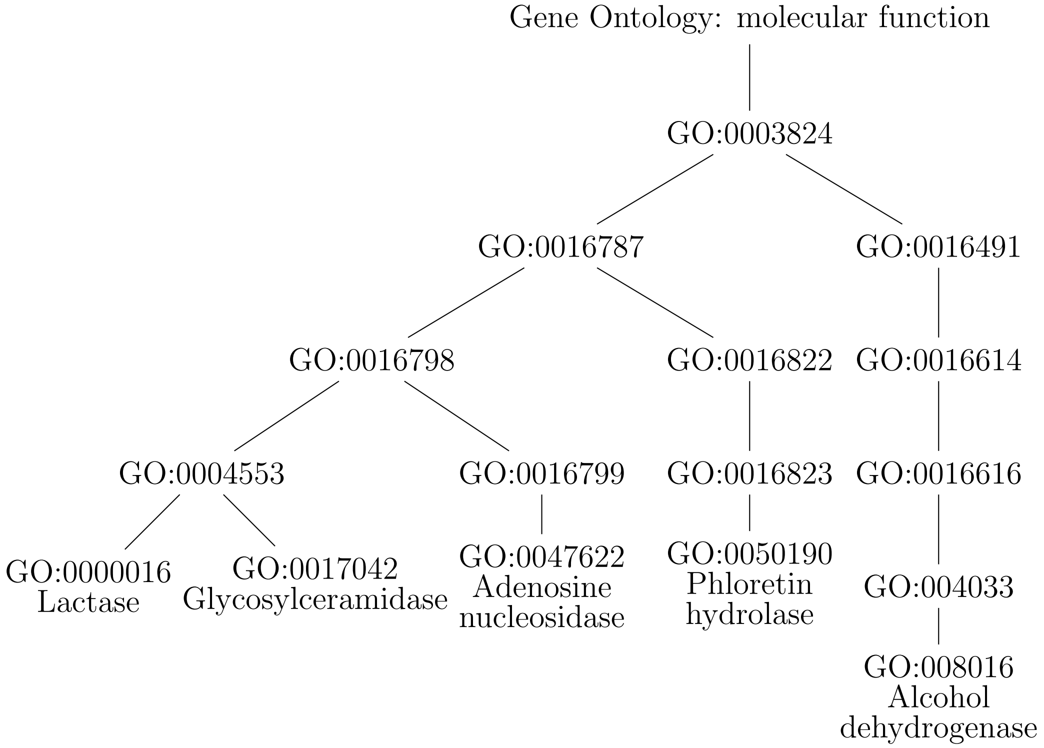

Gene Ontology: molecular function
GO:0003824
GO:0016787
GO:0016491
GO:0016798
GO:0016822
GO:0016614
GO:0004553
GO:0016799
GO:0016823
GO:0016616
GO:0000016 Lactase
GO:0017042 Glycosylceramidase
GO:0047622 Adenosine nucleosidase
GO:0050190 Phloretin hydrolase
GO:004033
GO:008016 Alcohol dehydrogenase

Figure 3.7: Gene ontology similarity between lactase and glycosylceramidase, adenosine nucleosidase and phloretin hydrolase

3.5 Similarity of compounds

In order to assess the similarity of compounds, we have just taken a similarity of 1 for identical compounds and 0 for distinct compounds. A more complex similarity measure based on shortest path distance among compounds using ChEBI, a chemical ontology [31], did not improve results presented in Chapter 4 and therefore was not included in this thesis.

As in previous studies [94, 153], we have discarded the so-called *current metabolites*, which function as cofactors in many reactions, namely: H_2O (KEGG id: C00001), ATP (C00002), NAD^+ (C00003), NADH (C00004), NADPH (C00005), $NADP^+$ (C00006), O_2 (C00007), ADP (C00008), Orthophosphate (C00009), CoA (C00010), CO_2 (C00011), Pyrophosphate (C00013), NH_3 (C00014), and UDP (C00015). Whether a metabolite is essential or not depends on the reaction in which it appears [94], but for simplicity we have considered current metabolites to function as cofactors in all reactions. Appendix D offers some more insights into the role of common metabolites and their possible influence in the results here presented.

3.6 Comparison with previous approaches

The method presented in this chapter has certain differences with those presented in Chapter 2, which can be summarized as follows:

- we do not use sequence information in our calculations. The motivation behind this decision is twofold: first, by not using sequence data we do not require full genomes of the organisms to be compared (as, for example, in [45]. And second, as we will see in Section 4.1, by not using sequence information we are trying to avoid problems associated with horizontal gene transfer [14, 66, 82][1].
- our method does not need to artificially establish a penalty score for "gaps" (i.e., reactions in a pathway that cannot be aligned to any reaction in the second pathway with similarity greater than zero).
- we use a metabolic pathway representation including both compounds and enzymes, which better reflects the relevance of both elements in metabolism, as opposed to those approaches based exclusively on catalytic enzymes.
- we can control the relative influence that metabolites and enzymes will have in the overall metabolic similarity score, by using the α parameter as described in Section 3.3. As we will see in Chapter 4, this parameter can be adjusted to improve results in phylogenetic reconstruction from metabolic similarity.
- our method is computationally faster than methods based on graph similarity [46, 62, 116], and at least as fast as methods based on algebra of sets [44, 130].

[1] Although the extension to which horizontal gene transfer is relevant in evolutionary process is still argued, see for instance [79].

Chapter 4

Applications

> *The nature of physical things is much more easily conceived when they are beheld coming gradually into existence, then when they are only considered as produced at once in a finished and perfect state*
>
> René Descartes [32]

4.1 Phylogenetic reconstruction

The understanding of evolutionary relationships among species has recently shifted from more conventional studies that exploit polymorphism information in DNA or protein sequence to assess the phylogenetic relationship among species [54, 106], to new studies aimed at assessing the evolution of complete biological processes [1][1].

There has also been much interest in the phylogenetic analysis of metabolic pathways, and several techniques have been conceived to extend the similarity assessment of these pathways into phylogenies for different organisms. Previous phylogenetic analyses have been based on the number of common enzymes between two organisms [45, 46], on profiles of the presence and absence of the various metabolic pathways [88], and on the topology of the underlying enzyme-enzyme relational graphs [62]. The produced phylogenies have often been evaluated by comparing them against the NCBI taxonomy [138], which is based on 16s ribosomal RNA sequences, and the best results so far have been obtained by Heymans and Singh [62]. The phylogenetic analysis of metabolic pathways has also lead to the identification of conserved pathway modules in different organisms [76, 145, 146].

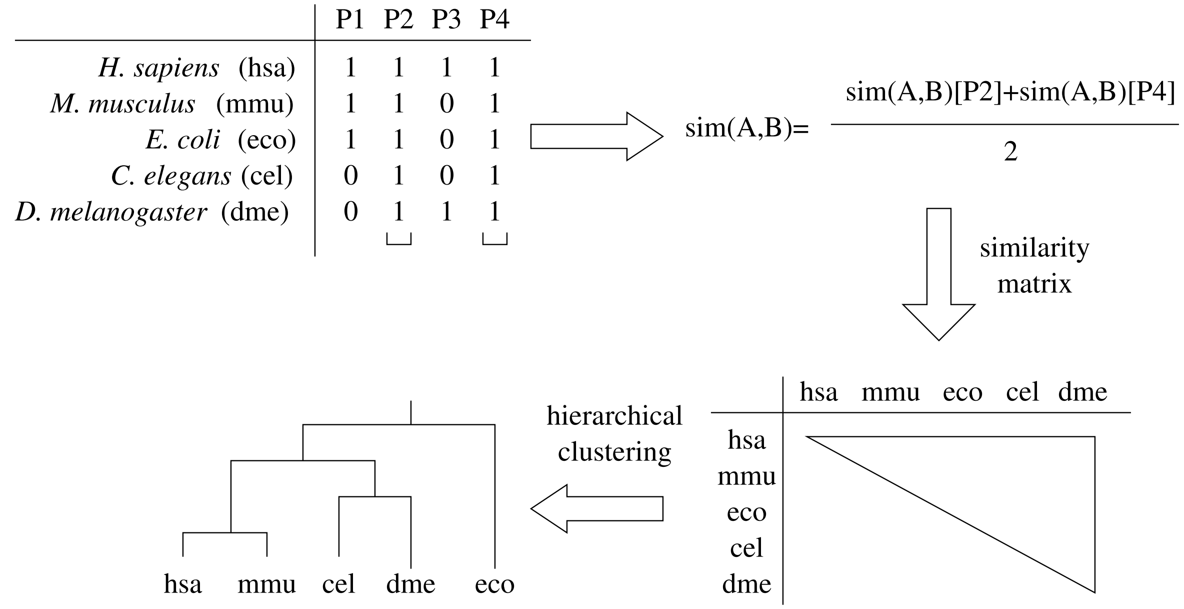
Phylogenetic reconstruction can be greatly affected by *horizontal gene transfer* (HGT), the process by which organisms transfer genetic material independent of reproduction [127]. The use of single genes as phylogenetic markers has been shown to be particularly unreliable [152]. Reconstruction of phylogenies from metabolic similarity of species provides therefore an alternative approach not burdened by problems associated to HGT.

Co-analysis of phylogeny and metabolic pathways can, from a more general perspective, provide valuable insights into the problem of explaining the appearance and development of complex networks of interacting proteins and chemical molecules [36, 120]. Several theories have been proposed to explain the evolution of such networks (see [84] for

[1] The interested reader is referred to [42] for a complete review on phylogenetics

a review), and although current research seems to support the so-called *patchwork evolution* model [83], it is still unclear whether other biological mechanisms play a significant role in the emergence of metabolic pathways.

Figure 4.1 explains the process for phylogenetic reconstruction from metabolic similarity. Given a set of organisms, we start by obtaining all their shared pathways. We compute then the similarity among all organisms by aligning their pathways (as described in Chapter 3) and averaging the obtained similarity scores for all pathways, producing a triangular similarity matrix for the set of species. Finally, a hierarchical clustering algorithm is applied to the similarity matrix in order to obtain a phylogenetic tree.



	P1	P2	P3	P4
H. sapiens (hsa)	1	1	1	1
M. musculus (mmu)	1	1	0	1
E. coli (eco)	1	1	0	1
C. elegans (cel)	0	1	0	1
D. melanogaster (dme)	0	1	1	1

$$\text{sim(A,B)=}\ \frac{\text{sim(A,B)[P2]+sim(A,B)[P4]}}{2}$$

Figure 4.1: Phylogenetic reconstruction from metabolic similarity

We will first address the reconstruction of phylogenetic relationships among a set of organisms by using their metabolic similarity obtained from the structural comparison of the glycolysis pathway (Section 4.1.1) [23]. We will then present three different methods to improve results on the reconstructed phylogenetic trees: fuzzy clustering (Section 4.1.2) [17], incremental use of shared metabolic pathways (Section 4.1.3) [25], and filtering “noisy” pathways (Section 4.1.4) [25]. Section 4.1.5 presents some results for phylogenetic reconstruction with photosynthetic organisms [25].

4.1.1 Phylogenetic reconstruction from the glycolysis pathway

Experimental setup

Glycolysis is a metabolic pathway that serves, among other functions, to generate high-energy ATP molecules. This pathway has been thoroughly studied in the literature, being highly conserved in the genetic code and occurring in most species. Because of these characteristics, similarity among different organisms can be studied by analyzing the similarity of their respective glycolysis pathways.

Using the method presented in Chapter 3, we determined the structural similarity of the glycolysis pathway among 73 organisms [23][2]. Table 4.1 presents the full list of organisms, together with their corresponding biological domain and NCBI identifier. Results were obtained for the three enzyme similarity measures described in Section 3.4 (hierarchical, information content, and gene ontology similarity), and for values of the α parameter (see Section 3.3) ranging from 0 to 1 at increments of 0.01.

Table 4.1: Organisms studied, classified by domain (E: eukaryota, B: bacteria, A: archaea), together with their identifier in the NCBI taxonomy

Code	Organism	Domain	NCBI
ATH	*A.thaliana*	E	3702
CEL	*C.elegans*	E	6239
DME	*D.melanogaster*	E	7227
HSA	*H.sapiens*	E	9606
MMU	*M.musculus*	E	10090
RNO	*R.norvegicus*	E	10116
SCE	*S.cerevisiae*	E	4932
SPO	*S.pombe*	E	4896
AAE	*A.aeolicus*	B	224324
ANA	*Anabaena*	B	103690
ATC	*A.tumefaciens_C*	B	176299
ATU	*A.tumefaciens*	B	176299
BHA	*B.halodurans*	B	86665
BME	*B.melitensis*	B	224914
BSU	*B.subtilis*	B	224308
CAC	*C.acetobutylicum*	B	272562
CCR	*C.crescentus*	B	190650
CJE	*C.jejuni*	B	192222
CMU	*C.muridarum*	B	243161
CPA	*C.pneumoniae_AR39*	B	115711
CPJ	*C.pneumoniae_J138*	B	138677
CPN	*C.pneumoniae*	B	115713
CTR	*C.trachomatis*	B	272561
DRA	*D.radiodurans*	B	243230
ECE	*E.coli_O157*	B	155864
ECJ	*E.coli_J*	B	83333
ECO	*E.coli*	B	83333
ECS	*E.coli_O157J*	B	83334
FNU	*F.nucleatum*	B	190304
HIN	*H.influenzae*	B	71421
HPJ	*H.pylori_J99*	B	85963
HPY	*H.pylori*	B	85962
LIN	*L.innocua*	B	272626
LLA	*L.lactis*	B	272623
LMO	*L.monocytogenes*	B	169963
MLE	*M.leprae*	B	272631
MLO	*M.loti*	B	266835
MTC	*M.tuberculosis_CDC1551*	B	83331
MTU	*M.tuberculosis*	B	83332
NMA	*N.meningitidis_A*	B	122587
NME	*N.meningitidis*	B	122586
PAE	*P.aeruginosa*	B	208964
PMU	*P.multocida*	B	272843
RPR	*R.prowazekii*	B	272947
RSO	*R.solanacearum*	B	267608
SAU	*S.aureus_N315*	B	158879
SAV	*S.aureus_Mu50*	B	158878
SCO	*S.coelicolor*	B	100226
SME	*S.meliloti*	B	266834
SPN	*S.pneumoniae*	B	170187
STM	*S.typhimurium*	B	99287
STY	*S.typhi*	B	220341
SYN	*Synechocystis*	B	1148
TMA	*T.maritima*	B	243274
TTE	*T.tengcongensis*	B	273068
VCH	*V.cholerae*	B	243277
XCC	*X.campestris*	B	190485
XFA	*X.fastidiosa*	B	160492
YPE	*Y.pestis*	B	214092
AFU	*A.fulgidus*	A	224325
APE	*A.pernix*	A	56636
HAL	*Halobacterium*	A	64091
MAC	*M.acetivorans*	A	188937
MJA	*M.jannaschii*	A	243232
MMA	*M.mazei*	A	192952
MTH	*M.thermoautotrophicum*	A	187420
PAB	*P.abyssi*	A	272844
PAI	*P.aerophilum*	A	13773
PFU	*P.furiosus*	A	186497
SSO	*S.solfataricus*	A	273057
STO	*S.tokodaii*	A	273063
TAC	*T.acidophilum*	A	273075
TVO	*T.volcanium*	A	50339

From the similarity matrix so obtained, we clustered the organisms using UPGMA (Unweighted Pair Group Method with Arithmetic mean) hierarchical clustering [98], producing phylogenetic trees for each of the previously described parameter settings.

[2] Those which have at least three enzymes annotated to their glycolysis pathway in KEGG

Results and discussion

In order to evaluate the effectiveness of our method, we have compared the produced phylogenies with the NCBI taxonomy (the gold standard for this and following experiments) and the phylogeny for the same organisms in [62]. Figure 4.2 shows one of the produced phylogenies, together with the NCBI taxonomy [138] and the phylogenetic tree from [62, Fig. 2] (best published results to our knowledge).

In order to facilitate comparison of results, we have used the `cousins` algorithm [151] to compute similarity measures between phylogenies. This tool compares unordered trees with labeled leaves, like phylogenetic trees, by comparing in a specific way, and up to a certain cousin distance, the sets of *cousin pairs*, triples consisting of a pair of leaves and their *cousin distance*: 0 if they are siblings (they share the same parent), 0.5 if the parent of one of them is the grandparent of the other, 1 if they are cousins (they share the same grandparent but not the same parent), 1.5 if their last common ancestor is the grandparent of one of them and the great-grandparent of the other one, 2 if they are second cousins (they share the same great-grandparent but not the same grandparent) and so on.

We were unable to reproduce the 0.19 similarity claimed by Heymans and Singh in [62, Table 2] for any parameter of the cousins tool[3], and have therefore adopted the one that gives the closest result (similarity up to second cousins of the trees) for our experiments. Table 4.2 presents results for the glycolysis pathway on 73 organisms when comparing both our method and the method presented in [62] against the NCBI taxonomy. As it can be observed, our method achieves a significantly higher similarity to the NCBI taxonomy.

Table 4.2: Similarity measures based on the NCBI taxonomy for the glycolysis pathway

Technique	Similarity
Our technique	0.1709924
Heymans and Singh's technique	0.1346749

As previously described, the measure of metabolic pathway similarity presented in this thesis is parameterized by the relative weight of compounds and enzymes in the assessment of reaction similarity (Section 3.4). The influence of this parameter in the final phylogenies was also studied, and it was found that results vary with the underlying enzyme similarity measure: while hierarchical enzyme similarity yields the highest metabolic pathway similarity for a weight of about 30%, gene ontology enzyme similarity yields the highest metabolic pathway similarity for a weight between 45% and 65%, and information content enzyme similarity yields the highest metabolic pathway similarity for a weight close to 40% (see Figure 4.3).

From a qualitative point of view, the phylogenetic trees obtained include several biologically relevant clusters. In Figure 4.2 (right), we can appreciate how archaea organisms are clustered in two groups: MTH, MJA, PFU and PAB in the first cluster (with the thermococci PFU and PAB forming a subcluster), and APE, PAI, TVO, TAC, SSO, HAL, AFU, STO, MMA, and MAC in the second (with methanosarcinales MMA and MAC

[3] Both authors were contacted concerning this issue

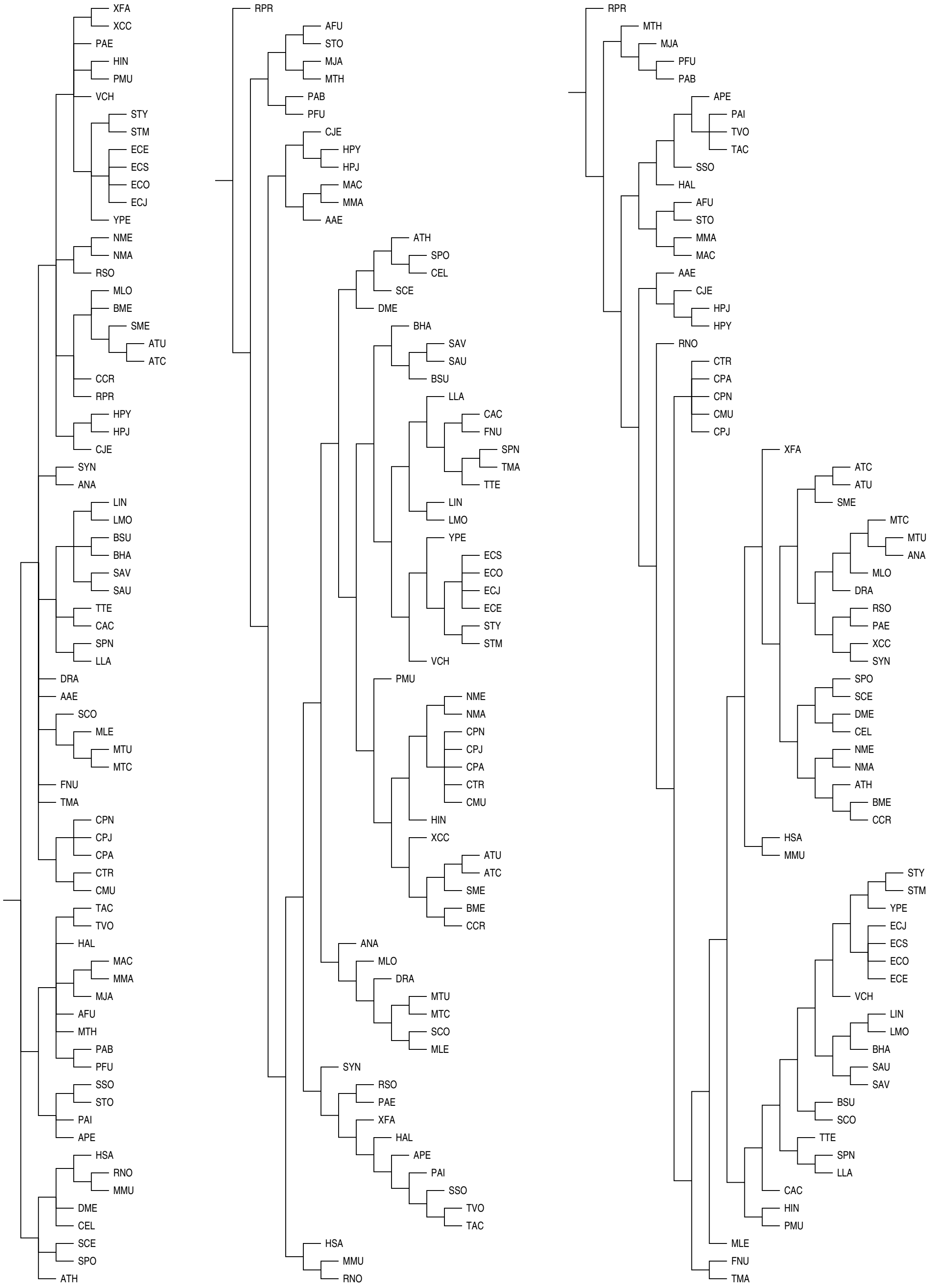


Figure 4.2: Phylogenetic trees obtained from the glycolysis pathway for 73 organisms: NCBI (left), Heymans and Singh (middle), and gene ontology enzyme similarity (right, average-link hierarchical clustering, $\alpha = 40\%$)

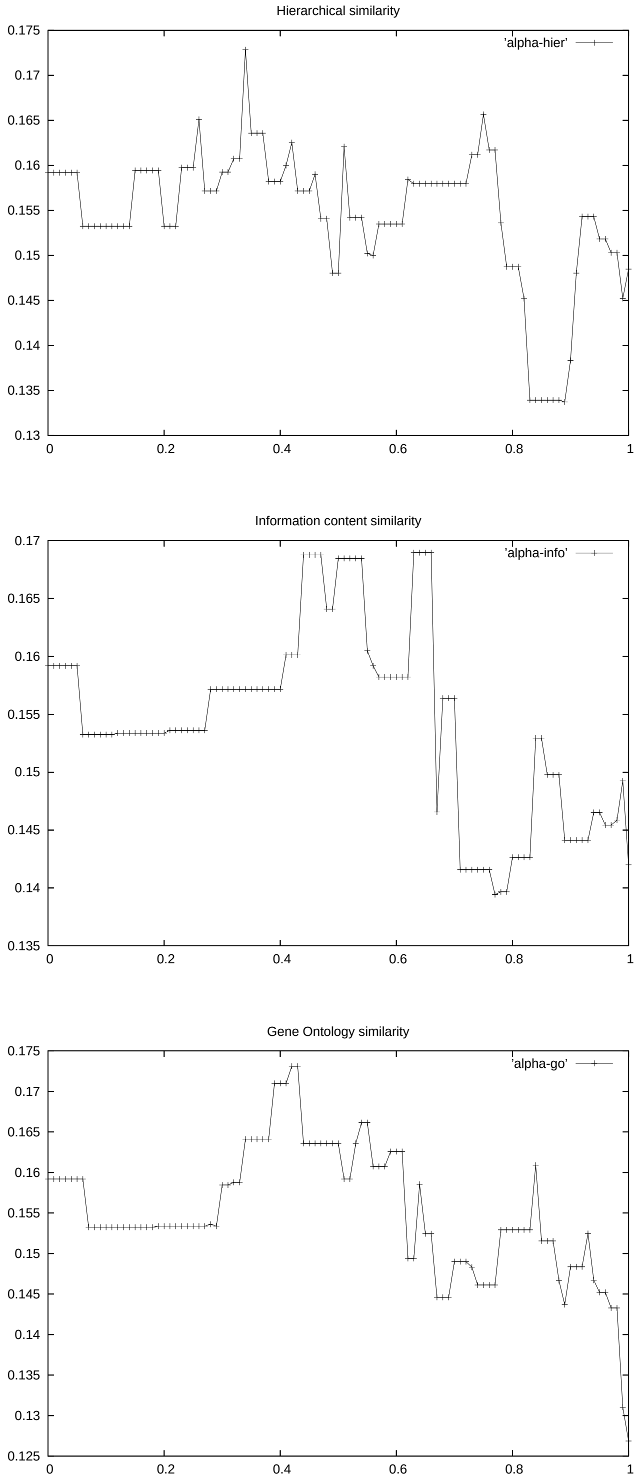


Figure 4.3: Influence of the α weight parameter on metabolic pathway similarity: hierarchical (top), information content (middle), and gene ontology (bottom) enzyme similarity

forming a subcluster). Regarding bacteria, the Chlamydia CPN, CPJ, CPA, CTR, and CMU are also clustered together, as well as the proteobacteria gamma STY, STM, YPE, ECJ, ECS, ECO, and ECE (with the Escherichia in one subcluster and the Salmonella STY and STM in another one). Firmicutes bacillus appear in two main clusters: one for the Bacillales LIN, LMO, BHA, SAU, and SAV, and another one for the Lactobacillales and Clostridia TTE, SPN, and LLA. The proteobacteria delta are also clustered in one group, HPJ, HPY, and CJE.

Despite the goodness of our approach to find relevant clusters, detailed inspection shows we are still far from a fully significative taxonomy. Heymans and Singh's method (Figure 4.2, center), is capable of clustering together all the proteobacteria alpha but one (MLO). The eukaryota also appear grouped into two clusters: mammals (HSA, RNO, MMU), and the remaining eukaryota (DME, SCE, CEL, SPO, and ATH).

Figure 4.4 presents the best trees obtained for each of the three enzyme similarity measures (hierarchical, information content, and gene ontology). Any of these trees produces phylogenies more similar to the NCBI taxonomy than those obtained with previous approaches, and the three of them have a similar clustering of related organisms, showing the robustness of our approach.

4.1.2 Fuzzy clustering improves phylogenetic reconstruction

Although results presented in Section 4.1.1 showed how our method is more similar to the NCBI taxonomy than previous approaches, in this and following sections we will investigate which inconsistencies found in the phylogenies are actually due to our method and which are related to some of the techniques utilized in the reconstruction process (for instance, the hierarchical clustering algorithm) or due to noise present in the datasets.

In this section we will show how by substituting the UPGMA clustering by a fuzzy equivalence relations-based (FER) hierarchical clustering method [101, §4.2], the resulting phylogenetic trees can be further improved [17]. Fuzzy clustering has been successfully used in bioinformatics, mostly through variants of the *fuzzy c-means* (FCM) clustering method. For instance, [113] introduced a method for DNA-based phylogenetic tree reconstruction based on FCM and Markov models.

FCM-based hierarchical clustering methods have the disadvantage of requiring the desired number of clusters be given *a priori* in each step. Alternatively, all possible number of clusters must be tried and then the optimal number chosen according to some "least fuzzy partitions" criterion, although this method is computationally expensive. The FER clustering method overcomes these drawbacks: it is faster, logically simpler, and naturally hierarchical [148]. Although it has found several applications in health sciences (see [101, Ch. 4] and the references therein), to our knowledge it has only been used once to produce phylogenetic trees [93].

In the FER clustering method, we determine a fuzzy similarity relation S (reflexive and symmetric) on the set of objects and compute the fuzzy equivalence relation E generated by this similarity, as the max-min transitive closure of the matrix of S. Then, for each t appearing in E's matrix, the *t-cut* crisp equivalence relation obtained by replacing every entry in E's matrix smaller than t by 0 and every entry greater than or equal to t by 1, induces a crisp partition of the set of objects: each element of the partition is a maximal subset of objects that have "E-equivalence value" $\geqslant t$ with each other. These partitions, together with the hierarchy induced by the increasing order of the values t, yields a

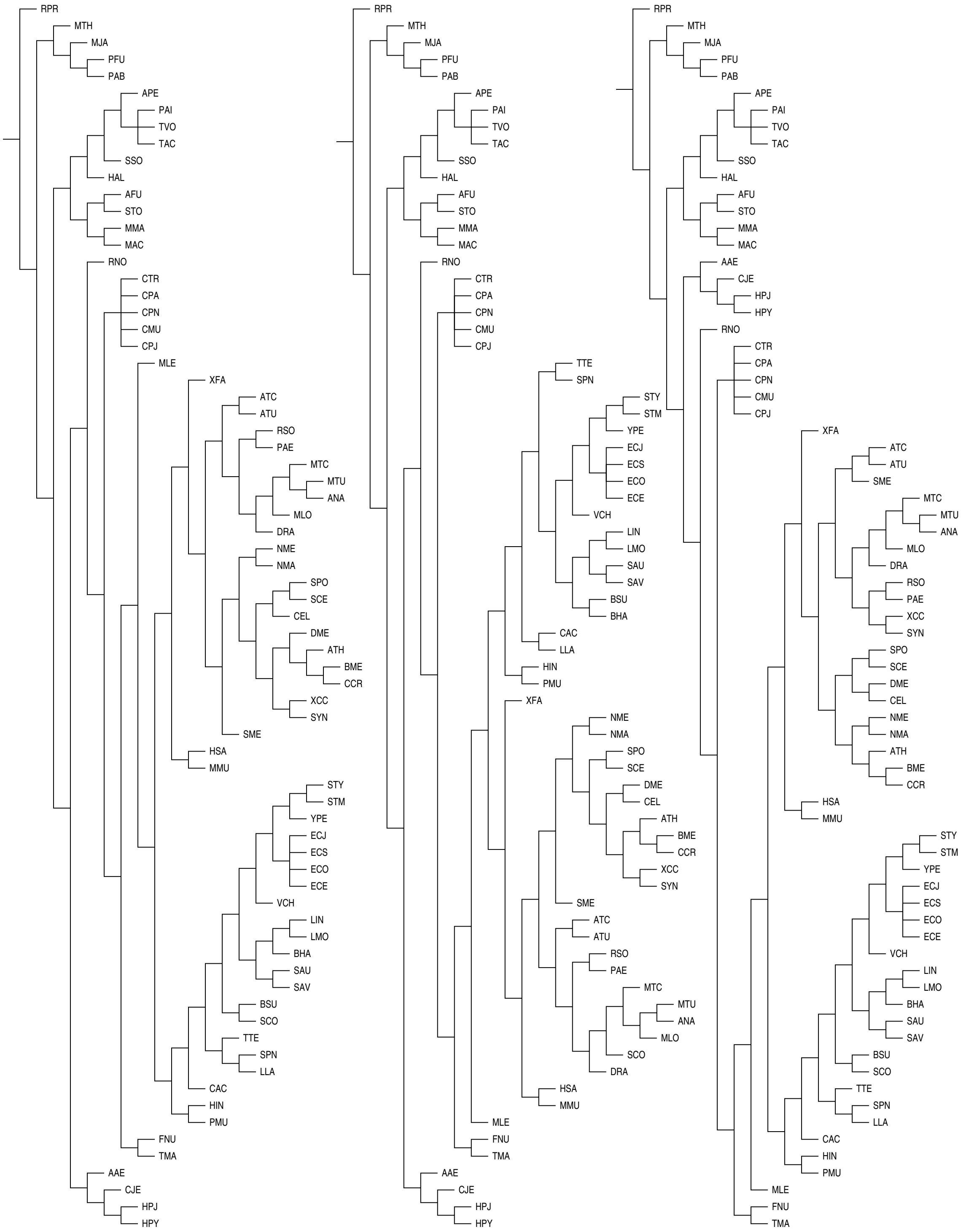


Figure 4.4: Phylogenetic trees obtained from the glycolysis pathway for 73 organisms using average-link hierarchical clustering: hierarchical enzyme similarity (left, $\alpha = 30\%$), information content enzyme similarity (middle, $\alpha = 50\%$), and gene ontology enzyme similarity (right, $\alpha = 40\%$)

classification tree for the objects.

For instance, consider the similarity matrix

	MGE	HIN	MTU	MJA	ECO	AFU
MGE	1.00	0.33	0.07	0.02	0.17	0.22
HIN	0.33	1.00	0.33	0.32	0.34	0.27
MTU	0.07	0.33	1.00	0.09	0.20	0.20
MJA	0.02	0.32	0.09	1.00	0.18	0.24
ECO	0.17	0.34	0.20	0.18	1.00	0.32
AFU	0.22	0.27	0.20	0.24	0.32	1.00

on the set of organisms

$$\{\mathrm{MGE}, \mathrm{HIN}, \mathrm{MTU}, \mathrm{MJA}, \mathrm{ECO}, \mathrm{AFU}\}$$

(see Table 4.3). The fuzzy equivalence relation generated by this similarity is given by the matrix

$$\begin{pmatrix} 1.00 & 0.33 & 0.33 & 0.32 & 0.33 & 0.32 \\ 0.33 & 1.00 & 0.33 & 0.32 & 0.34 & 0.32 \\ 0.33 & 0.33 & 1.00 & 0.32 & 0.33 & 0.32 \\ 0.32 & 0.32 & 0.32 & 1.00 & 0.32 & 0.32 \\ 0.33 & 0.34 & 0.33 & 0.32 & 1.00 & 0.32 \\ 0.32 & 0.32 & 0.32 & 0.32 & 0.32 & 1.00 \end{pmatrix}$$

The hierarchy of partitions of the set of organisms defined by the t-cuts of this fuzzy equivalence relation is:

t	Partition corresponding to the t-cut
1.00	{MGE} {HIN} {MTU} {MJA} {ECO} {AFU}
0.34	{HIN, ECO} {MGE} {MTU} {MJA} {AFU}
0.33	{MGE, HIN, MTU, ECO} {MJA} {AFU}
0.32	{MGE, HIN, MTU, MJA, ECO, AFU}

This hierarchical clustering yields a classification tree, depicted as a dendogram in Figure 4.5, and which is very close to the NCBI taxonomy tree for these six organisms, the only difference being that in the latter the archaea MJA and AFU are also clustered (that is, they should be at cousin distance 1 from MGE and MTU, instead of 0.5).

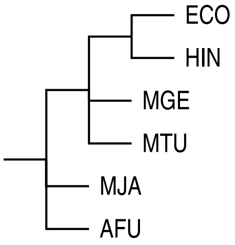


Figure 4.5: A dendogram for MGE, HIN, MTU, MJA, ECO, and AFU similar to their NCBI taxonomy

Experimental setup

To compare the performance of FER clustering and UPGMA clustering, we computed the similarities defined in 3 of the glycolysis pathways of a model set of 16 organisms presented in table 4.3. From this information, we reconstructed the corresponding phylogenetic trees and computed their similarity to the NCBI taxonomy corresponding to those 16 organisms using the `cousins` tool up to cousin distance 2, for each of the three enzyme similarity measures, and for each $\alpha = 0, 0.1, 0.2, \ldots, 0.9, 1$. This produced 33 16×16 matrices with entries in $[0, 1]$. These matrices are symmetrical and all entries in their main diagonal are 1. We have then clustered the 16 organisms based on these similarities, using UPGMA clustering [98].

To use the FER hierarchical clustering on these matrices, we computed their max-min transitive closure using the algorithm derived from [101, Thm. 4.2.1]. In this way we obtained the matrix of the fuzzy equivalence generated by each one of the 33 similarity matrices on the set of the 16 organisms. We have then computed the classification tree given by each one of these fuzzy equivalences, and considered them as the phylogenetic trees for the set of organisms.

Table 4.3: Organisms studied, classified by domain (A: archaea, B: bacteria, E: eukaryota), together with their identifier in the NCBI taxonomy

AFU	*A. fulgidus*	A	224325
MJA	*M. jannaschii*	A	243232
CPN	*C. pneumoniae*	B	115713
MGE	*M. genitalum*	B	243273
MPN	*M. pneumoniae*	B	272634
HIN	*H. influenzae*	B	71421
SYN	*Synechocystis*	B	1148
DRA	*D. radiodurans*	B	243230
MTU	*M. tuberculosis*	B	83332
TPA	*T. pallidum*	B	243276
BSU	*B. subtilis*	B	224308
AAE	*A. aeolicus*	B	224324
TMA	*T. maritima*	B	243274
ECO	*E. coli*	B	83333
HPY	*H. pylori*	B	85962
SCE	*S. cerevisiae*	E	4932

Results and discussion

We were unable to reproduce the 0.27 similarity claimed in [61, Table 5] for any parameter of the `cousins` tool, and we have therefore adopted the parameter setting that provides the closest result (similarity up to second cousins, that is, up to cousin distance 2) for our experiments, which yields a similarity of 0.1935484 between NCBI's and Heymans-Singh's trees.

Table 4.4: Similarity values for both clustering methods and all similarity measures

	UPGMA	*FER*		*UPGMA*	*FER*		*UPGMA*	*FER*
$go_{0.0}$	0.2386	0.2604	$hier_{0.0}$	0.2386	0.2604	$info_{0.0}$	0.2386	0.2604
$go_{0.1}$	0.2386	0.2736	$hier_{0.1}$	0.2222	0.2736	$info_{0.1}$	0.2386	**0.3125**
$go_{0.2}$	0.2222	**0.3195**	$hier_{0.2}$	0.2222	**0.3020**	$info_{0.2}$	0.2222	0.2526
$go_{0.3}$	0.2222	0.2736	$hier_{0.3}$	0.2222	0.2903	$info_{0.3}$	0.2222	0.3020
$go_{0.4}$	0.2222	0.2659	$hier_{0.4}$	0.2222	0.2553	$info_{0.4}$	0.2222	0.2234
$go_{0.5}$	0.2222	0.2659	$hier_{0.5}$	0.2222	0.2842	$info_{0.5}$	0.2386	0.2105
$go_{0.6}$	0.2222	0.2307	$hier_{0.6}$	0.2222	0.2340	$info_{0.6}$	0.2386	0.1827
$go_{0.7}$	0.2222	0.2127	$hier_{0.7}$	0.2222	0.2197	$info_{0.7}$	0.2386	0.1413
$go_{0.8}$	0.2222	0.2947	$hier_{0.8}$	**0.2527**	0.2340	$info_{0.8}$	**0.2777**	0.1505
$go_{0.9}$	0.2222	0.2600	$hier_{0.9}$	**0.2527**	0.1935	$info_{0.9}$	**0.2777**	0.1630
$go_{1.0}$	**0.2527**	0.2043	$hier_{1.0}$	**0.2527**	0.1956	$info_{1.0}$	**0.2777**	0.1868

Table 4.4 shows the similarity values to the NCBI taxonomy tree of the phylogenetic trees obtained through UPGMA clustering (column *UPGMA*) and through FER hierarchical clustering (column *FER*) for each of the similarity measures go_α, $hier_\alpha$ and $info_\alpha$, $\alpha = 0, 0.1, 0.2, \ldots, 0.9, 1$.

As it can be seen in Table 4.4, the gene ontology similarity yields better results when using the FER clustering method. For all values of α except $\alpha = 1$ (that is, except when compound similarity is not taken into account) and $\alpha = 0.7$, the FER tree is closer to the NCBI taxonomy than the UPGMA tree. The greatest similarity is $go_{0.2}$ with FER (almost 0.32), while the maximum with UPGMA is $go_{1.0}$ (slightly under 0.253). The average similarity of the FER trees to the NCBI taxonomy is 0.260, while the average similarity of the UPGMA trees is 0.228.

FER clustering also generates better trees than UPGMA for the hierarchical similarity. The FER tree is closer to the NCBI taxonomy than the UPGMA tree for all values of α except for all $\alpha \geqslant 0.7$. The greatest similarity is reached again for $hier_{0.2}$ using FER (slightly over 0.3), while the maximum with UPGMA is reached for $hier_\alpha$ with $\alpha \geqslant 0.8$ (slightly under 0.253). The average similarity of the FER trees to the NCBI taxonomy is in this case 0.249, while the average similarity of the UPGMA trees is 0.232.

Interestingly, FER behaves worse than UPGMA for the information content similarity: the FER tree is more similar to the NCBI taxonomy than the UPGMA tree for all $\alpha \leqslant 0.4$, while the UPGMA tree is better when $\alpha \geqslant 0.5$. The greatest similarity is reached in this case for $info_{0.1}$ with FER (slightly over 0.31), while the maximum with UPGMA is obtained for $info_\alpha$ with $\alpha \geqslant 0.8$ (slightly above 0.277).

The better performance of FER when using gene ontology and hierarchical similarity might be explained by the fact that these measures are conceptually similar, since both are based on shortest path distance among enzymes (in the GO graph and the EC tree, respectively). On the other hand, information content similarity (where FER obtained worse results) is based on EC subtree size, which results in a more fine-grained measure than gene ontology or hierarchical similarity.

It is also evident from Table 4.4 that for all three types of enzyme similarity the best results are obtained using FER and low values of α. Indeed, if we only take into account the values $\alpha = 0, \ldots, 0.4$, the average similarity of the phylogenetic trees to the NCBI taxonomy is 0.2786602 for *go*, 0.27636518 for *hier* and 0.27020718 for *info*.

The best phylogenetic trees obtained with FER clustering for each one of the three enzyme similarity measures are shown in Figure 4.6.

4.1.3 Incremental reconstruction

In Section 4.1.1 we presented results for reconstruction of phylogenetic trees from the similarity of the glycolysis pathway for a set of organisms. It comes as a natural idea to use information from more than one single pathway to reconstruct phylogenetic relationships among species. In this section, we will extended previous results by performing a series of additional experiments that will take into account all shared pathways for a set of organisms containing at least one reaction. Intuitively, we would expect to obtain better results (phylogenetic trees closer to the NCBI taxonomy) since we now use a whole range of metabolic processes rather than focusing only in the glycolysis pathway [25].

Experimental setup

A first experiment consisted in the phylogenetic reconstruction of a set of organisms based on the comparison of all of their common metabolic pathways. The organisms chosen were *Archaeoglobus fulgidus* (afu), *Clostridium perfringens* (cpe), *Haemophilus influenzae* (hin), *Listeria innocua* (lin), *Methanocaldococcus jannaschii* (mja), *Mus musculus* (mmu), *Neisseria meningitidis* (nme), and *Rattus norvegicus* (rno). In a second experiment, we have performed the incremental phylogenetic reconstruction of the same set of 8 organisms by adding their 53 common metabolic pathways one by one.

The *alpha* parameter was set to 0.5 (equal weight for enzymes and compounds in the reaction similarity measure), similarity of enzymes was calculated using hierarchical similarity, and the phylogenetic trees were produced using UPGMA. Similarity of obtained phylogenies were compared against the NCBI taxonomy (Figure 4.7) for the set of chosen organisms using the `cousins` tool, up to second cousins.

Results and discussion

As can be seen in Figure 4.7, the phylogenetic reconstruction obtained in the first experiment based on the 53 common metabolic pathways produced a phylogeny which is very close to the NCBI taxonomy.

In the second experiment, where pathways were gradually incorporated and then similarity to the NCBI taxonomy calculated, it was observed how the addition of a metabolic pathway does not always have a consequence on the resulting phylogeny and, as illustrated in Figure 4.8, incremental phylogenetic reconstruction produces only 11, instead of 53, different phylogenies for the set of 8 organisms. Figure 4.8 also shows that second cousins similarity to the NCBI taxonomy increases gradually when adding metabolic pathways until reaching a highest similarity value.

4.1.4 Filtering of *noisy* pathways

Analysis of results presented in Section 4.1.3 showed how some of the shared metabolic pathways among the organisms tend to be too dissimilar (suggesting incomplete metabolic

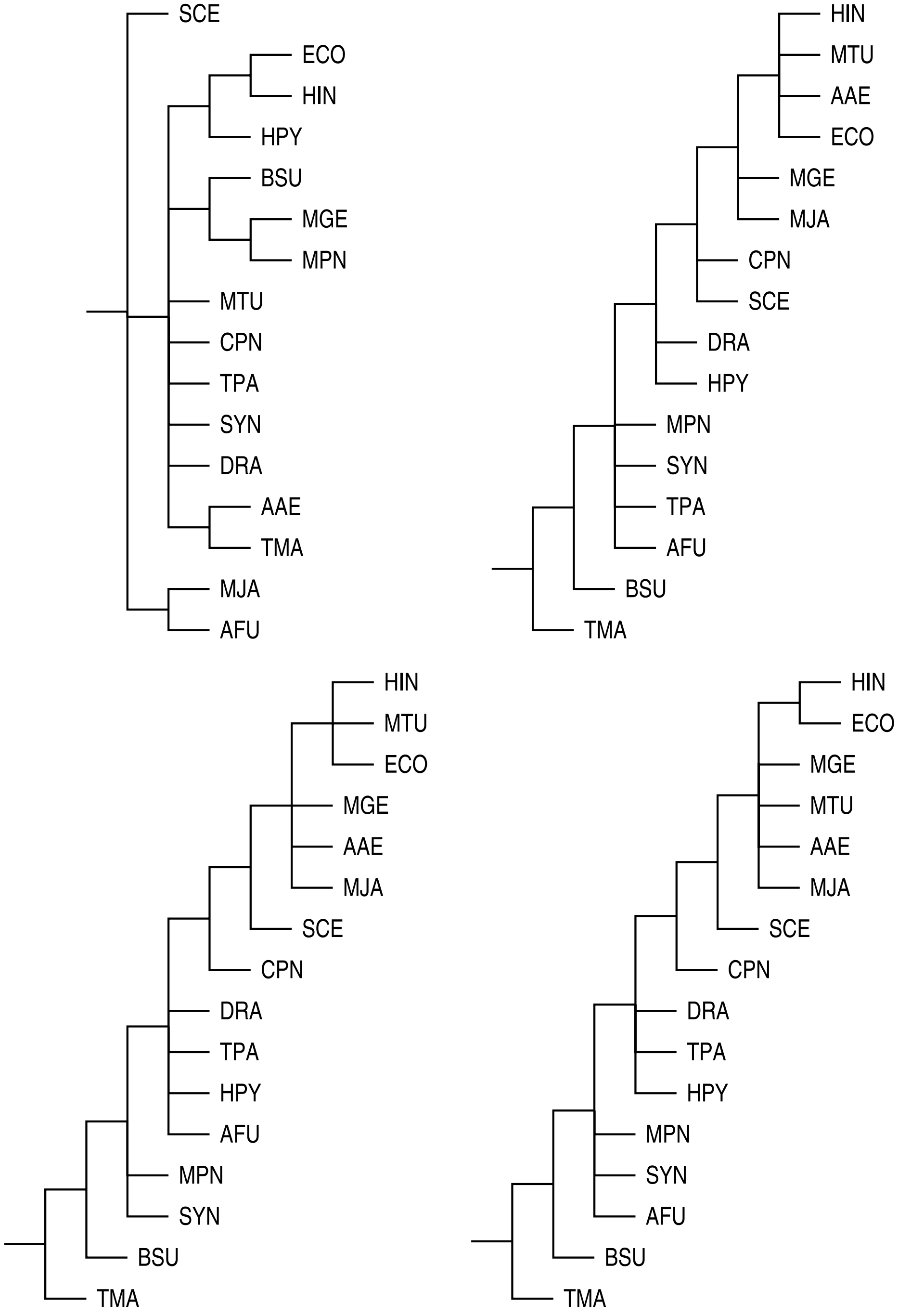


Figure 4.6: Phylogenetic tree for the set of 16 organisms (NCBI taxonomy, top left) and best trees obtained with FER clustering from the similarity of their glycolysis pathways, using gene ontology (top right), hierarchical (bottom left), and information content enzyme similarity (bottom right)

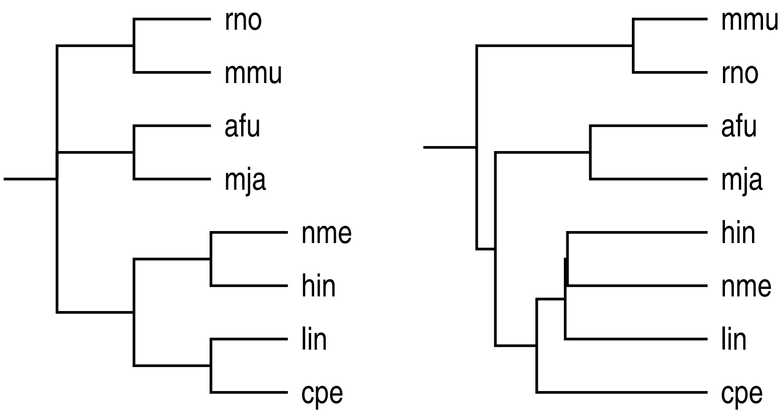


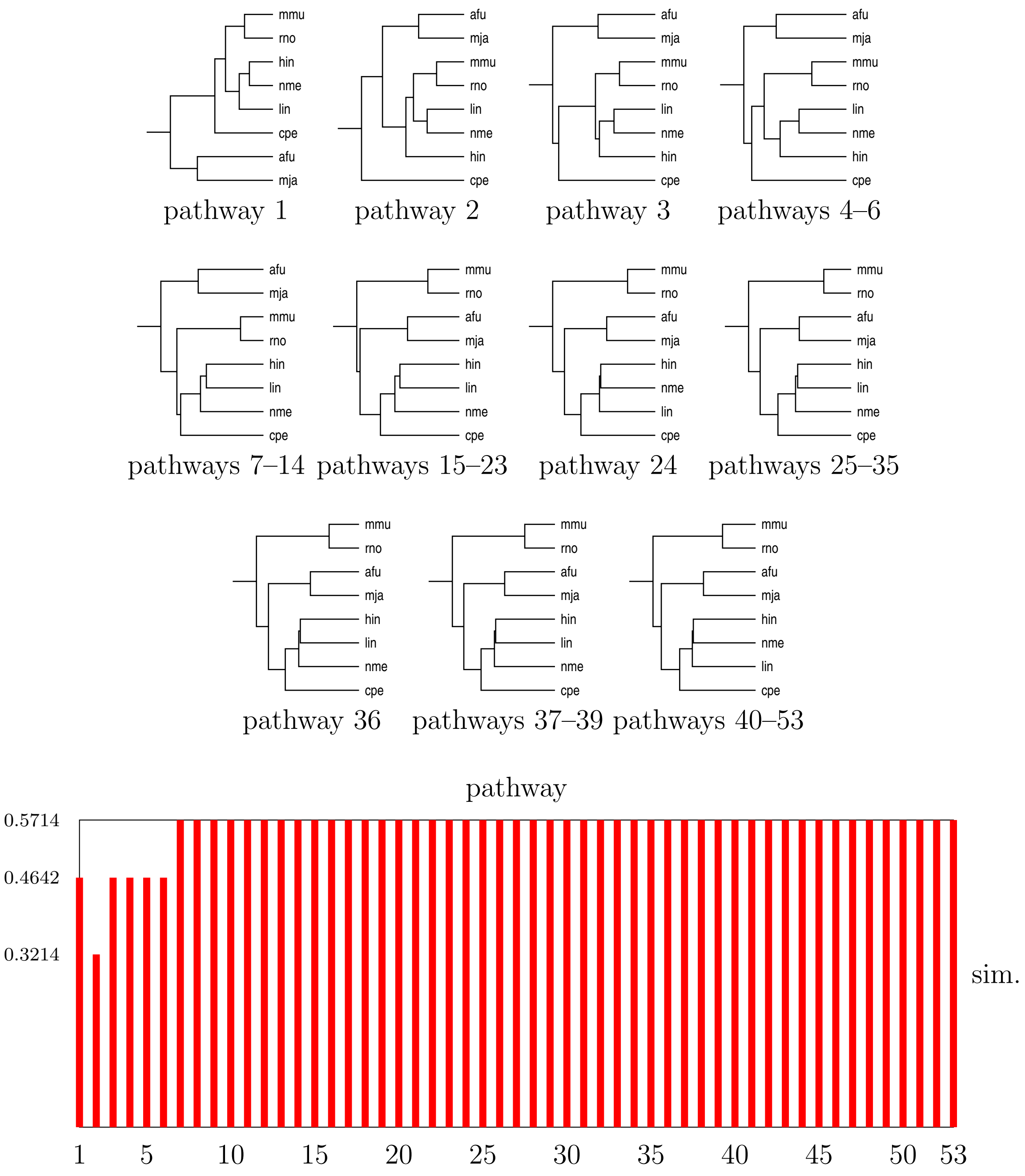
Figure 4.7: Left: NCBI taxonomy for the 8 organisms *Archaeoglobus fulgidus* (afu), *Clostridium perfringens* (cpe), *Haemophilus influenzae* (hin), *Listeria innocua* (lin), *Methanocaldococcus jannaschii* (mja), *Mus musculus* (mmu), *Neisseria meningitidis* (nme), and *Rattus norvegicus* (rno). Right: phylogenetic reconstruction based on the 53 metabolic pathways common to the 8 organisms

information in organisms with extremely few reactions) or too similar (even for organisms phenotypically very different, suggesting the metabolic pathway data is still not complete in all organisms, and is therefore useless for phylogenetic reconstruction purposes). The main repository of information on metabolic pathways used in this thesis, KEGG, is an ongoing effort that automatically incorporates data which is afterwards manually reviewed by annotators. Information stored in KEGG should therefore be considered as tentative only, and in this section we present a basic method to detect and discard pathways which are suspected to be incomplete and would therefore introduce “noise” in the phylogenetic reconstruction process [25].

Experimental setup

We repeated the incremental phylogenetic reconstruction experiment on the same set of 8 organisms presented in Section 4.1.3, discarding some of the 53 common metabolic pathways according to one of three following criteria:

- Discard those metabolic pathways where the distance of an organism to its closest neighbour is more than a certain threshold, for a percentage of all organisms. The intuition behind this criteria is to remove pathways where some of the organisms have very few annotated reactions, producing a non-meaningful mapping among reactions and making the organisms appear more dissimilar than they are in reality (and hence the artificially high distance).
- Discard those metabolic pathways where the distance of an organism to its closest neighbour is equal to zero, for a percentage of all organisms. Information about some metabolic pathways appears to be still incomplete in most organisms, producing an artificially high metabolic similarity score between clearly different organisms.
- Discard those metabolic pathways where either the number of reactions is less than a certain threshold, or one organism has significantly less reactions than the or-



pathway 1 pathway 2 pathway 3 pathways 4–6

pathways 7–14 pathways 15–23 pathway 24 pathways 25–35

pathway 36 pathways 37–39 pathways 40–53

Figure 4.8: Incremental phylogenetic reconstruction for the 8 organisms *Archaeoglobus fulgidus* (afu), *Clostridium perfringens* (cpe), *Haemophilus influenzae* (hin), *Listeria innocua* (lin), *Methanocaldococcus jannaschii* (mja), *Mus musculus* (mmu), *Neisseria meningitidis* (nme), and *Rattus norvegicus* (rno), from their 53 common metabolic pathways (top). Second cousins similarity of the reconstructed phylogenies with respect to the NCBI taxonomy (bottom)

ganism with more reactions in the metabolic pathway. This criteria is somehow a combination of the two previous ones, trying to eliminate those pathways for which information is suspected to be incomplete.

Parameters for our metabolic similarity measure were set to the same values as in previous sections: $\alpha = 0.5$, UPGMA clustering, and enzyme hierarchical similarity.

Results and discussion

Best results for the first filtering criteria were obtained with a distance threshold of 0.4 and 10% of the organisms, where the remaining 25 metabolic pathways produced a top second cousins similarity score of 0.714, improving those presented in 4.8. Notice though how the similarity curve now reaches a maximum after adding 7 pathways, and drops to the previous 0.571 once we add more than 10 metabolic routes (Figure 4.9, left).

For the second criteria, best results were obtained using a threshold of 20% of the organisms having distance 0 to its closes neighbour. The remaining 37 metabolic pathways produce similar results to the previous criteria, with best second cousins similarity score of 0.714 using 10 to 12 and 30 to 37 pathways (Figure 4.9, center).

For the last criteria, a reaction threshold of 0.3 reduces the set to 19 metabolic pathways, with best second cousins similarity of score of 0.714 using 5 to 8 pathways (Figure 4.9, right).

Results using these three criteria show how the method for metabolic pathway similarity presented in this thesis can produce phylogenetic trees quite similar to the correct NCBI taxonomy, provided that information on enough metabolic pathways is available, and that "noisy" pathways are removed by some criteria. Figure 4.10 presents the best phylogeny obtained after filtering.

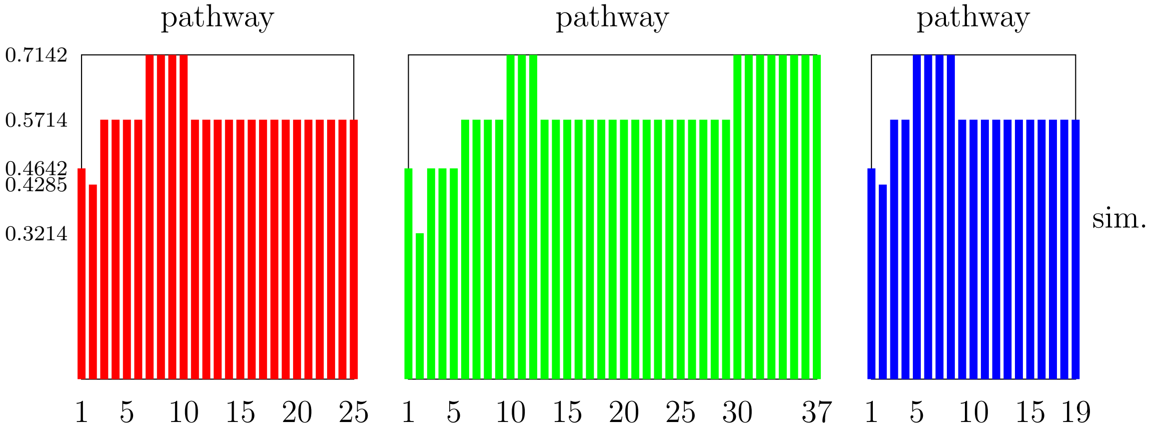


Figure 4.9: Incremental phylogenetic reconstruction for 8 organisms and a subset of their common metabolic pathways chosen according to three criteria: *too dissimilar pathways* (left), *too similar pathways* (center), and *too few reactions* (right)

As shown in [21], the use of filters to eliminate sequences that are phylogenetically discordant can improve the reconstruction of genome trees. Following a similar reasoning, we argue that filtering of noisy metabolic pathways can help in reconstructing better phylogenies. It is unclear though if even removing such discordant pathways we will ever

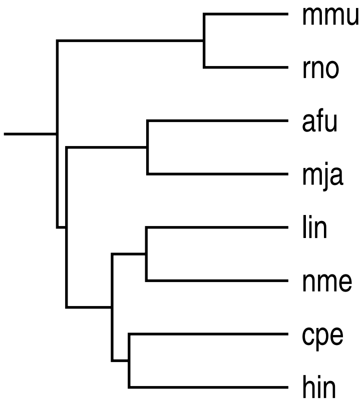


Figure 4.10: Most similar tree to NCBI taxonomy for set of 8 organisms, after removing *noisy* pathways

beat results obtained using the 16s rRNA gene. In fact, it has been suggested that despite expanding data sets and alternative markers for inferring phylogenetic relationships, 16s rRNA will remain as the gold standard in the field [92] [97, Ch. X,XI]. Nevertheless, it should be noticed that this work is focused on evolution of metabolism rather than in genome evolution, and therefore different results with trees reconstructed from sequence data are to be expected.

4.1.5 Photosynthetic organisms

The experiments discussed in previous sections used a set of organisms belonging to the three domains of life, eukarya, bacteria and archaea [142][4]. It could be argued that reconstructing phylogenetic trees from organisms with such different metabolic processes is a somehow non-complex problem, and we have therefore performed one more experiment to reconstruct the phylogenetic relationships among a set of 10 photosynthetic organisms: eight bacteria, and two eukaryota containing chloroplasts [25].

Chloroplasts descended from cyanobacteria [95], and despite horizontal gene transfer of many ancestral cyanobacterial genes to the plant nuclear genome, a selective set of metabolic pathways is maintained in chloroplasts [135]. As a result, these eight photosynthetic bacteria share a large number of metabolic pathways with the two photosynthetic eukaryotes, which makes it particularly hard to distinguish them by comparison of their metabolism only.

Experimental setup

We utilized the 61 shared metabolic pathways containing at least one reaction annotated to the set of selected organisms: the photosynthetic bacteria *Anabaena* (ana), *Gloeobacter violaceus* (gvi), *Prochlorococcus marinus marinus* (pma), *Prochlorococcus marinus pastoris* (pmm), *Prochlorococcus marinus* (pmt), *Synechocystis* (syn), *Synechococcus* (syw),

[4] Whether there are two or three domains, and how are they subdivided, has been extensively discussed in the literature [96, 139, 142]. In this work we will assume the existence of three domains as postulated in [143], which is currently recognized as the most plausible hypothesis

and *Thermosynechococcus elongatus* (tel), and the eukaryota *Arabidopsis thaliana* (ath), *Cyanidioschyzon merolae* (cme). We then applied our method for metabolic similarity by incrementally adding all the pathways and reconstructing a set of phylogenetic trees from them. Parameters for our algorithm were set to the same values as in previous sections: $\alpha = 0.5$, UPGMA clustering, and enzyme hierarchical similarity.

Results and discussion

As it can be seen in Figure 4.11, the incremental phylogenetic reconstruction based on 61 common metabolic pathways produced a phylogeny that clearly separates the photosynthetic eukaryotes from the photosynthetic bacteria. Inside the bacteria cluster, our method was not able to separate the Chroococcales (*Synechocystis*, *Synechococcus*, *Thermosynechococcus elongatus*) from the Prochlorales (*Prochlorococcus marinus marinus*, *Prochlorococcus marinus pastoris*, *Prochlorococcus marinus*), the Nostocales (*Anabaena*), and the Gloeobacterales (*Gloeobacter violaceus*). It should be noticed though that parameter fine-tuning (election of enzyme similarity measure, α value, hierarchical clustering method) and removal of "noisy" pathways could improve the results in the final phylogeny.

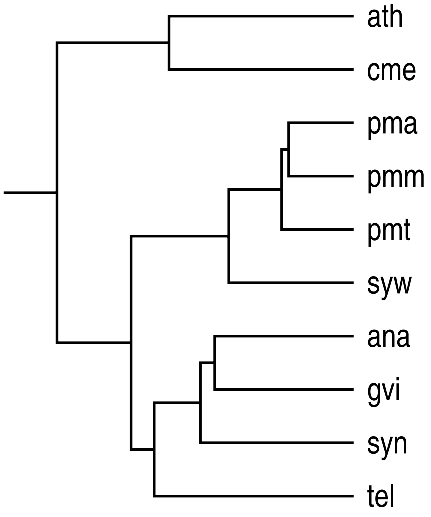


Figure 4.11: Tree reconstructed for the eight photosynthetic bacteria *Anabaena* (ana), *Gloeobacter violaceus* (gvi), *Prochlorococcus marinus marinus* (pma), *Prochlorococcus marinus pastoris* (pmm), *Prochlorococcus marinus* (pmt), *Synechocystis* (syn), *Synechococcus* (syw), *Thermosynechococcus elongatus* (tel) and the two photosynthetic eukaryotes *Arabidopsis thaliana* (ath), *Cyanidioschyzon merolae* (cme) when using 61 of their 68 common pathways

4.2 Conserved and non-conserved reactions

In previous sections, we have utilized our algorithm to obtain metabolic similarity values among organisms. Using those similarity values, we can reconstruct phylogenetic trees depicting the common metabolic history of a group of organisms, or find suitable model organisms in the study of diseases related to certain metabolic conditions. In this section we will further expand the range of possible applications of our method by investigating the significance behind the pathway alignments that we produce, and how to detect conserved and non-conserved metabolic reactions in a set of organisms [24].

Let us first consider how our approach produces an alignment between different reactions in the metabolism of two organisms. Borrowing terminology from sequence alignment, we will define the alignment of reactions as a *perfect match*, a *substitution* or a *gap*. In a perfect match, both reactions are composed of the same set of metabolites and enzymes, and our algorithm would score their similarity as 1. Figure 4.12 has several examples of perfect matches between reactions in the TCA cycle of *Archaeoglobus fulgidus* and *Listeria innocua* marked in blue/solid (00268, 00344, 00351, 00412, 01082 and 01899). Substitutions occur when the reactions are not exactly the same, but share some compounds or enzymes. In such case, the similarity value will be greater than 0 but less than 1. Figure 4.12 contains three such substitutions: reactions 00342, 00405 and 01197 in *Archaeoglobus fulgidus* are respectively substituted by 01082, 00412 and 00268 in *Listeria innocua* (alignment marked in green/dashed lines). Finally, gaps occur when a reaction in one of the organisms cannot be mapped with similarity greater than 0 in the other organism. Figure 4.13 presents an example of a gap: reaction 01698 (underlined) in *Listeria innocua* has similarity 0 to any reaction in *Archaeoglobus fulgidus*, and therefore creates a gap in the alignment. Biologically, a gap between species A and B represents a reaction which, in the course of evolution, has been gained by A ("insertion" event) or lost by B ("deletion"). Generalizing for a set of organisms $\{A, B, C, \ldots, Z\}$, a reaction is said to be a gap when it is present in one of the organisms and cannot be aligned with any of the remaining organisms with similarity greater than 0.

Notice that because we are aligning unordered sets of reactions instead of ordered sequences of nucleotides or amino acids, our algorithm produces a directed alignment since the mapping of reactions is not symmetric. Reaction 00342 in *Archaeoglobus fulgidus* is aligned to reaction 01082 in *Listeria innocua* (Figure 4.12), but the opposite is not true since reaction 01082 in *Listeria innocua* gets aligned to 01082 in *Archaeoglobus fulgidus* (Figure 4.13). Although for simplicity we have described an alignment between two organisms only, the same method can be applied to a set of species by calculating all their respective alignments.

In this section, we will study in detail perfect matches and gaps. Perfect matches are interesting since they represent highly conserved reactions in the metabolism of different organisms. Vital biological processes in a group of related species (taxons such as bacteria or archaea) should be conserved and expressed by a significant number of reactions in all the organisms of the group. We will validate this hypothesis by studying perfect matches among bacteria, archaea and eukarya in Section 4.2.2.

Gap reactions, on the other hand, are interesting since they imply the complete absence of a certain group of metabolites and enzymes in one of the organisms being compared. More specifically, if the organisms being compared are known to be similar we would not expect to find many gaps in the alignment of their metabolism. We will test this hypothesis by studying the alignment of a set of strains (genetic variants of an organism) in Section 4.2.3.

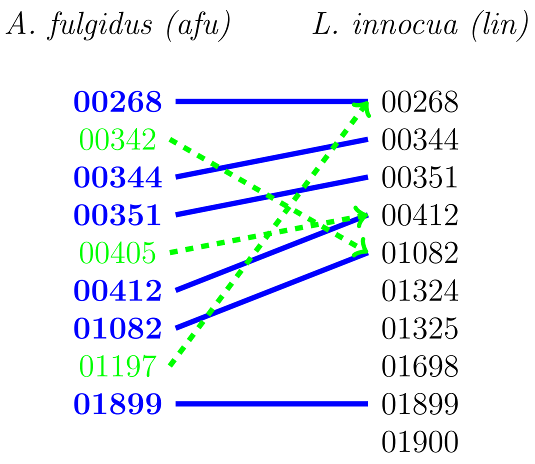


Figure 4.12: Alignment of TCA cycle from *Archaeoglobus fulgidus* to *Listeria innocua*. Reactions 00342, 00405 and 01197 in *afu* (green/dashed) are substituted in *lin* for 01082, 00412 and 00268 with similarity < 1; rest of reactions are perfectly aligned, similarity = 1 (blue/solid)

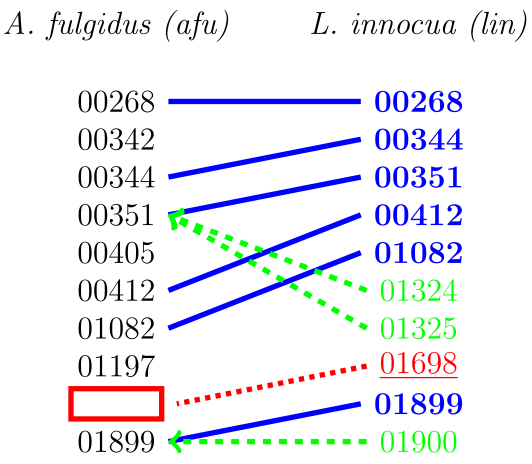


Figure 4.13: From *Listeria innocua* to *Archaeoglobus fulgidus*: 01324, 01325 and 01900 (green/dashed) are substituted for 00351, 00351 and 01899 with similarity < 1. 01698 (red/dotted) cannot be mapped with similarity > 0. Remaining reactions are perfectly aligned (blue/solid)

4.2.1 Experimental setup

All data used was obtained from KEGG release 39.0 (July 2006). For each experiment, we selected a set of organisms and retrieved all shared pathways which contained at least one reaction. We then applied our metabolic pathway alignment algorithm using hierarchical enzyme similarity and parameter $\alpha = 0.5$.

Experiments described in section 4.2.2 utilized the obtained alignment to select all reactions that were aligned with similarity 1 (perfect alignments) for all organisms in a taxon: bacteria, archaea, eukarya, mammals, and plants. With this set of highly conserved reactions, we then calculated the percentage of each pathway that was conserved as the number of conserved reactions appearing in the pathway divided by the total number of reactions of the pathway.

Experiments in section 4.2.3 used results from the pathway alignment to obtain gap reactions in any strain, i.e., reactions that could not be mapped with similarity greater than 0 to any of the remaining strains, thus inducing a gap in the alignment. Finally, we calculated the number and type of enzymes involved in such reactions by counting the number of reactions in which each enzyme appeared.

4.2.2 Conserved reactions

Given a set of organisms, we expect reactions conserved with high similarity to perform functions related to biologically relevant processes in the set. We performed experiments in a group of bacteria, archaea and eukarya and their shared metabolic pathways (Table 4.5). Ideally, reactions conserved with high similarity only in one of the domains and not in the others should reflect some property exclusive to that specific domain.

Table 4.5: Conserved reactions among bacteria, archaea and eukarya: organisms and shared pathways

Taxon	**Organisms**	**Shared path.**
Bacteria	*aae ana atc atu bha bme bsu cac ccr cje cmu cpa cpj cpn ctr dra ece ecj eco ecs fnu hin hpj hpy lin lla lmo mle mlo mtc mtu nma nme pae pmu rpr rso sau sav sco sme spn stm sty syn tma tte vch xcc xfa ype*	00010 00061 00190 00230 00240 00251 00252 00260 00271 00272 00280 00290 00310 00330 00400 00450 00500 00550 00564 00620 00630 00640 00650 00670 00710 00790 00860 00900 00970
Archaea	*afu ape hal mac mja mma mth pab pai pfu sso sto tac tvo*	
Eukarya	*ath cel dme hsa mmu rno sce spo*	

As seen in Table 4.6(a), Bacteria were very clearly characterized. Most of the reactions exclusively conserved with high similarity in bacteria belong to the fatty acid biosynthesis pathway (00061), with 64% of the total number of reactions in the pathway being conserved. The peptidoglycan biosynthesis pathway (00550) also has over 37% of its reactions highly conserved in bacteria. Remnant pathways were conserved in lower proportions.

Archaea had a significant number of conserved reactions related to the phenylalanine pathway (00400), with 80% of the total number of reactions in the pathway. The pyrimidine metabolism pathway (00860) is also partially conserved, with over 27% of its reactions present (Table 4.6(b)).

Results in eukaryota, as seen in Table 4.6(c), show how oxidative phosphorylation (00190), carbon fixation (00710), and glyoxylate and dicarboxylate metabolism (00630) pathways have over 60% of their total number of reactions highly conserved. The valine, leucine and isoleucine biosynthesis pathway (00290) also had a significant number of reactions highly conserved.

We ran two more experiments using the previous archaea and bacteria organism sets, but recalculating the shared pathways among all organisms. In the first one, we substituted the eukaryota organism set by all the mammals present in KEGG except *Rattus norvegicus*[5]. Table 4.7 presents the selected organism and shared pathways among them.

Results in Table 4.9(a) shows the oxidative phosphorylation pathway (00190) still ranks as the most relevant, with nearly 43% of its reactions highly conserved. In the second experiment, eukaryota were substituted by a subset of all plants[6] present in KEGG. Table 4.8 describes the set of organisms and their shared pathways. In this case, the carbon fixation in photosynthetic organisms pathway was selected as the most relevant (nearly 80% of its reactions are conserved), while oxidative phosphorylation is also highly conserved (Table 4.9(b)).

[5] Information in KEGG about this organism seems to be incomplete, unpublished results.

[6] Results for plants were based on metabolic pathways obtained from Expressed Sequence Tag (EST) data, which are of lower quality

Table 4.6: Conserved pathways in bacteria, archaea, and eukaryota

(a) Bacteria

Path	Cons. reacs.
00061	**64% (16/25)**
00550	**37.5% (6/16)**
00670	16.66% (1/6)
00190	14.28% (1/7)

(b) Archaea

Path	Cons. reacs.
00400	**80% (16/20)**
00860	27.27% (3/11)
00670	25% (1/4)
00252	22.22% (2/9)

(c) Eukaryota

Path	Cons. reacs.
00190	**66.66% (4/6)**
00710	**64.28% (9/14)**
00630	**62.5% (5/8)**
00290	50% (3/6)

Table 4.7: Conserved reactions among bacteria, archaea and mammals: organisms and shared pathways

Taxon	Organisms	Shared path.

Bacteria	*aae ana atc atu bha bme bsu cac ccr cje cmu cpa cpj cpn ctr dra ece ecj eco ecs fnu hin hpj hpy lin lla lmo mle mlo mtc mtu nma nme pae pmu rpr rso sau sav sco sme spn stm sty syn tma tte vch xcc xfa ype*	00010 00190
		00230 00240
		00251 00280
		00310 00500
		00650 00710

Archaea	*afu ape hal mac mja mma mth pab pai pfu sso sto tac tvo*	
Mammals	*bta cfa hsa mmu ptr ssc*	

Table 4.8: Conserved reactions among bacteria, archaea and plants: organisms and shared pathways

Taxon	Organisms	Shared path.

Bacteria	*aae ana atc atu bha bme bsu cac ccr cje cmu cpa cpj cpn ctr dra ece ecj eco ecs fnu hin hpj hpy lin lla lmo mle mlo mtc mtu nma nme pae pmu rpr rso sau sav sco sme spn stm sty syn tma tte vch xcc xfa ype*	00010 00061 00190 00230
		00240 00251 00252 00260
		00271 00272 00280 00290
		00300 00310 00330 00400
		00450 00500 00550 00564
		00620 00630 00640 00650
		00670 00710 00760 00860
		00900 00970

Archaea	*afu ape hal mac mja mma mth pab pai pfu sso sto tac tvo*	
Plants	*ebna ecsi egar egma egra ehan ehvu elco eles elsa emtr eosa epba epta esbi esof estu etae evvi ezma*	

Table 4.9: Conserved pathways in mammals and plants

(a) Mammals

Path	Cons. reacs.
00190	**42.85% (3/7)**
00240	6.66% (5/75)
00280	3.03% (1/33)

(b) Plants

Path	Cons. reacs.
00710	**79.16% (19/24)**
00190	**71.42% (5/7)**
00272	60% (6/10)

Table 4.10: Non-conserved reactions: strains and shared pathways

Organism	Strains	Shared path.
S. pyogenes	*spy spz spm spg sps sph spi spj spk spa spb*	00010 00030 00040 00051 00052 00061 00071 00072 00100 00190 00220 00230 00240 00251 00252 00260 00271 00272 00280 00290 00300 00310 00330 00340 00350 00360 00380 00400 00430 00450 00460 00471 00473 00480 00500 00520 00521 00523 00530 00550 00561 00562 00564 00590 00620 00624 00630 00632 00640 00650 00670 00710 00740 00760 00770 00780 00790 00860 00900 00903 00910 00920 00960 00970
E. coli	*eco ecj ece ecs ecc eci ecp*	00010 00020 00030 00040 00051 00052 00053 00061 00062 00071 00100 00120 00130 00190 00220 00230 00240 00251 00252 00260 00271 00272 00280 00290 00300 00310 00330 00340 00350 00360 00361 00380 00400 00401 00410 00430 00450 00460 00471 00473 00480 00500 00520 00521 00523 00530 00540 00550 00561 00564 00600 00603 00620 00624 00627 00630 00632 00640 00650 00660 00670 00680 00710 00720 00730 00740 00750 00760 00770 00780 00790 00860 00900 00903 00910 00920 00930 00950 00960 00970 00980 01053
S. aureus	*sau sav sam sar sas sac sab saa sao*	00010 00020 00030 00040 00051 00052 00061 00100 00120 00190 00220 00230 00240 00251 00252 00260 00271 00272 00280 00290 00300 00310 00330 00340 00350 00360 00380 00400 00410 00430 00440 00450 00460 00471 00472 00473 00480 00500 00520 00530 00550 00561 00564 00600 00602 00604 00620 00624 00630 00632 00640 00650 00660 00670 00680 00710 00720 00730 00740 00760 00770 00780 00790 00860 00900 00903 00910 00970

4.2.3 Non-conserved reactions

We also investigated how different strains of the same organism can be aligned, focusing in gap reactions which are not conserved among the strains (that is, cannot be aligned with similarity higher than 0). Since the strains are all genetic variants of a single organism, we expect to have a high-score alignment among strains, with few or no gap reactions. We performed experiments with strains of *Streptococcus pyogenes*, *Escherichia coli* and *Staphylococcus aureus* and their respective sets of shared metabolic pathways (Table 4.10).

In the following sections we describe results for these experiments as presented in tables B.1, B.2 and B.3, where gap reactions annotated to a strain S indicate that the reactions are present in some of the other strains but cannot be aligned with similarity greater than 0 in S.

Streptococcus pyogenes

Streptococcus pyogenes is a Gram-positive bacteria associated with different diseases through the release of toxins, as in scarlet fever and toxic shock syndrome. We used 11 strains from *Streptococcus pyogenes* currently stored in KEGG, and aligned 64 of their shared pathways which contained at least one reaction.

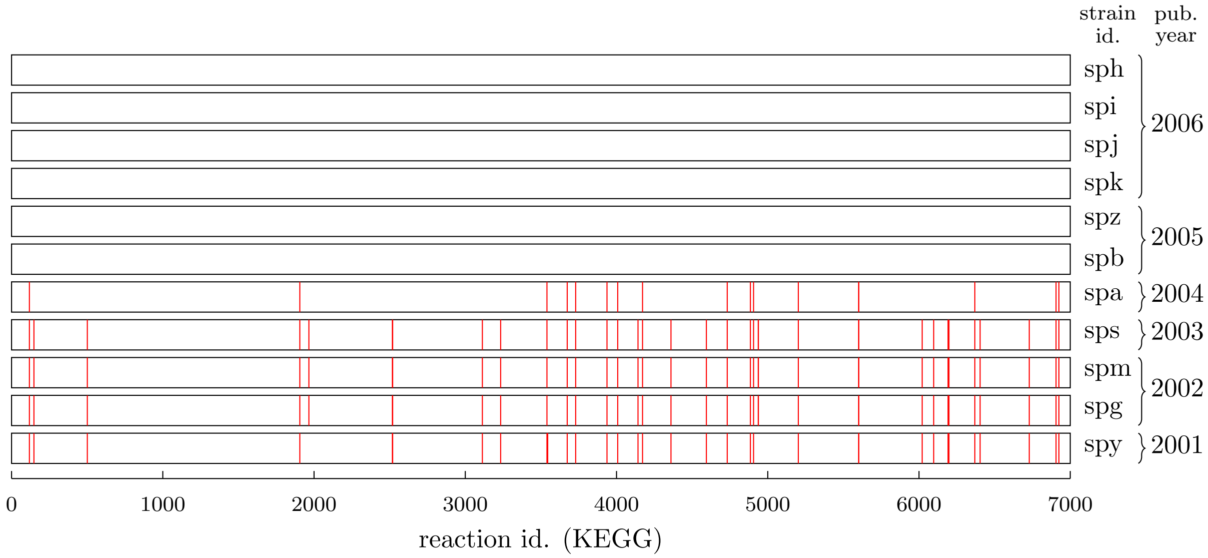


Figure 4.14: Gaps in the alignment of *Streptococcus pyogenes* strains (ordered by publication year)

Figure 4.14 represents the gaps induced in the alignment between different strains. As it can be seen, recently published strains (years 2005 and 2006) tend to have a larger number of reactions that cannot be aligned to “older” strains. In particular, the serotype M3 strains *sph*, *spi*, *spj* and *spk* [13], introduced in KEGG in 2006, contain many reactions that cannot be aligned and therefore appear as gaps in other strains. These four strains can have all their reactions aligned with no gaps among them, which indicates a high degree of similarity.

Strains *spg* and *sps* are also serotype M3, but they have some gaps in their alignment to previous M3 strains. Serotype M18 strain *spm*, associated with acute rheumatic fever outbreaks, shares the gaps with both *spg* and *sps*. Strains *spy* and *spz* are serotype M1. The more recent strain *spz* (2005) has no gaps, while *spy* (2001) contains several. Serotype M28 strain *spb* (2005) also presents no gaps. Table B.1 details the exact reactions introducing gaps in each of the strains.

We also investigated what enzymes catalyze the gap reactions (Table B.4). Most of these enzymes are still not fully characterized, as indicated by the "-" symbol in their EC identifier. Only enzymes 3.1.3.73 and 1.14.18.1 are fully characterized, and they appear in few reactions.

Escherichia coli

Escherichia coli, a bacteria present in the lower intestine of mammals, is a common bacterial model organism responsible for different kinds of infections, as well as for food-poisoning in contaminated meat. We used 7 strains stored in KEGG, and aligned 82 shared pathways containing at least one reaction.

As with *Streptococcus pyogenes*, strains of *Escherichia coli* recently incorporated in KEGG seem to have less gap reactions than older ones. Figure 4.15 shows how the two most recent strains, *eci* and *ecp*, introduce a significant number of gaps in the alignment with older strains. The number of gaps does not strictly correspond with the year of publication though: strain *ecc* was introduced after *ecj*, *ece* or *eco*, but has a larger number of gaps.

Strains representing the toxigenic *Escherichia coli* O157:H7 (*ece* and *ecs*) share their set of gap reactions. K-12 type strains *eco* (MG1655) and *ecj* (W3110), on the other hand, do not fully share their gap reaction set. These two strains were published respectively in 1997 and 2006 (the web entry in KEGG's organism list incorrectly marks 2001), with the second publication correcting some of the entries of the first one [58]. Inspection of enzymes involved in gap reactions in *Escherichia coli* shows again a large predominance of not fully-characterized enzymes, as it can be seen in Table B.5. Table B.2 presents the detailed list of gaps for each strain.

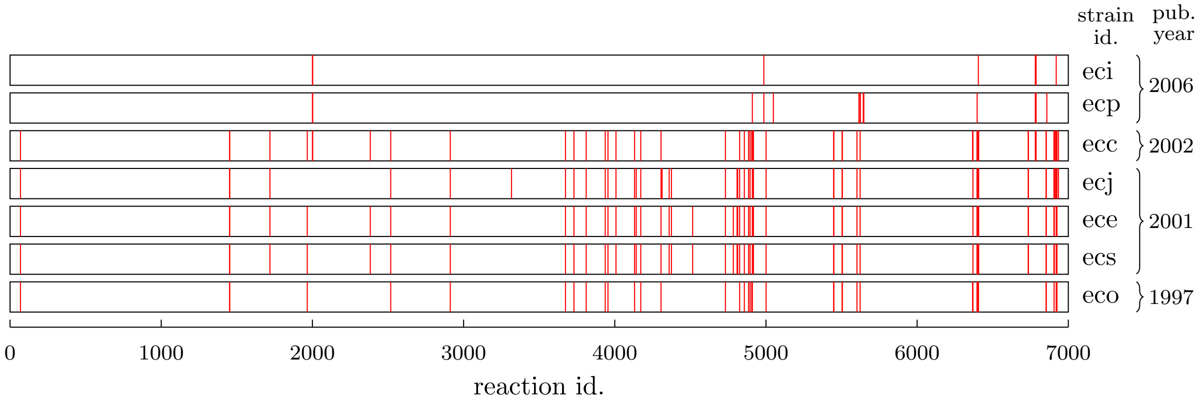


Figure 4.15: Gaps in the alignment of *Escherichia coli* strains (ordered by publication year)

Staphylocossus aureus

Staphylococcus aureus is a Gram-positive bacterium that can cause a wide range of diseases, such as pneumonia, meningitis, or toxic shock syndrome. The bacteria has become resistant to many antibiotics in the last years, and may be fatal in cases of severe infections. KEGG currently contains data on 9 different strains, and for our experiments we used 68 of their shared pathways containing at least one reaction.

Figure 4.16 presents how, as in the two previous sections, recently published strains tend to have less gap reactions. In this case, *saa* (2006) and *sab* (2005) have the least number, closely followed by *sao* (2006). Methicillin-resistant (MRSA) and methicillin-susceptible (MSSA) strains *sar* and *sas* were published together [64] and have an identical set of gap reactions. The three strains for which the primary repository is the NITE/Juntendo database (*sau*, *sav* and *sam*) have the same gap set as well. A detailed revision of the papers associated with the first two strains [80] and the last one [9] reveals that the group that published the three strains is in fact the same. Finally, the strain *sac* has a significant number of gaps despite its recent publication. Table B.3 summarizes the list of gap reactions. Most of the enzymes present in gap reactions in *Staphylococcus aureus* were again not fully-characterized (Table B.6).

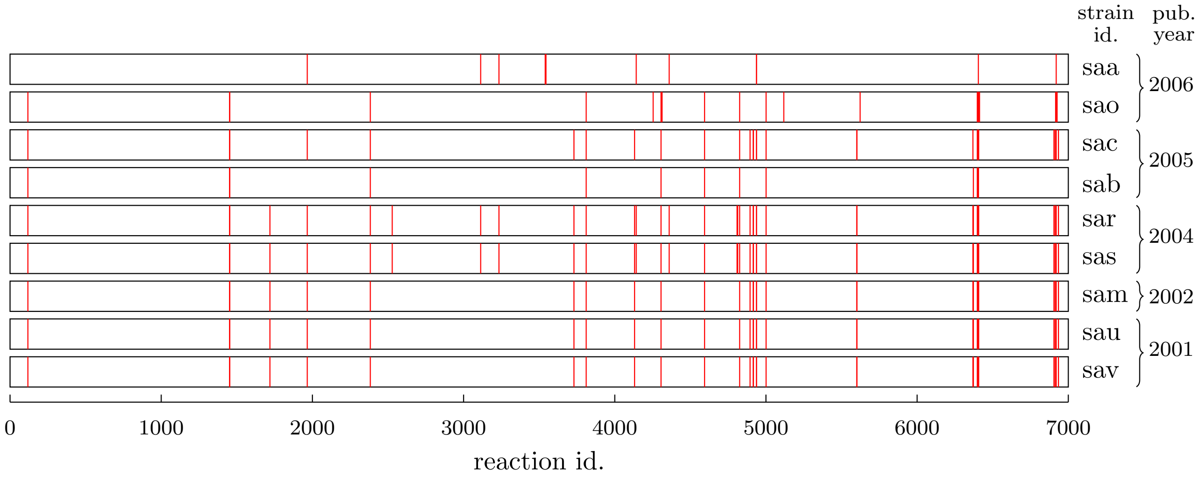


Figure 4.16: Gaps in the alignment of *Staphylococcus aureus* strains (ordered by publication year)

4.2.4 Results and discussion

Understanding which metabolic processes are fundamental for a group of organisms can be of great utility [11]. In this section we have presented an approach based on our metabolic pathway alignment algorithm to detect what pathways have a significant number of reactions conserved with high similarity. Results show how can we establish which are the most relevant metabolic pathways for taxa such as bacteria, archaea, mammals or plants.

We also investigated how an alignment among strains of three different bacteria reveals a significant number of reactions that belong exclusively to some of the strains, and therefore produce a gap in the alignment with other strains. We found out such gap reactions to be much more common in species recently introduced in KEGG. Also, most gap reactions seem to be catalyzed by enzymes which are not yet fully determined. These evidence seem to imply that these reactions are in fact misannotations.

KEGG is probably the most complete resource on metabolic information publicly available, and despite its high quality standards it is to be expected that misannotations are introduced due to the automated nature of the annotation process. Methods to detect such errors are well-known in the case of sequence data, and have been recently introduced also for metabolic data [40, 41]. The work presented in this section provides a different approach towards detecting such misannotations.

Conserved reactions in a group of organisms

By studying conserved reactions in a group of organisms, we expected to find biological processes which are relevant for some of the organisms. The fatty acid biosynthesis pathway was significantly conserved in bacteria when compared to the other two domains (Table 4.6(a)). It is known that enzymes of fatty acid biosynthesis represent excellent targets for drugs against bacteria [16, 102], therefore the relevance of this pathway and its high number of conserved reactions in our experiments.

Our method was also able to detect the relevance of the peptidoglycan biosynthesis pathway (Table 4.6(a)). These polymers enable bacteria to withstand high osmotic pressures. Highly conserved in bacteria, peptidoglycans have no parallels in eukaryota, and disruption of the peptidoglycan pathway can be lethal for bacteria [134].

We also found out the phenylalanine pathway was significantly conserved in Archaea (Table 4.6(b)). Phenylalanine is an essential alpha amino acid that cannot be synthesized by animals, which have to obtain it from their diet. It is produced from prephenate, an intermediate on the shikimate pathway. Interestingly, although this pathway is present in archaea, bacteria, fungi, and plants, there are two kinds of shikimate kinases: archaeal and non-archaeal. Archaeal shikimate kinases are, by sequence similarity, distantly related to homoserine kinases (GHMP kinase domain superfamily) [30], while all non-archaeal shikimate kinases (the typical form) belong to the (structurally unrelated) NMP kinase domain superfamily [77]. We found 6 entries for enzymes related to shikimate in KEGG: 1.1.1.25, 1.1.1.282, 1.14.13.36, 2.3.1.133, 2.5.1.19 and 2.7.1.71. Except for 1.14.13.36 and 2.3.1.133, which are not annotated to any pathway, all enzymes belong to the phenylalanine pathway.

Eukaryotes were more complex to analyze. The first experiment (see Table 4.6(c)) shows a series of conserved pathways which cannot be directly linked to fundamental metabolic processes in all eukarya: carbon fixation (00710), for instance, is vital only for plants. It is not clear either how the glyoxylate and dicarboxylate metabolism (00630) are of relevance to the selected eukaryotes. After detailed inspection of results we found out reactions conserved in these pathways are actually annotated to more than one metabolic pathway. Additionally, the definition of a metabolic pathway in KEGG does not necessarily correspond to the traditional understanding expressed in the literature, and KEGG pathways are known to overlap. Several reactions in the Calvin cycle are also present in the pentose phosphate pathway, which might explain the high value for conserved reactions in the carbon fixation pathway among all eukaryotes.

The most conserved pathway in eukaryota was oxidative phosphorylation, which is the final pathway of cellular respiration after glycolysis and the cytric acid cycle. Its basic function is to transfer electrons from NADH or $FADH_2$ to molecular oxygen through protein complexes located in the mitochondria. Arguably, the fact that mitochondria are not found in bacteria or archaea [60] could explain that metabolic pathways related to activity in the mitochondria would be relevant to eukaryotic organisms only. Tables 4.9(a) and 4.9(b) show the relevance of this pathway both in mammals and plants, which is consistent with this hypothesis. Additionally, Table 4.9(b) also describes how carbon fixation in photosynthetic organisms is effectively conserved for plants, as would be expected.

It should be noticed that genome sequences for nearly all bacteria and archaea present in KEGG are fully determined, which only happens for a few of the eukarya. Given that metabolic information in KEGG is obtained by analyzing sequence data from GenBank, this might explain the poorer quality of results for eukaryota.

Non-conserved reactions in a group of strains

Non-conserved reactions (gaps) are those reactions which are contained in one strain but cannot be aligned with similarity greater than 0 to any other reaction in different strains. This means that the set of enzymes and compounds contained in such reactions do not appear in any other reaction. Since strains are variants of one single organism, gap reactions are worth studying because they represent a set of enzymes and compounds unique to a certain strain. We argue that such reactions are most probably an annotation mistake.

Sections 4.2.3, 4.2.3 and 4.2.3 describe how strains recently included in KEGG contain a significant number of reactions that appear as gaps in older strains. Specifically, strains introduced in years 2006 and 2005 in *Streptococcus pyogenes* (*sph*, *spi*, *spj*, *spk*, *spz*, *spb*; Table B.1), *Escherichia coli* (*eci*, *ecp*; Table B.2), and *Staphylococcus aureus* (*sca*, *sco*, *scc*, *scb*; Table B.3) contained the smallest number of gaps and the largest number of reactions that are gaps in older strains. This suggests that such strains contain misannotated reactions due to the automated nature of the annotation process. Although our experiments are limited to alignment among strains of an organism, it is reasonable to expect similar results for any organism in general. Information about species recently included in KEGG should therefore be considered as tentative, as it has been confirmed by other authors [40, 41].

Additionally, most gap reactions are catalyzed by enzymes not fully determined (Tables B.4, B.5 and B.6). Analysis of the references associated with each of the strains in KEGG did not explain either why these enzymes should be annotated to those strains, which further implies that these reactions might indeed be misannotations. Even for reactions with fully determined enzymes, such as 3.1.3.73 in *Streptococcus pyogenes*, 1.18.1.4 in *Escherichia coli*, and 3.1.3.73 in *Staphylococcus aureus*, we found out that they have no genes annotated to them. EC numbers are not always supported by experimental validation, and they can introduce errors in metabolic pathway repositories as those presented in this work. Further details on errors associated with EC numbers can be found in [53].

4.3 Model organisms

A different set of questions of biological relevance can also be answered using our method for metabolic pathway alignment. Experimentation of novel disease treatments on humans is costly, can imply risks for the subject, and might raise ethical issues. The alternative use of model organisms is a common practice in biology to avoid these problems. Identifying appropriate model organisms for a specific experiment can be of great use for the biologist. Our method for metabolic similarity can provide this functionality by simply choosing the candidate model organisms and the set of pathways involved in the experiments [25].

As an example of such use, we performed an experiment to find a suitable model organism in a hypothetical condition related to pyruvate kinase deficiency in humans, which is usually related to haemolytic anaemias [12, 149][7].

Given a set of candidate organisms (*Escherichia coli*, *Caenorhabditis elegans*, *Rattus norvegicus*, *Drosophila melanogaster*, and *Homo sapiens*), we started by aligning all their common pathways to reconstruct their phylogenetic relations. The result, as seen in Figure 4.17 (left), is a phylogenetic tree with one cluster for animals, further divided into vertebrate (*Rattus norvegicus*, *Homo sapiens*) and non-vertebrate (*Caenorhabditis elegans* and *Drosophila melanogaster*), and a second cluster with the bacterium *Escherichia coli.* This tree corresponds with the NCBI taxonomy for this set of organisms. We then considered those KEGG pathways that contain pyruvate kinase and are common to the candidate organisms, namely: glycolysis, pyruvate metabolism, and purine metabolism. Figure 4.17 (right) presents the produced tree in this case, which places *Drosophila melanogaster* closer to *Homo sapiens* than *Rattus norvegicus.* Although we could not find conclusive confirmation in the literature, this result suggests that organisms with a most similar global metabolism might not necessarily be the best option as model organisms for specific biological experiments.

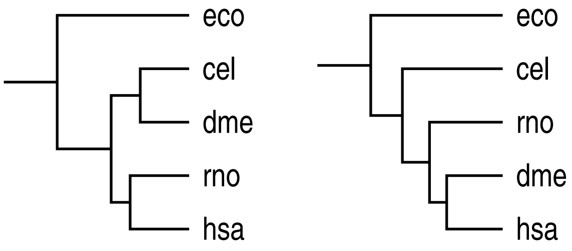


Figure 4.17: Tree reconstructed for *Escherichia coli*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Rattus norvegicus*, and *Homo sapiens* when using all their common pathways (left) and common pathways that contain pyruvate kinase (right)

[7] Interestingly, deficiency in this enzyme has also been linked to protection against malaria in mice [99]

Chapter 5

Implementation

> *The competent programmer is fully aware of the limited size of his own skull. He therefore approaches his tasks with full humility, and avoids clever tricks like the plague*
>
> E. W. Dijkstra [34]

5.1 Phylogenetic web server

We have implemented the method described in Chapter 3 as a web service that can be used to reconstruct the phylogeny of a set of organisms from the observed similarity of their common metabolic pathways. Although some tools have been previously described for querying pathways [86, 116, 124], comparing metabolic connectivity [118], and reconstructing phylogenetic trees from genome sequence [75], we are not aware of any publicly available web server that allows the user to reconstruct phylogenetic trees from similarity of metabolic processes.

Information about organisms and pathways in the server is periodically retrieved from KEGG through a Perl script, and then updated into a local SQLite database that contains information on all organisms, pathways, reactions, enzymes, and chemical compounds. In order to save computation time in scoring the similarity of pathways, we precalculate and store distances among all enzymes, compounds, and reactions to provide results in linear time. Figure 5.1 describes the overall implementation process of the web server. Currently, our database stores information for 13 organisms, 25 metabolic pathways, and over 1,000 enzymes, compounds, and reactions, as well as precalculated distances for 180,000 enzyme pairs and over 1,000,000 reaction pairs. Figure 5.2 presents the database schema used. Our web server can be accessed at `http://www.jaist.ac.jp/~clemente/cgi-bin/phylo.pl`.

The user can select among a set of organisms and pathways, three different enzyme similarity measures (hierarchical, information content and gene ontology), two clustering methods (UPGMA [98] and neighbor-joining [121]), the choice of full organism names or abbreviations in the output, and the value of the α parameter (relative weight of compounds and enzymes in assessing the similarity of enzymatic reactions). The interface is designed in a way such that, when selecting an organism, all the pathways not present in that organism are disabled, and vice versa. This is to ensure that the alignment of pathways is done in a meaningful way, that is, we are not aligning against an empty set

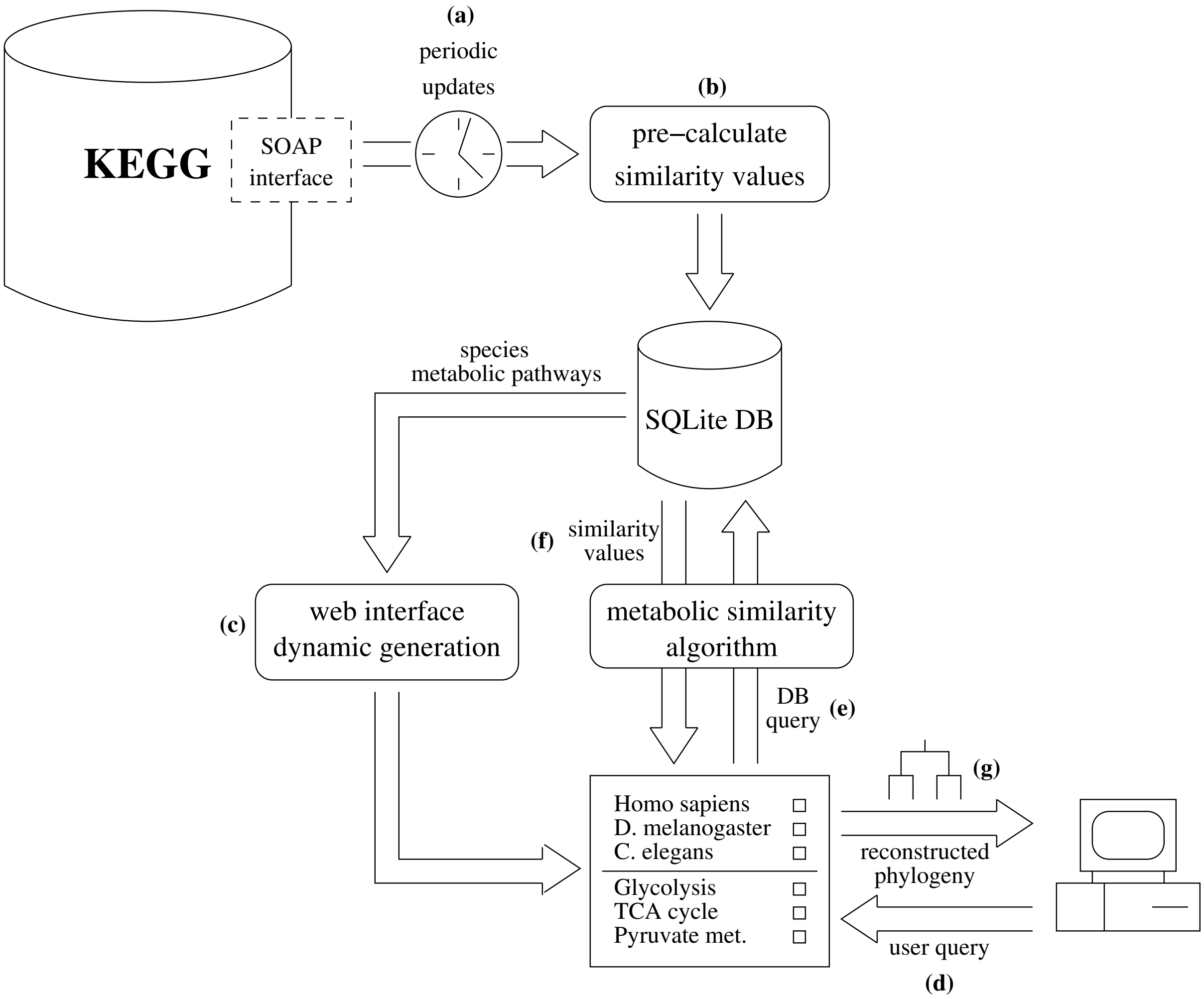


Figure 5.1: Web server implementation. Information from KEGG is retrieved periodically through its SOAP interface **(a)** to update the local SQLite database with new metabolic data, precalculating distances among enzymes, compounds and reactions **(b)**. From the list of metabolic pathways and species contained in the database, a web interface is dynamically generated **(c)**. When a user makes a query by selecting a set of organisms, pathways and control parameters **(d)**, a set of scripts implementing our metabolic similarity algorithm make the relevant queries to the database **(e)** retrieving the necessary similarity values **(f)** and returning the calculated phylogenetic tree obtained from the metabolic similarity of the selected species **(g)**

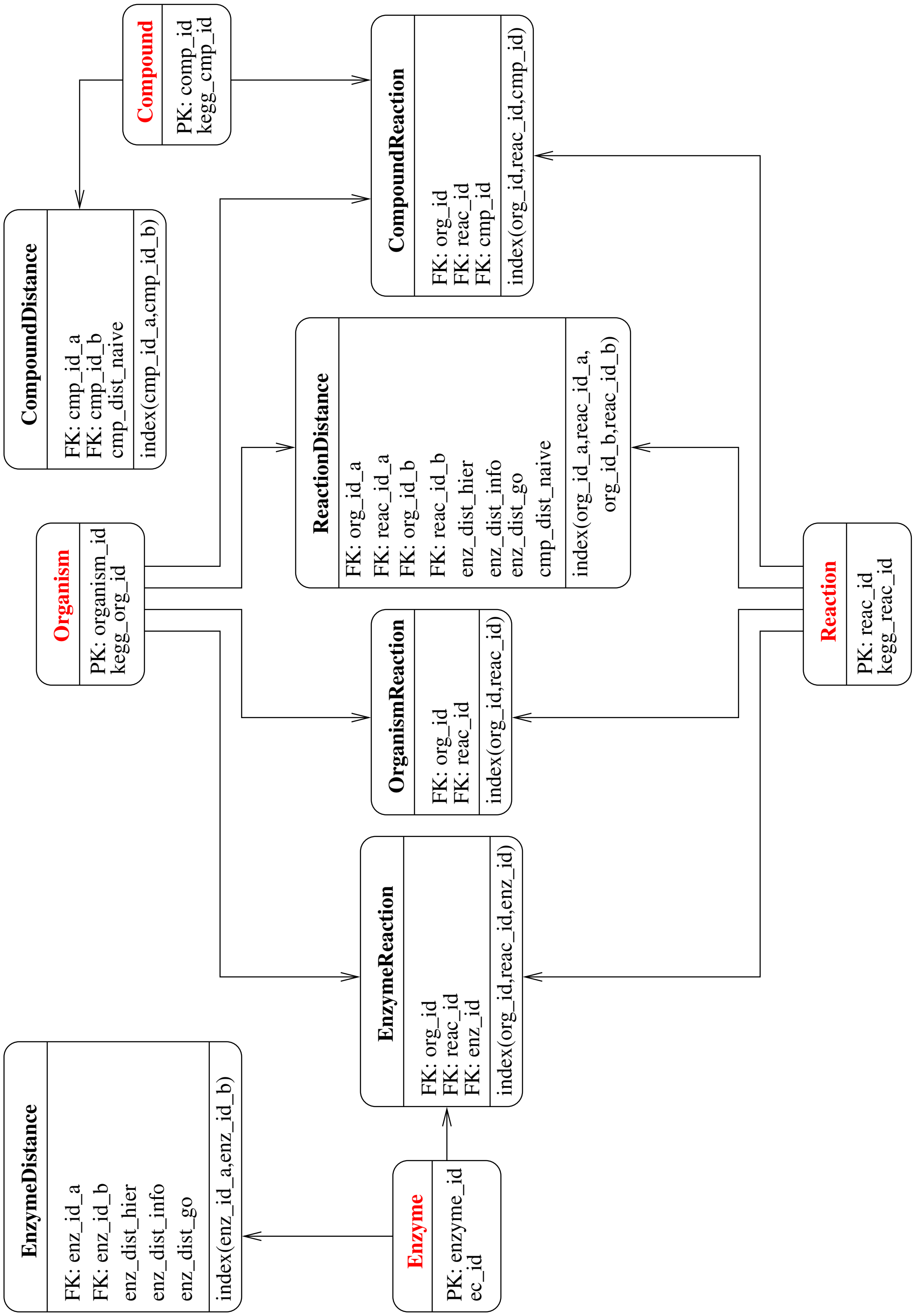
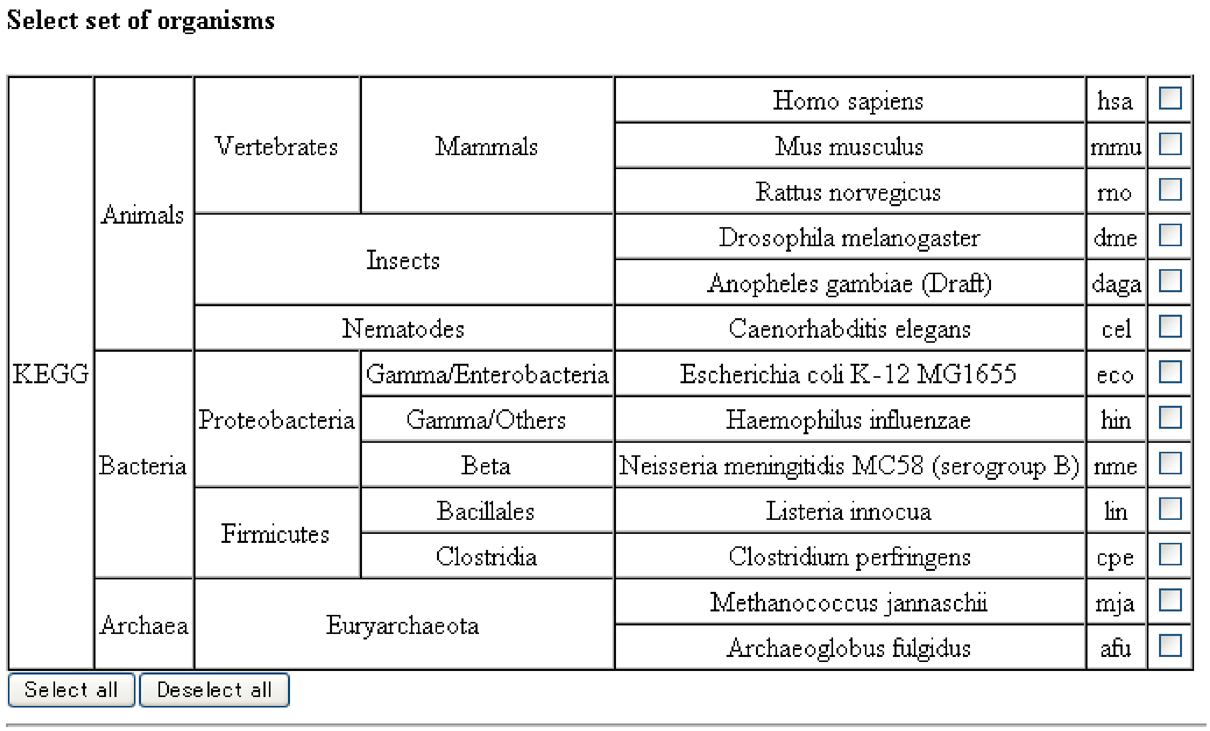


Figure 5.2: SQLite database schema for the web server. Data in the "primary" tables (Organism, Reaction, Enzyme and Compound) is directly retrieved from KEGG, with each entry having a unique identifier. Tables CompoundReaction, EnzymeReaction and OrganismReaction relate compounds, enzymes and organisms to reactions. Each entry is defined by exporting the appropriate primary key, and indices are defined to speedup lookups. Distance tables (EnzymeDistance, CompoundDistance and ReactionDistance) contain pairs of objects for which the precalculated distance is greater than zero. Indices were also created in these tables to improve efficiency

of reactions. Two screenshots of our web server can be seen in figures 5.3, 5.4, and 5.5, showing the list of currently available organisms, metabolic pathways, and parameters.



Select set of organisms

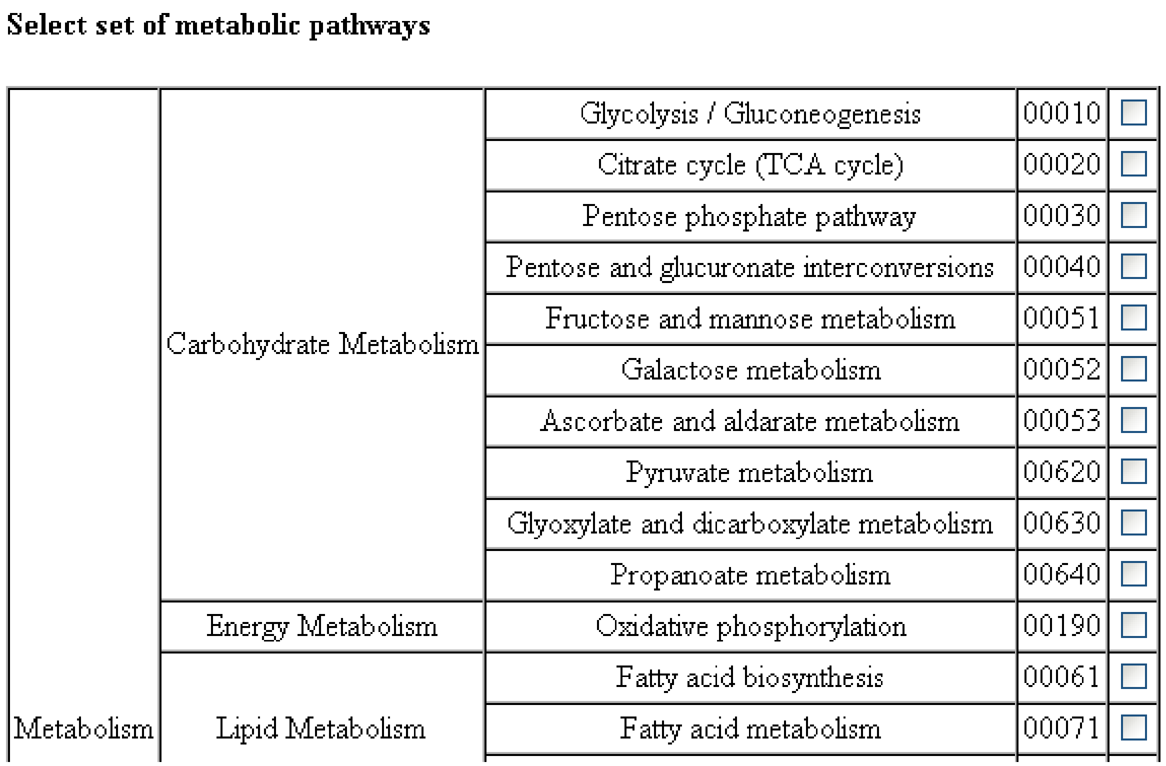
KEGG	Animals	Vertebrates	Mammals	Homo sapiens	hsa	□
				Mus musculus	mmu	□
				Rattus norvegicus	rno	□
		Insects		Drosophila melanogaster	dme	□
				Anopheles gambiae (Draft)	daga	□
		Nematodes		Caenorhabditis elegans	cel	□
	Bacteria	Proteobacteria	Gamma/Enterobacteria	Escherichia coli K-12 MG1655	eco	□
			Gamma/Others	Haemophilus influenzae	hin	□
			Beta	Neisseria meningitidis MC58 (serogroup B)	nme	□
		Firmicutes	Bacillales	Listeria innocua	lin	□
			Clostridia	Clostridium perfringens	cpe	□
	Archaea	Euryarchaeota		Methanococcus jannaschii	mja	□
				Archaeoglobus fulgidus	afu	□

Figure 5.3: Web server screenshot: organism selection

Once a query is submitted, we calculate the phylogenetic tree for the selected organisms, pathways, and parameters, as previously described. Results are presented both in Newick and graphic format, including branch lengths. A Postscript file of the produced tree can be downloaded by clicking on the displayed image (see Figure 5.6).

The server includes several optimizations to reduce the time required to calculate a phylogenetic tree. As explained above, similarity values are precalculated offline and then stored in the database. Whenever a new organism or metabolic pathway is retrieved from KEGG all similarity values not already calculated are computed and updated in the database if the value is greater than 0[1]. Queries to the database are also optimized for speed, at the cost of retrieving data which might not be used. For instance, if we need the similarity value of objects A and B (where A and B might be enzymes, compounds or reactions), instead of obtaining only the value $sim(A, B)$ we retrieve all similarity values for A. We found out that retrieving more data and post-processing it afterwards is faster, since many of the similarity values for A would be required at some point later in the calculations anyways. These optimizations allow our web server to answer user queries in time almost linear in the number of enzymes, compounds and reactions.

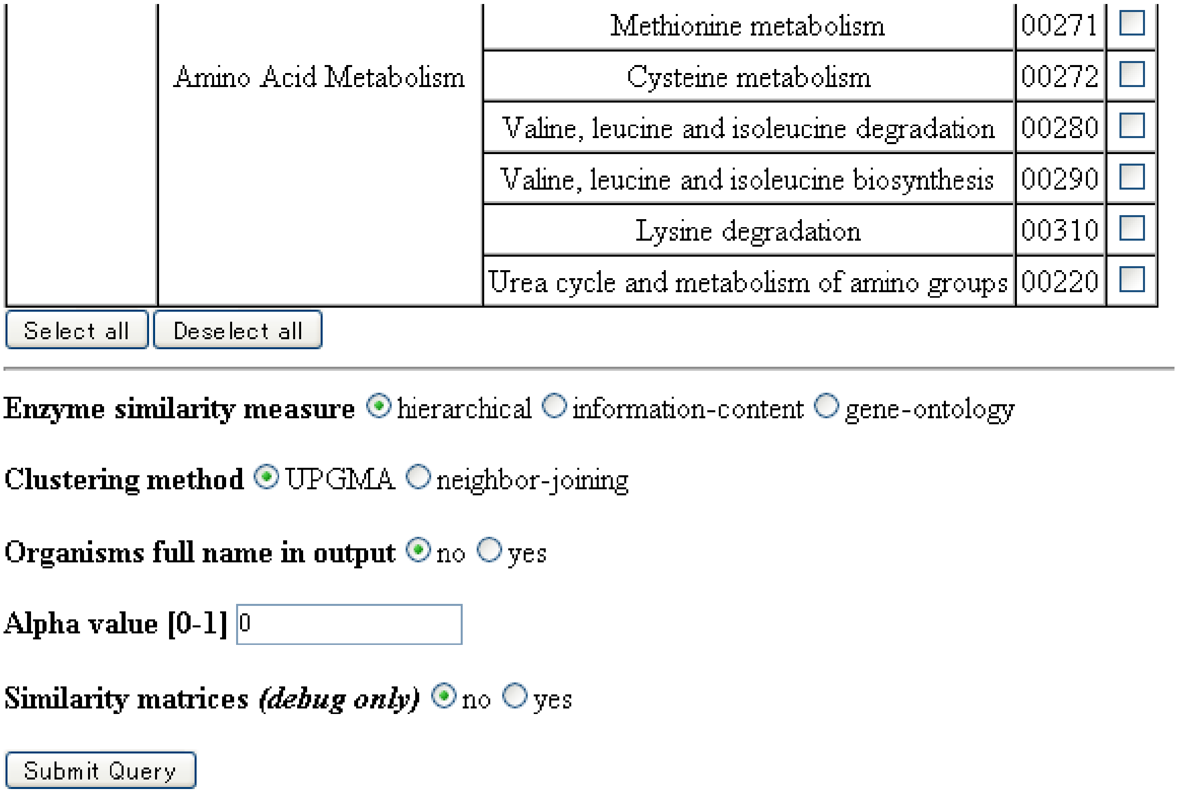
[1] In order to save disk space, since most similarity values are in fact 0.



Select set of metabolic pathways

Metabolism	Carbohydrate Metabolism	Glycolysis / Gluconeogenesis	00010	□
		Citrate cycle (TCA cycle)	00020	□
		Pentose phosphate pathway	00030	□
		Pentose and glucuronate interconversions	00040	□
		Fructose and mannose metabolism	00051	□
		Galactose metabolism	00052	□
		Ascorbate and aldarate metabolism	00053	□
		Pyruvate metabolism	00620	□
		Glyoxylate and dicarboxylate metabolism	00630	□
		Propanoate metabolism	00640	□
	Energy Metabolism	Oxidative phosphorylation	00190	□
	Lipid Metabolism	Fatty acid biosynthesis	00061	□
		Fatty acid metabolism	00071	□

Figure 5.4: Web server screenshot: metabolic pathway selection



	Amino Acid Metabolism	Methionine metabolism	00271	□
		Cysteine metabolism	00272	□
		Valine, leucine and isoleucine degradation	00280	□
		Valine, leucine and isoleucine biosynthesis	00290	□
		Lysine degradation	00310	□
		Urea cycle and metabolism of amino groups	00220	□

Enzyme similarity measure ● hierarchical ○ information-content ○ gene-ontology

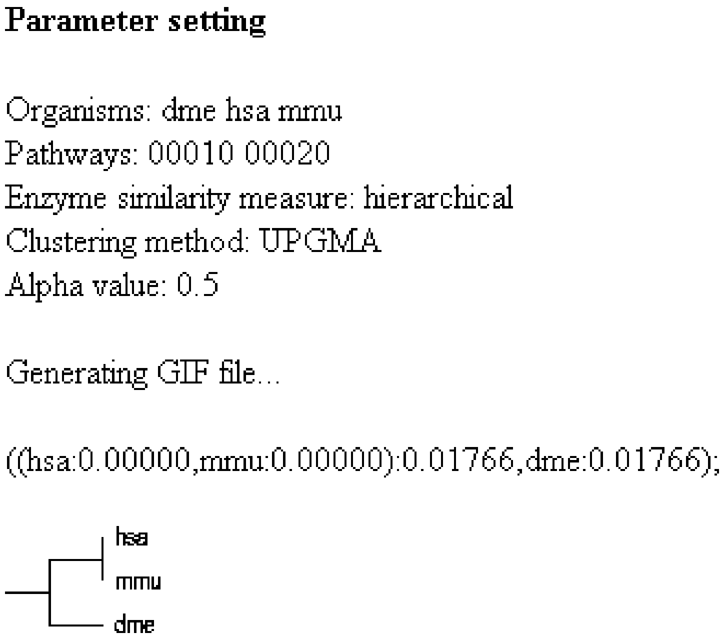
Clustering method ● UPGMA ○ neighbor-joining

Organisms full name in output ● no ○ yes

Alpha value [0-1] 0

Similarity matrices *(debug only)* ● no ○ yes

Figure 5.5: Web server screenshot: parameter selection



Parameter setting

Organisms: dme hsa mmu
Pathways: 00010 00020
Enzyme similarity measure: hierarchical
Clustering method: UPGMA
Alpha value: 0.5

Generating GIF file...

((hsa:0.00000,mmu:0.00000):0.01766,dme:0.01766);

Figure 5.6: Web server screenshot: produced phylogenetic tree

5.2 Standalone distribution

We have also developed a series of Perl modules to ease the manipulation of metabolic data and implementing the algorithms presented in this thesis. A total of 5 modules represent each of the objects to be found in metabolism: `Compound, Enzyme, Reaction, Pathway`, and `Organism`. Three auxiliary modules are used for input/output operations (`PathwayIO`), user queries (`Query`), and calculation of similarity values (`Similarity`). These modules make use of the `Graph` module and the `BioPerl` collection, which should be previously installed by the user[2]. All modules are currently available from the author upon request, and will be made publicly available in the near future at CPAN (http://www.cpan.org).

The `Compound` module represents metabolites used in metabolism. The module provides operations to create, assign name, convert into a string and compare with other metabolites (determine whether metabolites are exactly the same or not).

The `Enzyme` module represents catalytic enzymes acting upon reactions. Operations to create, assign name, convert into a string and compare with other enzymes are similar to those in the `Compound` module. Since enzymes are represented by their EC identifier, we also implemented a function to determine when two enzymes are *similar*: either all their four digits are the same, or they share their most significant digits and less significant digits are replaced by dashes (“-”). For instance, enzymes 3.1.2.4 and $3.1.-.-$ are similar under this definition, but 3.1.2.4 and $4.1.-.-$ are not. Since some enzymes in KEGG are not fully determined, this function is useful to calculate similarity among partially determined enzymes. This function should not be confused with the comparison of enzymes previously described. In fact it provides a more general functionality: instead of strict equality, partial equality of enzymes is checked by using wildcards (dashes).

[2] We recommend installing `BioPerl` 1.4. Later versions seem to have several problems when using the `Ontology` module which we require for the calculations in our `Similarity` module.

The Reaction module implements metabolic reactions. The usual set of basic operations are included in this module: create, assign a name, convert into a string and compare with other reactions. Reactions are composed of compounds and enzymes, and in order for two reactions to be equal they should contain exactly the same compound and enzyme sets. Operations to declare and/or retrieve the compounds and enzymes of the reaction are also implemented, with additional functions to determine if a compound is either a substrate or a product of the reaction. A string-to-reaction conversion is also implemented to facilitate loading reaction data from standard KEGG text files.

Metabolic pathways are implemented in the Pathway module. Functions to create, name, convert into a string, convert from a string and compare with other pathways were implemented as with previously described modules. Operations to declare and/or retrieve the compounds, enzymes and reactions of a pathway are also included.

The metabolism of whole organisms can be specified by using the module Organism, which includes operations to create, name and compare organisms, as well as a function to declare/retrieve the pathways that compose it.

The PathwayIO module manages input/output processes, reading pathway data from a file or directly from KEGG's API, and printing metabolic information. The data format for files follows the regular expression:

```
(path:\w{3,4}:\d{5} (R\d{5} (C\d{5})+ (\d.\d.\d.\d)*)+)+
```

where `path:\w{3,4}:\d{5}` is the name of the organism and the KEGG pathway identifier, `R\d{5}` is one of the reactions included in the pathway, `(C\d{5})+` is the set of compounds (substrates and products) that take part in a reaction, and `(\d.\d.\d.\d)*` represents the enzymes catalyzing the reaction. For instance:

```
path:hsa:00010 R00947 C00103 C00001 C00267 C00009 3.1.3.10
```

corresponds to reaction R00947 (D-Glucose-1-phosphate phosphohydrolase) in the *Homo sapiens* glycolysis pathway (hsa:00010), which takes as substrates D-Glucose 1-phosphate (C00103) and water (C00001) to produce alpha-D-Glucose (C00267) and Orthophosphate (C00009), catalyzed by the enzyme glucose-1-phosphatase (3.1.3.10).

The Similarity module implements all similarity operations between metabolic objects (compounds, enzymes, reactions, pathways, and organisms). When creating a Similarity object, we initialize it with a list of references for the objects for which we would like to have their similarities computed, and with the parameters to calculate the similarity values as described in Chapter 3: enzyme similarity measure (hierarchical, information content or gene ontology) and α value (relative weight of enzymes and compounds in reaction similarity). Once similarity values are calculated, they can be accessed through a similarity matrix stored with the object. In order to save computation time when using the information content or gene ontology similarity measures, we provide a precomputed enzyme similarity matrix that is automatically loaded by the Similarity object if found present in the working directory. Figures 5.7 to 5.10 present code using this module to calculate similarity among enzymes, reactions, pathways and organisms. Figures 5.11 to 5.14 present the respective outputs for each of this programs.

Finally, the Query module is initialized with a set of organisms, a set of pathways, an alpha value, an enzyme similarity method, and a clustering method (UPGMA or neighbor-joining), and then proceeds to reconstruct a phylogenetic tree in the same way the web server does. For each pathway, we calculate the similarity matrix containing all organisms which are annotated to it. We then average all the similarity matrices and

```
my @enz_list = ("1.1.1.1","1.1.2.3");
print "...enzyme␣similarity␣with␣hierarchical␣measure...\n";
my $hier_sim_obj = new Metabolism::Similarity(
  ’-method’ => "hierarchical",’-objects’ => \@enz_list);
my $aa = $hier_sim_obj->similarity_matrix;
my $bb = $hier_sim_obj->object_index;
for ( my $i = 0 ; $i < scalar(@$bb) - 1 ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$bb) ; $j++ ) {
        print "Similarity␣";
        print $bb->[$i] . "␣,␣" .
              $bb->[$j] . "=␣$$aa[$i][$j]\n";
    }
}
print "...enzyme␣similarity␣with␣information␣content...\n";
my $info_sim_obj = new Metabolism::Similarity(
  ’-method’ => "information",’-objects’ => \@enz_list);
$aa = $info_sim_obj->similarity_matrix;
$bb = $info_sim_obj->object_index;
for ( my $i = 0 ; $i < scalar(@$bb) - 1; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$bb) ; $j++ ) {
        print "Similarity␣";
        print $bb->[$i] . "␣,␣" .
              $bb->[$j] . "=␣$$aa[$i][$j]\n";
    }
}
print "...enzyme␣similarity␣with␣GO...\n";
my $go_sim_obj = new Metabolism::Similarity(
  ’-method’ => "go",’-objects’ => \@enz_list);
$aa = $go_sim_obj->similarity_matrix;
$bb = $go_sim_obj->object_index;
for ( my $i = 0 ; $i < scalar(@$bb) - 1; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$bb) ; $j++ ) {
        print "Similarity␣";
        print $bb->[$i] . "␣,␣" .
              $bb->[$j] . "=␣$$aa[$i][$j]\n";
    }
}
```

Figure 5.7: `Similarity` module sample session for enzyme similarity: code

```
my $reac_c = new Metabolism::Reaction(
  ’-name’ => ’R00003’,’-subst’ => [’sC00001’,’sC00002’],
  ’-prod’ => [’C00008’,’C00004’],’-enz’ => [’2.1.1.1’]);
my $reac_d = new Metabolism::Reaction(
  ’-name’ => ’R00004’,’-subst’ => [’sC00001’,’sC00009’],
  ’-prod’ => [’C00008’,’C00004’],’-enz’ => [’1.1.1.1’]);
my @reac_list = ( $reac_c , $reac_d);
print "...reaction␣similarity␣with␣hierarchical␣measure...\n"
   ;
my $reac_sim_obj = new Metabolism::Similarity(
  ’-method’ => "hierarchical",’-alpha’ => 0.5,’-objects’ => \
     @reac_list);
my $cc = $reac_sim_obj->similarity_matrix;
my $dd = $reac_sim_obj->object_index;
for ( my $i = 0 ; $i < scalar(@$dd) - 1 ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$dd) ; $j++ ) {
        print $dd->[$i]->name . "␣-␣" .
              $dd->[$j]->name . "=␣$$cc[$i][$j]\n";
    }
}
print "...reaction␣similarity␣with␣information␣content...\n";
$reac_sim_obj = new Metabolism::Similarity(
  ’-method’ => "information",’-alpha’ => 0.5,’-objects’ => \
     @reac_list);
$cc = $reac_sim_obj->similarity_matrix;
$dd = $reac_sim_obj->object_index;
for ( my $i = 0 ; $i < scalar(@$dd) - 1 ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$dd) ; $j++ ) {
        print $dd->[$i]->name . "␣-␣" .
              $dd->[$j]->name . "=␣$$cc[$i][$j]\n";
    }
}
print "...reaction␣similarity␣with␣GO...\n";
$reac_sim_obj = new Metabolism::Similarity(
  ’-method’ => "go",’-alpha’ => 0.5,’-objects’ => \@reac_list
     );
$cc = $reac_sim_obj->similarity_matrix;
$dd = $reac_sim_obj->object_index;
for ( my $i = 0 ; $i < scalar(@$dd) - 1; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$dd) ; $j++ ) {
        print $dd->[$i]->name . "␣-␣" .
              $dd->[$j]->name . "=␣$$cc[$i][$j]\n";
    }
}
```

Figure 5.8: `Similarity` module sample session for reaction similarity: code

```
my @path_list;
$path_a = new Metabolism::Pathway(’TCA’,[$reac_c]);
$path_b = new Metabolism::Pathway(’Photosynthesis’,[$reac_d])
   ;
push(@path_list,$path_a);
push(@path_list,$path_b);
print "...pathway␣similarity␣with␣hierarchical␣measure...\n";
my $sim_obj3 = new Metabolism::Similarity(
  ’-method’=>"hierarchical",’-alpha’=>0.5,’-objects’=>\
     @path_list);
my $ee = $sim_obj3->similarity_matrix;
my $ff = $sim_obj3->object_index;
for ( my $i = 0 ; $i < scalar(@$ff) ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$ff) ; $j++ ) {
        print $$ff[$i]->name . "␣-␣" .
              $$ff[$j]->name . "=␣$$ee[$i][$j]\n";
    }
}
print "...pathway␣similarity␣with␣information␣content...\n";
$sim_obj3 = new Metabolism::Similarity(
  ’-method’=>"information",’-alpha’=>0.5,’-objects’=>\
     @path_list);
$ee = $sim_obj3->similarity_matrix;
$ff = $sim_obj3->object_index;
for ( my $i = 0 ; $i < scalar(@$ff) ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$ff) ; $j++ ) {
        print $$ff[$i]->name . "␣-␣" .
              $$ff[$j]->name . "=␣$$ee[$i][$j]\n";
    }
}
print "...pathway␣similarity␣with␣GO...\n";
$sim_obj3 = new Metabolism::Similarity(
  ’-method’=>"go",’-alpha’=>0.5,’-objects’=>\@path_list);
$ee = $sim_obj3->similarity_matrix;
$ff = $sim_obj3->object_index;
for ( my $i = 0 ; $i < scalar(@$ff) ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$ff) ; $j++ ) {
        print $$ff[$i]->name . "␣-␣" .
              $$ff[$j]->name . "=␣$$ee[$i][$j]\n";
    }
}
```

Figure 5.9: `Similarity` module sample session for pathway similarity: code (reac_c and reac_d are the same as in Figure 5.8)

```
my @path_list_g;
push(@path_list_g,$path_a);
push(@path_list_g,$path_b);
my $org_a = new Metabolism::Organism(
  '-name'=>"hsa",'-path'=>\@path_list_g);
my @path_list_b;
my $reac_z = new Metabolism::Reaction(
  '-name'=>'R00003','-subst'=>['sC00001','sC00002'],
  '-prod'=>['C00008','C00004'],'-enz'=>['2.1.5.10']);
my $path_z = new Metabolism::Pathway('TCA',[$reac_z]);
push(@path_list_b,$path_z);
push(@path_list_b,$path_b);
my $org_b = new Metabolism::Organism(
  '-name'=>"mmu",'-path'=>\@path_list_b);
my $reac_a = new Metabolism::Reaction(
  '-name' => 'R00003','-subst' => ['sC00001','sC00002'],
  '-prod' => ['C00008','C00004'],'-enz' => ['2.2.1.1']);
my $reac_w = new Metabolism::Reaction(
  '-name'=>'R00004','-subst' => ['sC00001','sC00002'],
  '-prod'=>['C00008','C00004'],'-enz' => ['1.1.1.10']);
my $path_w = new Metabolism::Pathway('Photosynthesys',[
   $reac_w]);
my $path_t = new Metabolism::Pathway('TCA',[$reac_a]);
my @path_list_c;
push(@path_list_c,$path_t);
push(@path_list_c,$path_w);
my $org_c = new Metabolism::Organism(
  '-name'=>"dme",'-path'=>\@path_list_c);
my @org_list;
push(@org_list,$org_a);
push(@org_list,$org_b);
push(@org_list,$org_c);
print "...organism␣similarity␣with␣hierarchical␣measure...\n"
   ;
my $sim_obj4 = new Metabolism::Similarity('-method'=>"
   hierarchical",'-alpha' => 0.5,'-objects'=>\@org_list);
my $ee = $sim_obj4->similarity_matrix;
my $ff = $sim_obj4->object_index;
for ( my $i = 0 ; $i < scalar(@$ff) ; $i++ ) {
    for ( my $j = 0 ; $j < scalar(@$ff) ; $j++ ) {
        print $$ff[$i]->name . "␣-␣" . $$ff[$j]->name . "=␣
           $$ee[$i][$j]\n";
    }
}
```

Figure 5.10: `Similarity` module sample session for organism similarity: code (path_a and path_b are the same as in Figure 5.9)

```
...enzyme similarity with hierarchical measure...
Similarity 1.1.1.1 , 1.1.1.1= 1
Similarity 1.1.1.1 , 1.1.2.3= 0.5
...enzyme similarity with information content...
Similarity 1.1.1.1 , 1.1.1.1= 1
Similarity 1.1.1.1 , 1.1.2.3= 0.318356190414535
...enzyme similarity with GO...
Similarity 1.1.1.1 , 1.1.1.1= 1
Similarity 1.1.1.1 , 1.1.2.3= 0.789473684210526
```

Figure 5.11: `Similarity` module sample session for enzyme similarity: output to Figure 5.7

```
...reaction similarity with hierarchical measure...
R00003 - R00003= 1
R00003 - R00004= 0.3
...reaction similarity with information content...
R00003 - R00003= 1
R00003 - R00004= 0.33737670243871
...reaction similarity with GO...
R00003 - R00003= 1
R00003 - R00004= 0.563157894736842
```

Figure 5.12: `Similarity` module sample session for reaction similarity: output to Figure 5.8

```
...pathway similarity with hierarchical measure...
TCA - TCA= 1
TCA - Photosynthesis= 0.3
Photosynthesis - TCA= 0.3
Photosynthesis - Photosynthesis= 1
...pathway similarity with information content...
TCA - TCA= 1
TCA - Photosynthesis= 0.356760901089303
Photosynthesis - TCA= 0.356760901089303
Photosynthesis - Photosynthesis= 1
...pathway similarity with GO...
TCA - TCA= 1
TCA - Photosynthesis= 0.563157894736842
Photosynthesis - TCA= 0.563157894736842
Photosynthesis - Photosynthesis= 1
```

Figure 5.13: `Similarity` module sample session for pathway similarity: output to Figure 5.9

```
hsa:TCA Photosynthesis
mmu:TCA Photosynthesis
dme:TCA Photosynthesys
...organism similarity with hierarchical measure...
hsa - hsa= 1
hsa - mmu= 0.875
hsa - dme= 0.625
mmu - hsa= 0.875
mmu - mmu= 1
mmu - dme= 0.625
dme - hsa= 0.625
dme - mmu= 0.625
dme - dme= 1
```

Figure 5.14: `Similarity` module sample session for organism similarity: output to Figure 5.10

obtain a phylogenetic tree by applying the clustering method to the averaged matrix. Notice that the user can annotate a different number of pathways to each organism. In such cases, whenever two organisms are being compared and a certain pathway is present in one of the organisms but not in the other, the similarity for that pathway would be 0. All experiments presented in this thesis and the web server implementation previously described use only those pathways common to the set of organisms under study. The standalone distribution, on the other hand, allows the use of non-common pathways.

Chapter 6

Conclusion

Since we now witness its end, some past moment must have witnessed its beginning

William James [67]

6.1 Summary

Metabolism is the transformation of chemical compounds inside the cell. Biosynthesis of complex molecules to perform cellular functions (anabolism), and their breakdown to generate energy (catabolism) are fundamental processes in any living organism. These processes are usually grouped into collections of enzymatic reactions called metabolic pathways. The study of the differences among such pathways in different species is a problem which has only been recently addressed in a systematic way. In this thesis, we have specifically focused on how to align pathways in a biologically relevant way. We have presented a new approach for metabolic pathway alignment based on a measure of metabolic similarity, and a series of applications of this method to biological problems.

Chapter 3 described our approach for pathway alignment. This approach is based on algebra of sets, considering pathways as sets of reactions, which are composed of compounds and enzymes. By using a measure of compound similarity, three different enzyme similarity measures, and a weight parameter α to establish the relative weight of compounds and enzymes, we can compute how similar two given reactions are. A maximum-scoring alignment of all reactions involved in the metabolic pathways under study thus produces an alignment and a similarity score of the pathways.

Three different applications of our method were presented in Chapter 4. Horizontal gene transfer events can hinder phylogenetic reconstruction from sequence data, and we therefore showed how our measure of metabolic similarity can be used to reconstruct robutst phylogenies in section 4.1. Section 4.2 described how to detect conserved (perfectly aligned) and non-conserved (non-alignable) reactions, and the relevance of such reactions. Finally section 4.3 introduces the idea of applying our approach to detect model organisms similar to humans under certain metabolic conditions.

6.2 Contributions

In this thesis, we have made three main contributions. First, we introduce a *new method for metabolic pathway alignment* based on a measure of similarity of the enzymes, compounds and reactions involved in the pathway, which has several advantages over previous approaches. We do not utilize sequence information, thus avoiding problems with horizontal gene transfer events. There is no need to artificially introduce a gap penalty for missing reactions, since our algorithm allows for alignment of non-identical reactions. In contrast with methods based on an enzyme-enzyme relational representation, we make use of both enzymes and metabolic compounds to align metabolic pathways. The use of a weight parameter allows us to balance the relevance of enzymes and compounds in the final assessment of pathway similarity. Our method is also computationally faster than those based on graph similarity.

Second, we introduced *applications of this method to problems of biological interest.* Phylogenetic reconstruction from metabolic similarity was more accurate than previous approaches, and different modifications to the original algorithm were introduced showing how to further improve results. We also detected conserved and non-conserved reactions in a group of organisms, linking the first to fundamental biological processes, and the latter to possible misannotations in KEGG. Furthermore, we proposed a method to identify appropriate model organisms for the study of specific metabolic conditions in humans.

Finally, we present a *web server to reconstruct phylogenetic relationships* among a set of organisms *by using their metabolic similarity* as calculated by our algorithm. This server, which is optimized to answer queries in linear time, can be useful to bioinformaticians interested in phylogenetics and metabolism evolution. A series of publicly available Perl modules implementing our algorithm were also introduced to help in the manipulation of metabolic data.

6.3 Future directions

Several lines of research could be conducted from the work presented in this thesis. First, it would be interesting to infer the metabolic characteristics of the common ancestor of a group of species to understand the mechanisms of emergence and evolution of biochemical pathways. Different models of metabolic evolution have been proposed (see [84] for a review), and although current observations tend to support the patchwork evolution model [147], certain research questions still remain open [120]. Furthermore, there is much controversy on whether current organisms evolved from one single common universal ancestor [35, 47, 111, 114, 119, 141], and if so, what characteristics would such last common ancestor possess [6, 87, 112]. We could address these questions as follows: given a phylogeny for a set of species, and using the metabolic alignment produced by our method, we would move from the tips (species) to the root (common ancestor) of the phylogenetic tree transferring up only those reactions conserved with a certain similarity in all nodes below. In such way we would obtain a tree having both tips and inner nodes annotated with metabolic reactions that would depict the evolutionary history of the set of species.

We could also investigate how significative the alignments between different biological networks are. *Signal transduction networks* [56], for instance, are a set of reactions inside the cell by which some kind of stimulus (such as heat or light) is converted into a response (activation of genes, initiation of metabolic processes, etc.). In bacteria, the variety of signal transduction processes directly influences in how many ways it can respond to its environment, while in plants or animals this also holds although in a less direct way. It could therefore be of interest to understand which parts of the transduction network are conserved among species and how they are linked to responses fundamental for the survival of organisms.

The study of similarities among species in *gene regulatory networks* [26, 81] could also be of great interest. Gene networks are groups of DNA segments interacting with each other and controlling whether and how genes will be transcribed into mRNA. These networks vary extremely in complexity, from the relatively simple bacterial structures to the most complex structures in higher species. Understanding similarities among such networks can be of great usefulness, for instance in the identification of gene functions or therapeutic compounds in certain diseases [129].

Appendix A

Algorithm Order Analysis

> *For many years I have been convinced that computer science is primarily the study of algorithms*
>
> Donald E. Knuth [72]

A.1 Worst-case analysis

The worst-case complexity of the algorithm presented in Chapter 3 can be calculated as a function of the complexity of calculating similarity among set of compounds, enzymes, and reactions.

The similarity of two compounds can be calculated in order $O(1)$, since we are only establishing whether the compounds are the same or not. The similarity of two sets of p and q compounds can therefore be calculated by finding the intersection of the sets, which is $O(p+q)$.

Similarity of enzymes depends on the enzyme similarity measure to be used. Hierarchical similarity can be directly calculated in $O(1)$. For information content similarity we calculate the normalized size of the subtree rooted at the least common ancestor. Precalculating all subtree sizes simply requires a traversal of the EC hierarchy, which takes $O(V_{ec})$, with V_{ec} being the number of nodes. The similarity of any two enzymes can then be calculated in $O(1)$ using the precalculated values. For the gene ontology similarity measure, we can again precalculate all distances in the Gene Ontology using Dijkstra's algorithm [28, §24.3] and then lookup the required value in $O(1)$. Since the Gene Ontology is a sparse graph, Dijkstra's algorithm can be efficiently implemented using a Fibonacci heap [28, §20] in $O(E_{go} + V_{go} log V_{go})$, with V_{go} and E_{go} being the number of edges and vertices respectively. Calculating all versus all distances using this procedure takes then $O(V_{go}^2 E_{go} + V_{go}^3 log V_{go})$.

The similarity of two reactions R and R', with C and C' compounds, and E and E' enzymes depends on calculating the intersection of the set of compounds, the intersection of the set of enzymes, and the difference of the set of enzymes (see Section 3.3). Given two ordered sets, the intersection and difference can be calculated in order equivalent to the sum of the cardinal of the sets by simultaneously traversing them. Since the similarity of compounds and enzymes can be calculated in order $O(1)$ as explained above, the order of calculating the similarity of two reactions will be $O(C + C') + O(E + E')$ plus the cost of ordering all the sets, $O(ClogC) + O(C'logC') + O(ElogE) + O(E'logE')$ [73]. In general, for any given reaction the number of compounds is greater than the number of enzymes, so the order of comparing two reactions has an upperbound order $O(ClogC)$, assuming $|C| > |C'|$.

Given two pathways A and B with m and n reactions, we can calculate the intersection and difference of the reaction sets using an algorithm of order $O(m+n)$. Following the equations presented in 3.2, we need now to calculate the similarity of each reaction in $A \setminus B$ versus all reactions in B, the cardinal of $A \cap B$, and the similarty of each reaction in $B \setminus A$ versus all reactions in A. If we define $|A \cap B| = p$, $|A \setminus B| = q$, and $|B \setminus A| = r$, the total number of calculations needed is $O(nq + p + mr)$. The value p can be previously obtained when calculating the set intersection, resulting in a final cost of $O(nq + mr) \approx O(m \times n)$.

A.2 Average-case analysis

Using results from the previous section, we will essume order $O(1)$ for the comparison of compounds and enzymes. KEGG reactions have an average of 4.35 compounds ($\sigma = 2.33$) and 0.84 enzymes ($\sigma = 0.46$). The average order of comparing two reactions is therefore bounded by $O(4.35log4.35) \approx O(6.4)$.

In order to calculate the average complexity of comparing two pathways, we want to show now that the cost of calculating the similarity values among the reaction sets, $O(nq + mr)$, does not grow substantially faster than the order of calculating the reaction sets intersection and difference, $O(m+n)$. Figure A.1 shows the evolution of these values for the glycolysis pathway in all organisms stored in KEGG compared to linear, loglinear, and quadratic growing functions. As it can be seen, the relation between $O(m+n)$ and $O(nq + mr)$ is clearly lower than quadratic in the average case. The average complexity calculated using pathways different to glycolysis does not vary substantially (results not shown).

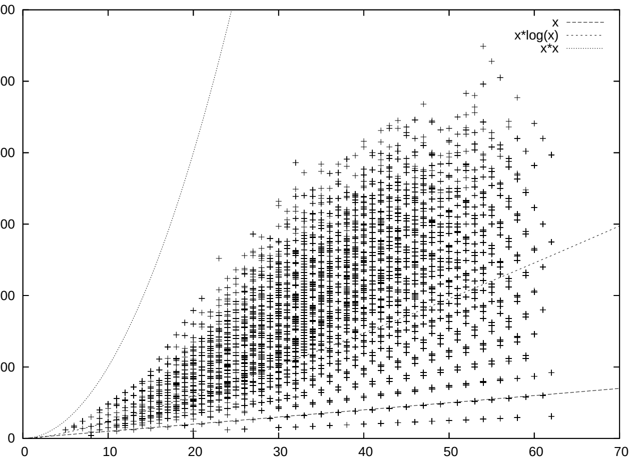


Figure A.1: Average complexity of pathway comparison for glycolysis in all KEGG organisms, compared against linear, loglinear, and quadratic functions

Appendix B

Non-conserved reactions in *S. pyogenes*, *E. coli*, and *S. aureus*

Table B.1: Gap reactions: *Streptococcus pyogenes*

Strain	**Year**	**serotype**	**Gap reactions**
sph	2006	M3	(none)
spi	2006	M3	(none)
spj	2006	M3	(none)
spk	2006	M3	(none)
spz	2005	M1	(none)
spb	2005	M28	(none)
spa	2004	M6	00118 01906 03540 03674 03730 03937 04008 04172 04594 04732 04885 04906 05202 05601 05602 06369 06906 06925
sps	2003	M3	00118 00148 00501 01906 01966 02518 02520 03113 03234 03540 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 04937 04938 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925
spm	2002	M18	00118 00148 00501 01906 01966 02518 02520 03113 03234 03540 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 04937 04938 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925
spg	2002	M3	00118 00148 00501 01906 01966 02518 02520 03113 03234 03540 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 04937 04938 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925
spy	2001	M1	00118 00148 00501 01906 02518 02520 03113 03234 03540 03544 03545 03674 03730 03937 04008 04142 04172 04360 04594 04732 04885 04906 05202 05601 05602 06021 06097 06192 06198 06369 06404 06729 06906 06925

Table B.2: Gap reactions: *Escherichia coli*

Strain	**Year**	**Gap reactions**
eci	2006	02000 02002 04986 06405 06782 06783 06784 06785 06786 06787 06920
ecp	2006	02000 02002 04910 04986 05049 05615 05617 05625 05644 05645 05646 05647 06397 06782 06783 06784 06785 06786 06787 06858
ecc	2002	00069 01452 01453 01719 01966 02000 02002 02383 02518 02912 03674 03730 03811 03937 03955 04008 04131 04172 04306 04732 04826 04857 04885 04895 04906 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06367 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06782 06783 06784 06785 06786 06787 06853 06854 06905 06906 06907 06914 06920 06925 06935
ecj	2001	00069 01452 01453 01719 02518 02912 03317 03674 03730 03811 03937 03955 04008 04131 04142 04172 04306 04313 04360 04375 04732 04809 04813 04826 04857 04885 04895 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06853 06854 06905 06906 06907 06914 06920 06925 06935
ece	2001	00069 01452 01453 01719 01966 02383 02518 02912 03674 03730 03811 03937 03955 04008 04131 04142 04172 04306 04360 04375 04515 04732 04784 04809 04813 04826 04857 04885 04895 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06853 06854 06905 06906 06907 06920 06925
ecs	2001	00069 01452 01453 01719 01966 02383 02518 02912 03674 03730 03811 03937 03955 04008 04131 04142 04172 04306 04360 04375 04515 04732 04784 04809 04813 04826 04857 04885 04895 04910 04916 04917 05001 05448 05449 05504 05505 05602 05623 06369 06396 06397 06398 06400 06405 06406 06407 06735 06736 06853 06854 06905 06906 06907 06920 06925
eco	1997	00069 01452 01453 01966 02518 02912 03674 03730 03811 03937 03955 04131 04172 04306 04732 04826 04857 04885 04895 04906 04910 05001 05448 05449 05504 05505 05602 05623 06367 06369 06396 06397 06398 06400 06405 06406 06853 06854 06906 06920 06925

Table B.3: Gap reactions: *Staphylococcus aureus*

Strain	**Year**	**Gap reactions**
saa	2006	01966 03113 03234 03540 03544 03545 04142 04360 04937 04938 06405 06920
sao	2006	00118 01452 01453 02383 03811 04254 04306 04313 04594 04826 05001 05118 05623 06398 06400 06405 06406 06413 06916 06920 06926
sac	2005	00118 01452 01453 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
sab	2005	00118 01452 01453 02383 03811 04306 04594 04826 05001 06372 06398 06400 06406
sar	2004	00118 01452 01453 01719 01966 02383 02528 03113 03234 03730 03811 04131 04142 04306 04360 04594 04809 04813 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
sas	2004	00118 01452 01453 01719 01966 02383 02528 03113 03234 03730 03811 04131 04142 04306 04360 04594 04809 04813 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
sam	2002	00118 01452 01453 01719 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
sau	2001	00118 01452 01453 01719 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935
sav	2001	00118 01452 01453 01719 01966 02383 03730 03811 04131 04306 04594 04826 04895 04916 04917 04937 04938 05001 05601 05602 06369 06372 06398 06400 06405 06406 06407 06905 06906 06914 06920 06935

Table B.4: Enzymes catalyzing gap reactions: *Streptococcus pyogenes*

Enzyme	**# reacs.**	**Enzyme**	**# reacs.**	**Enzyme**	**# reacs.**
1.1.1.-	49	3.1.1.-	7	6.3.2.-	5
4.1.1.-	28	3.5.1.-	6	2.7.1.-	5
3.2.1.-	24	3.1.3.73	5	1.14.18.1	4
1.13.12.-	14	3.1.3.-	5		
1.14.-.-	7	5.3.1.-	5		

Table B.5: Enzymes catalyzing gap reactions: *Escherichia coli*

Enzyme	**# reacs.**
1.14.13.-	36
4.1.1.-	23
1.2.1.-	19
4.2.1.-	18
1.1.1.-	16
2.5.1.-	15
2.7.-.-	10
1.3.1.-	9
1.18.1.4	8
1.14.12.-	8
1.14.12.19	8
5.-.-.-	6
2.3.1.-	6
3.4.-.-	6
3.1.-.-	6
1.1.-.-	5
1.18.1.1	4
2.1.1.-	4
1.2.1.71	4
1.14.18.1	4
1.2.1.24	4
6.2.1.-	3
3.6.1.-	3
3.2.1.-	3
1.14.13.95	2
2.7.1.-	2
4.2.1.107	2
1.97.1.-	2
6.3.2.-	2
1.3.99.-	2
3.5.3.-	1

Table B.6: Enzymes catalyzing gap reactions: *Staphylococcus aureus*

Enzyme	**# reacs.**
1.1.1.-	42
4.2.1.-	27
1.2.1.-	15
1.14.13.-	13
2.3.1.-	8
3.5.1.-	8
3.4.-.-	8
3.2.1.-	6
3.6.1.-	4
1.1.-.-	3
1.3.-.-	3
3.1.1.-	2
3.1.3.73	1
3.1.3.-	1
1.14.18.1	1
1.3.99.-	1

Appendix C

Algorithms for Pathway Similarity

Algorithm 1 SIMILARITY(P, P')

```
$R_P \leftarrow$ reactions of $P$
$R_{P'} \leftarrow$ reactions of $P'$
$pathwaysim \leftarrow 0$

%%% Pathway similarity: intersection %%%
for all $r \in R_P \cap R_{P'}$ do
  $pathwaysim++$
end for

%%% Pathway similarity: reactions in $P$ but not in $P'$ %%%
for all $r \in R_p \setminus R_{P'}$ do
  $max \leftarrow 0$
  for all $s \in R_{P'}$ do
    if $sim(r, s) > max$ then
      $max \leftarrow s$
    end if
  end for
  $pathwaysim \leftarrow pathwaysim + sim(r, max)$
end for

%%% Pathway similarity: reactions in $P$ but not in $P'$ %%%
for all $s \in R_{P'} \setminus R_P$ do
  $max \leftarrow 0$
  for all $r \in R_P$ do
    if $sim(s, r) > max$ then
      $max \leftarrow r$
    end if
  end for
  $pathwaysim \leftarrow pathwaysim + sim(s, max)$
end for
return $pathwaysim$
```

Algorithm 2 SIMILARITY(R, R')

```
$reacsim \leftarrow 0$
$C_R \leftarrow$ compounds of $R$
$C_{R'} \leftarrow$ compounds of $R'$
$E_R \leftarrow$ enzymes of $R$
$E_{R'} \leftarrow$ enzymes of $R'$

$reacsim \leftarrow (1-\alpha) \times compoundsim(C_R, C_{R'}) + \alpha \times enzymesim(E_R, E_{R'})$
return $reacsim$
```

Algorithm 3 SIMILARITY($C_R, C_{R'}$)

```
$compoundsim \leftarrow 0$
$C_R \leftarrow$ compounds of $R$
$C_{R'} \leftarrow$ compounds of $R'$

%%% Compound similarity: intersection %%%
for all $c \in C_R \cap C_{R'}$ do
  $compoundsim++$
end for

%%% Compound similarity: compounds in $R$ but not in $R'$ %%%
for all $c \in C_R \setminus C_{R'}$ do
  $max \leftarrow 0$
  for all $d \in C_{R'}$ do
    if $sim(c, d) > max$ then
      $max \leftarrow d$
    end if
  end for
  $compoundsim \leftarrow compoundsim + sim(c, max)$
end for

%%% Compound similarity: compounds in $R'$ but not in $R$ %%%
for all $d \in C_{R'} \setminus C_R$ do
  $max \leftarrow 0$
  for all $c \in C_R$ do
    if $sim(d, c) > max$ then
      $max \leftarrow c$
    end if
  end for
  $compoundsim \leftarrow compoundsim + sim(d, max)$
end for
return $compoundsim$
```

Algorithm 4 SIMILARITY($E_R, E_{R'}$)

```
$enzymesim \leftarrow 0$
$E_R \leftarrow$ enzymes of $R$
$E_{R'} \leftarrow$ enzymes of $R'$
%%% Enzyme similarity: intersection %%%
for all $e \in E_R \cap E_{R'}$ do
    $enzymesim + +$
end for
%%% Enzyme similarity: enzymes in $R$ but not in $R'$ %%%
for all $e \in E_R \setminus E_{R'}$ do
    $max \leftarrow 0$
    for all $f \in E_{R'}$ do
        if $sim(e, f) > max$ then
            $max \leftarrow f$
        end if
    end for
    $enzymesim \leftarrow enzymesim + sim(e, max)$
end for
%%% Enzyme similarity: enzymes in $R'$ but not in $R$ %%%
for all $f \in E_{R'} \setminus E_R$ do
    $max \leftarrow 0$
    for all $e \in E_R$ do
        if $sim(f, e) > max$ then
            $max \leftarrow e$
        end if
    end for
    $enzymesim \leftarrow enzymesim + sim(f, max)$
end for
return $enzymesim$
```

Appendix D

Pathway Alignment Through Context Similarity

For a long time completion was like a dream, I even felt such a thing would never be permitted to happen. On the other hand I could picture myself dancing barefoot over the garden on the morning when I finished.

Yukio Mishima [104]

D.1 Introduction

In Section 3.1 we presented an approach for metabolic pathway alignment using three different enzyme similarity measures. In this section we will introduce a new approach for pathway alignment based on the contextual similarity of reactions, this is, the more similar the context of two reactions is, the higher their similarity score. Contextual similarity is based on the distributional hypothesis by Zellig Harris [57]:

> The meaning of entities, and the meaning of grammatical relations among them, is related to the restriction on combinations of these entities relative to other entities

Roughly speaking, the distributional hypothesis states that words appearing in similar contexts tend to have similar semantics. By adapting this idea to metabolic similarity, we would like to test the hypothesis that *metabolic reactions playing similar functional roles tend to appear in similar contexts*. A functional role here is to be understood as the biological function of a reaction in a wide sense, either from an enzymatic point of view (is the reaction an oxidating process? a hydrolytic process?) or by considering the metabolites being used as substrates/products. A context will therefore contain those reactions catalyzed by similar enzymes or synthesizing/degrading similar compounds.

Figures D.1 and D.2 show examples of reactions with similar contexts according to their compounds and their enzymes. Given a reaction R, reaction S will be in the context of R if some of its products are used by R as substrates, or some of its substrates are produced by R as products. Therefore the compound context similarity tries to measure the topological similarity of the subnetwork in which the reactions are taking place. In

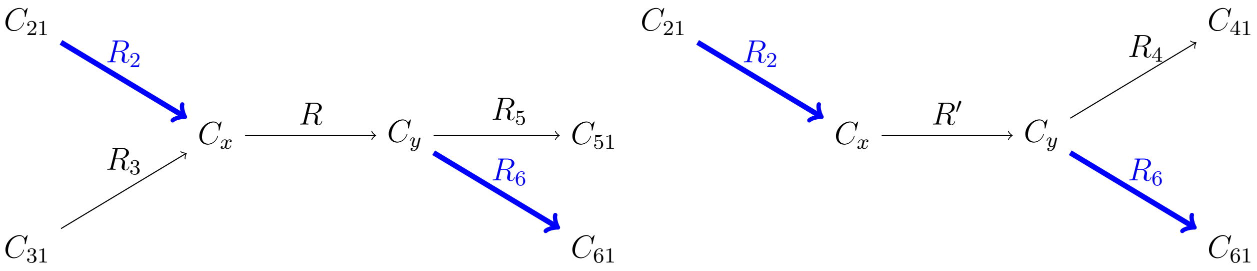


Figure D.1: Compound context of R and R'. The context of R (left) is the set $\{R_2, R_3, R_5, R_6\}$, while the context of R' (right) is $\{R_2, R_4, R_6\}$. The shared context is $\{R_2, R_6\}$ (in blue/solid)

Figure D.1, reactions R and R' have a significant degree of similarity, as they respectively share 2 out of 4, and 2 out of 3 of the reactions appearing in their compound contexts.

The enzymatic context, on the other hand, does not require for two reactions to share any compounds to be in each other's context. Reactions can be part of different metabolic processes and still be contextual, as long as they have an enzymatic similarity above a certain threshold k. Reactions R_7 and R_9 in Figure D.2 (up) do not share any compound with R but they belong to its enzymatic context since their enzymatic similarity is greater than the specified threshold. We are clustering into a context those reactions that share a certain metabolic functionality by using the EC identifiers of their catalyzing enzymes. Since EC identifiers are constructed to reflect a hierarchy of biochemical roles [137], we argue that our approach does in fact reconstruct sets of functionally related reactions.

In general, the context of a reaction will be defined as the set of reactions sharing some compounds with it or being catalyzed by similar enzymes (Section D.2 provides a more formal definition). In the rest of this chapter, we will test the hypothesis on contextual similarity presented above and compare results using contextual similarity to those obtained with the approach presented in Chapter 3 (non-contextual similarity).

D.2 Materials and methods

D.2.1 Contextual similarity

We will consider a reaction R as a set of enzymes, $enz(R)$, and a set of compounds, $cpd(R)$, which can be divided into the set of substrates, $sub(R)$, and the set products, $prod(R)$, with $cpd(R) = sub(R) \cup prod(R)$.

The *compound context* C_{cpd} of a reaction R in a fixed, but arbitrary metabolic pathway M, is the set of all reactions in M that share some substrate or product with it:

$$C_{cpd}(R, M) = \{R' \in M \mid R' \neq R \wedge ((prod(R) \cap sub(R') \neq \emptyset) \vee (prod(R') \cap sub(R) \neq \emptyset))\} \tag{D.1}$$

The *enzymatic context* $C_{enz}(R)_k$ of a reaction R in a pathway M is the set of all reactions that have enzymatic similarity greater than k with reaction R:

$$C_{enz}(R, M)_k = \{R' \in M \mid R' \neq R \wedge sim_{enz}(R, R') \geqslant k\} \tag{D.2}$$

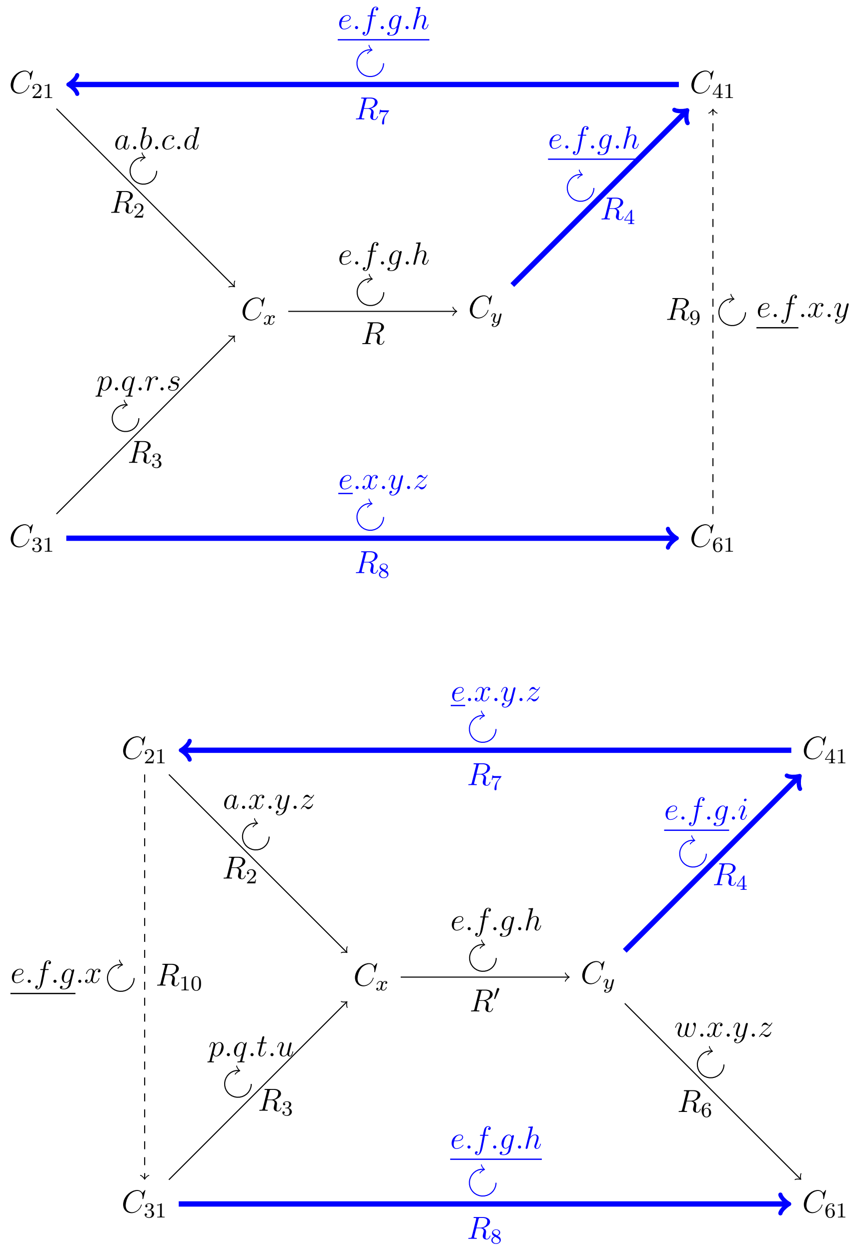


Figure D.2: Enzymatic context of reactions R (up) and R' (down) for hierarchical enzymatic similarity $\geqslant 0.25$. Since all reactions area catalyzed by a single enzyme, any reaction catalyzed by an enzyme sharing at least the most significant EC digit with R or R' will appear in their context. The context of R is $\{R_4, R_7, R_8, R_9\}$, while the context of R' is $\{R_4, R_7, R_8, R_{10}\}$. Reaction R_9, for instance, is in the context of R since it shares two EC digits of its catalyzing enzyme with R ($e.f$, underlined). The shared context of R and R' is then $\{R_4, R_7, R_8\}$ (in blue/solid). Notice how R_7 and R_8 do not share any compounds with R or R' but they are part of the shared context since the catalyzing enzymes of R_7 and R_8 are similar enough to those catalyzing R and R'. Reactions belonging to the context of R or R' but not in their shared context are presented with dashed lines (R_9 and R_{10})

Values of k can range from 0 to 1, since the enzymatic similarity measure is normalized. The higher the value of k, the more similar enzymes in two reactions should be to consider them contextual. For instance, if a reaction R is catalyzed by the enzyme $e = 1.2.3.4$ and k is set to 0.5, only reactions catalyzed by enzymes sharing at least the two most significative digits with e (that is, only 1.2.) would be contextual to R.[1] Setting $k = 1$ would require reactions to share exactly the same enzymes, and setting it to 0 would mean any reaction is enzymatically contextual to any other.

The *context* of a reaction $C(R, M)_k$ in a metabolic pathway M is defined as the union of its compound context $C_{cmp}(R, M)$ and its enzymatic context $C_{enz}(R, M)_k$.

$$C(R, M)_k = C_{cmp}(R, M) \cup C_{enz}(R, M)_k \tag{D.3}$$

Finally, we define the *contextual similarity* of reactions R_1 and R_2 in a metabolic pathway M for a given k as:

$$sim_c(R_1, R_2, M)_k = \frac{|C(R_1, M)_k \cap C(R_2, M)_k|}{|C(R_1, M)_k \cup C(R_2, M)_k|} \tag{D.4}$$

D.2.2 Experimental setup

All data used in our experiments was obtained from KEGG release 39.0 (July 2006). We chose organisms *Drosophila melanogaster* (*dme*), *Escherichia coli* (*eco*) and *Archaeoglobus fulgidus* (*afu*), since they represent the three domains of life and complete sequence information is available for all of them. We retrieved the 68 shared pathways among these organisms containing at least one reaction. To calculate the contextual similarity of reactions, we first create a context graph $G = (V, E)$ with all reactions present in KEGG, where a node v represents a reaction and an edge $e = (u, v)$ means that reactions u and v are contextual according to a certain value of the parameter k. We then calculate the contextual similarity for any two nodes of the graph and store it in a similarity matrix, which is then used to obtain the similarity for all reactions present in the shared pathways among *dme*, *eco* and *afu*. Non-contextual similarity among reactions was calculated using the algorithm described in Chapter 3, with parameter $\alpha = 0.5$. Similarity of enzymes was calculated using the hierarchical similarity measure.

Once similarity of reactions was calculated using both methods we compared results by studying the respective alignments produced. We sorted all reactions according to the difference between contextual and non-contextual similarity values, and analyzed those reactions which showed a larger similarity value difference.

D.3 Results and discussion

For each pair of reactions, we calculated their contextual and non-contextual similarity values, and then we obtained the absolute difference between these two measures. Figure D.3 is a histogram representing such differences for all pairs of reactions. As it can be seen, the great majority of calculated values are very similar with both measures: 2161 ($k = 0.5$) and 2214 ($k = 0.75$) values have a difference in similarity less than 0.1. The

[1] Since we are using hierarchical enzyme similarity, enzymes *a.b.c.d* and *a.b.x.y* would have exactly similarity 0.5

number of reaction pairs with large difference in their similarity values decreases as we look for larger differences: nearly 300 pairs have difference between 0.1 and 0.2, while there were 127 with difference between 0.6 and 0.7 and only 18 between 0.7 and 0.8 (for $k = 0.5$).

It is interesting to notice though the increase in number of pairs of reactions with differences between 0.9 and 1. This large difference means that while one of the measures scores the reactions as nearly identical, the other measure considers these reactions share little resemblance. We investigated what reactions appear often in pairs where the difference between contextual and non-contextual similarity is larger than 0.9, and results are presented in table D.1. For instance, for $k = 0.75$ reaction R4986 appears 11 times in reaction pairs with large disagreement between the similarity scores: R04732 (similarity difference: 0.9504), R04008 (0.9504), R04172 (0.9349), R06925 (0.9055), etc.

For those reactions pairs with large differences in similarity values, we found out contextual measure was always the one scoring high similarity, while non-contextual gave low similarity scores. In fact, we found that the contextual measure is systematically biased towards higher scores. Detailed analysis of the results shows that some reactions have an extremely large context due to the presence of metabolites such as H^+ (C00080), which were not previously identified as common metabolites. The presence of such metabolites in a large number of reactions incorrectly increases the similarity score, and the contextual similarity appears to be too sensitive to this noise. The non-contextual measure, on the other hand, is more resilient to the presence of those metabolites and can calculate more accurately the similarity among reactions. Given that common metabolites act as cofactors depending on the reaction, we therefore argue that the less any method for pathway alignment relies on lists of common metabolites, the better.

Because we have limited our experiments to reactions appearing in an extremely reduced set of organisms, we suspect several other common metabolites could have been missed. We plotted a histogram of compounds versus number of reactions in which they appear (Figure D.4), in order to visually determine what is a good “cut point” to establish which metabolites are common and which are not. From a practical point of view, this can help to eliminate bias towards high similarity scores for those reactions containing common metabolites.

A more sophisticated way to reduce this bias would be to determine which substrates in a reaction are being transformed into which products, and which metabolites are acting as cofactors (understood as non-protein compounds required to assist in biochemical transformations). Similarity scores could then be adjusted to prioritize those reactions where the transformed substrates or products are similar, rather than those where only the cofactors are.

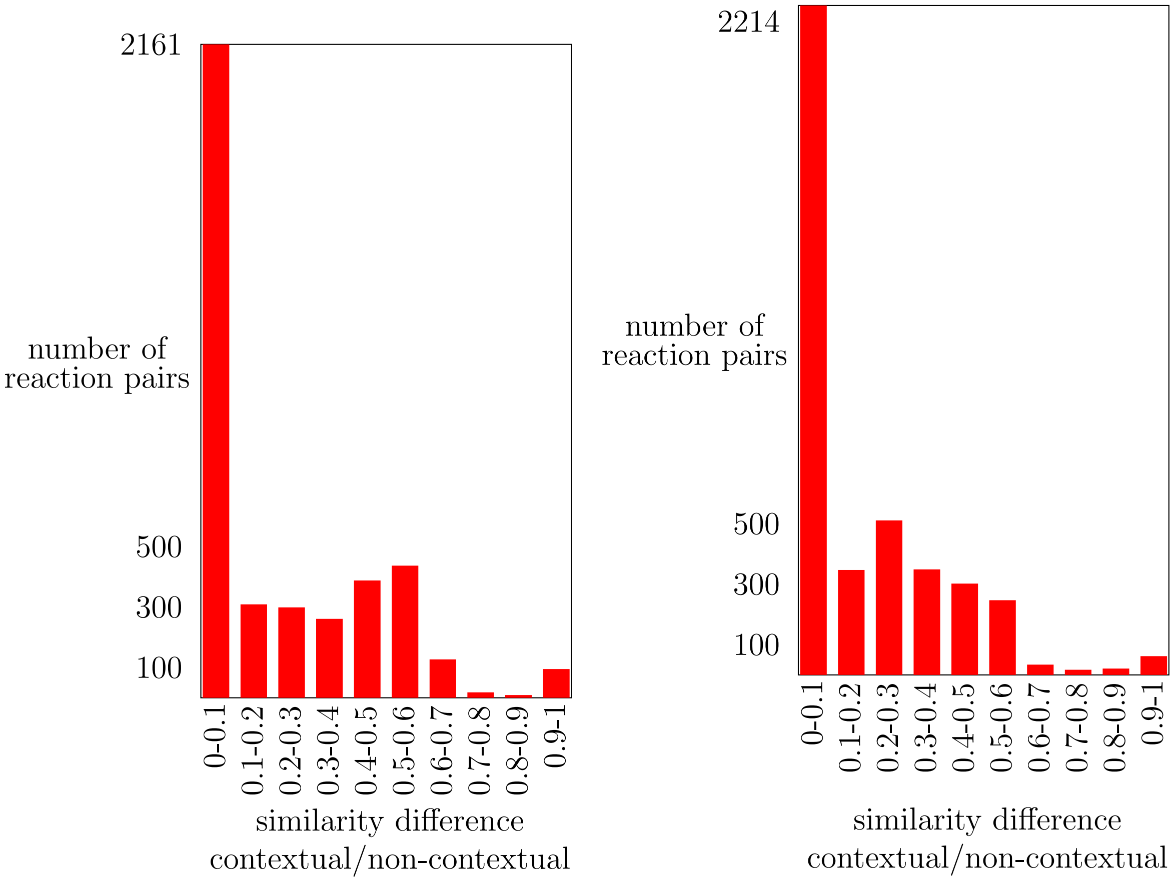


Figure D.3: Histogram for differences between contextual and non-contextual similarity measure, $k = 0.5$ (left) and $k = 0.75$ (right)

Reaction	**Large differences**
R04737	25
R04456	21
R03299	20
R01813	16
R04778	16
R04986	13
R03348	13
R04031	10
R06728	9
R04355	8
R04984	6
R00219	6
R00750	5

Reaction	**Large differences**
R04737	16
R04778	14
R04986	13
R04001	11
R04984	6
R03299	6
R00132	5
R04506	5
R04993	4
R04810	4
R06728	4
R00084	4
R04405	4

Table D.1: Top reactions for which the similarity value calculated using contextual and non-contextual methods is larger than 0.9 with $k = 0.5$ (left) and $k = 0.75$ (right). *Large differences* represents the number of reaction pairs where the difference is greater than 0.9

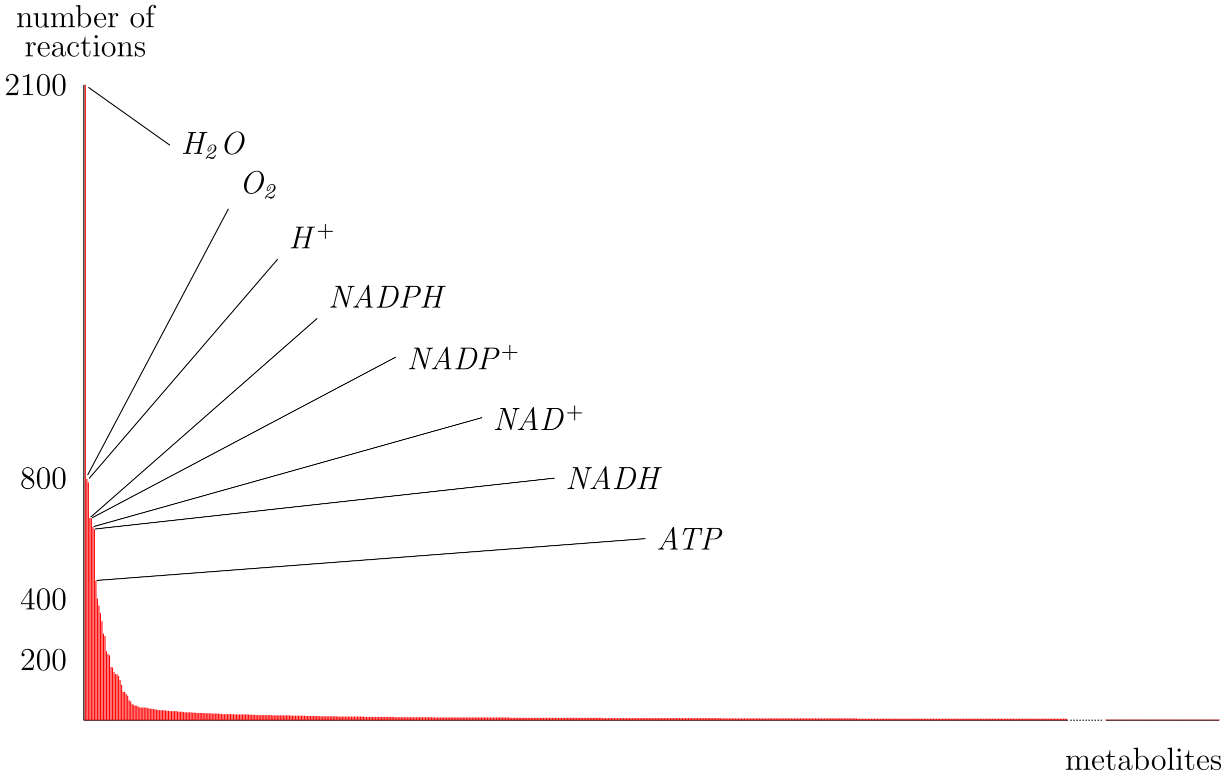


Figure D.4: Metabolite histogram: number of reactions in which a metabolite appears

Bibliography

[1] D. Aguilar, F. X. Aviles, E. Querol, and M. J. E. Sternberg. Analysis of Phenetic Trees Based on Metabolic Capabilities Across the Three Domains of Life. *Journal of Molecular Biology*, 340(3):491–512, 2004.

[2] T. Aittokallio and B. Schikowski. Graph-based methods for Analysing Networks in Cell Biology. *Briefings in Bioinformatics*, 7(3):243–255, 2006.

[3] G. Alberti. Noncommunicable Diseases: Tomorrow's Pandemics. *Bulletin of the World Health Organization*, 79(10):907, 2001.

[4] E. Almaas, Z. N. Oltvai, and A.-L. Barabási. The Activity Reaction Core and Plasticity of Metabolic Networks. *PLOS Computational Biology*, 1(7):0557–0563, 2005.

[5] S. F. Altschul et al. Basic Local Alignment Search Tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

[6] S. G. E. Andersson et al. The Genome Sequence of *Rickettsia prowazekii* and the Origin of Mitochondria. *Nature*, 396(6707):133–140, 1998.

[7] M. Arita. The Metabolic World of *Escherichia coli* is Not Small. *Proceedings of the National Academy of Sciences, USA*, 101(6):1543–1547, 2004.

[8] M. Ashburner et al. Gene Ontology: Tool for the Unification of Biology. *Nature Genetics*, 25(1):25–29, 2000.

[9] T. Baba et al. Genome and Virulence Determinants of High Virulence Community-Acquired MRSA. *The Lancet*, 359(9320):1819–1827, 2002.

[10] F. Bacon. *Wisdom of the Ancients*. Kessinger Publishing, New York NY, USA, 1997.

[11] Y. Balmer et al. A Complete Ferredoxin/Thiredoxin System Regulates Fundamental Processes in Amyloplasts. *Proceedings of the National Academy of Sciences, USA*, 103(8):2988–2993, 2006.

[12] L. Baronciani and E. Beutler. Analysis of Pyruvate Kinase-Deficiency Mutations that Produce Nonspherocytic Hemolytic Anemia. *Proceedings of the National Academy of Sciences, USA*, 90(9):4324–4327, 1993.

[13] S. B. Beres et al. Molecular Genetic Anatomy of Inter- and Intraserotype Variation in the Human Bacterial Pathogen Group *A. streptococcus*. *Proceedings of the National Academy of Sciences, USA*, 103(18):7059–7064, 2006.

[14] Y. Boucher et al. Lateral Gene Transfer and the Origins of Prokaryotic Groups. *Annual Reviews of Genetics*, 37:283–328, 2003.

[15] L. J. Brewerton, E. Fung, and F. F. Snyder. Polyethylene Glycol-Conjugated Adenosine Phosphorylase: Development of Alternative Enzyme Therapy for Adenosine Deaminase Deficiency. *Biochimica et Biophysica Acta - Molecular Basis of Disease*, 1637(2):171–177, 2003.

[16] J. W. Campbell and J. E. Cronan. Bacterial Fatty Acid Biosynthesis: Targets for Antibacterial Drug Discovery. *Annual Review of Microbiology*, 55(1):305–332, 2001.

[17] J. Casasnovas, J. C. Clemente, J. Miró-Julià, F. Rosselló, K. Satou, and G. Valiente. Fuzzy clustering improves phylogenetic relationships reconstruction from metabolic pathways. In *Proc. 11th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 2807–2814. Editions EDK, 2006.

[18] R. Caspi et al. MetaCyc: a Multiorganism Database of Metabolic Pathways and Enzymes. *Nucleic Acids Research*, 34 (Database issue):D511–D516, 2006.

[19] M.-J. Chung. $O(n^{2.5})$ Times Algorithms for the Subgraph Homeomorphism Problems on Trees. *Journal of Algorithms*, 8(1):106–112, 1987.

[20] A. G. Clark et al. Inferring Nonneutral Evolution from Human-Chimp-Mouse Orthologous Gene Trios. *Science*, 302(5652):1960–1963, 2003.

[21] G. D. P. Clarke, R. G. Beiko, M. A. Ragan, and R. L. Charlebois. Inferring Genome Trees by Using a Filter to Eliminate Phylogenetically Discordant Sequences and a Distance Matrix Based on Mean Normalized BLASTP Scores. *Journal of Bacteriology*, 184(8):2072–2080, 2002.

[22] J.-M. Claverie. What If There are Only 30,000 Human Genes? *Science*, 291(5507):1255–1257, 2001.

[23] J. Clemente, K. Satou, and G. Valiente. Reconstruction of Phylogenetic Relationships from Metabolic Pathways based on the Enzyme Hierarchy and the Gene Ontology. *Genome Informatics*, 16(2):45–55, 2005.

[24] J. Clemente, K. Satou, and G. Valiente. Finding Conserved and Non-Conserved Reactions Using a Metabolic Pathway Alignment Algorithm. *Genome Informatics*, 17(2), 2006. In press.

[25] J. Clemente, K. Satou, and G. Valiente. Phylogenetic Reconstruction from Non-Genomic Data. *Bioinformatics*, 2007. In press.

[26] J. Collado-Vides and R. Hofestädt, editors. *Gene Regulation and Metabolism: Postgenomic Computational Approaches*. The MIT Press, Cambridge MA, USA, 2004.

[27] F. S. Collins et al. A Vision for the Future of Genomics Research. *Nature*, 422(6934):835–847, 2003.

[28] T. H. Cormen, C. S. Leiserson, R. L. Rivest, and C. Stein, editors. *Introduction to Algorithms*. The MIT Press, Cambridge MA, USA, 2001.

[29] T. Dandekar, S. Schuster, B. Snel, M. jn Huynen, and P. Bork. Pathway Alignment: Application to the Comparative Alignment of Glycolytic Enzymes. *The Biochemical Journal*, 343(1):115–124, 1999.

[30] M. Daugherty, V. Vonstein, R. Overbeek, and A. Osterman. Archaeal Shikimate Kinase, a New Member of the GHMP-Kinase Family. *Journal of Bacteriology*, 183(1):292–300, 2001.

[31] P. de Matos et al. ChEBI - Chemical Entities of Biological Interest. *Nucleic Acid Research*, 2006. Database Summary Paper 646.

[32] R. Descartes. *Discourse on Method and Meditations on First Philosophy*. Hackett Publishing Company, 1998.

[33] Y. Deville, D. Gilbert, J. van Helden, and S. J. Wodak. An Overview of Data Models for the Analysis of Biochemical Pathways. *Briefings in Bioinformatics*, 4(3):246–259, 2003.

[34] E. W. Dijkstra. The Humble Programmer. *Communications of the ACM*, 15(10):859–866, 1972.

[35] W. F. Doolittle. Uprooting the Tree of Life. *Scientific American*, 282(2):90–95, 2000.

[36] O. Ebenhöh, T. Handorf, and R. Heinrich. Structural Analysis of Expanding Networks. *Genome Informatics*, 15(1):35–45, 2004.

[37] I. Ebersberger, D. Metzler, C. Schwarz, and S. Pääbo. Genomewide Comparison of DNA Sequences between Humans and Chimpanzees. *The American Journal of Human Genetics*, 70(6):1490–1497, 2002.

[38] L. B. M. Ellis, D. Roe, and L. P. Wackett. The University of Minnesota Biocatalysis/Biodegradation Database: the First Decade. *Nucleic Acids Research*, 34 (Database issue):D517–D521, 2006.

[39] W. E. Evans and H. L. McLeod. Pharmacogenomics – Drug Disposition, Drug Targets, and Side Effects. *The New England Journal of Medicine*, 348(6):538–549, 2003.

[40] L. Félix and G. Valiente. Efficient validation of metabolic pathway databases. In *Proc. 6th Int. Symp. Computational Biology and Genome Informatics*, pages 1209–1212, Salt Lake City, Utah, 2005.

[41] L. Fèlix and G. Valiente. Validation of Metabolic Pathway Databases based on Chemical Substructure Search. *Biomolecular Engineering*, 2006. In press.

[42] J. Felsenstein. *Inferring Phylogenies*. Sinauer Associates, Inc., Sunderland MA, USA, 2004.

[43] O. Fiehn. Metabolomics — the Link between Genotypes and Phenotypes. *Plant Molecular Biology*, 48(1-2):155–171, 2002.

[44] 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(67), 2005.

[45] C. V. Forst and K. Schulten. Evolution of Metabolisms: A New Method for the Comparison of Metabolic Pathways using Genomic Information. *Journal of Computational Biology*, 6(3–4):343–360, 1999.

[46] C. V. Forst and K. Schulten. Phylogenetic Analysis of Metabolic Pathways. *Journal of Molecular Evolution*, 52(1):471–489, 2001.

[47] P. Forterre and H. Philippe. Where is the Root of the Universal Tree of Life? *BioEssays*, 21(10):871–879, 1999.

[48] M. R. Garey and D. S. Johnson. *Computers and Intractability: a Guide to the Theory of NP-Completeness.* W. H. Freeman, New York NY, USA, 1979.

[49] E. Gasteiger et al. ExPASy: the Proteomics Server for In-Depth Protein Knowledge and Analysis. *Nucleic Acids Research*, 31(13):3784–3788, 2003.

[50] Y. Gilad, A. Oshlack, G. K. Smyth, T. P. Speed, and K. P. White. Expression Profiling in Primates Reveals a Rapid Evolution of Human Transcription Factors. *Nature*, 440(7081):242–245, 2006.

[51] G. Glazko, V. Veeramachaneni, M. Nei, and W. Makalowski. Eighty Percent of Proteins are Different between Humans and Chimpanzees. *Gene*, 346:215–219, 2005.

[52] S. Goto, H. Bono, H. Ogata, W. Fujibuchi, T. Nishioka, and K. Sato. Organizing and Computing Metabolic Pathway Data in Terms of Binary Relations. In *Proc. 2nd Pacific Symposium on Biocomputing*, pages 175–186, 1996.

[53] M. L. Green and P. D. Karp. Genome Annotation Errors in Pathway Databases due to Semantic Ambiguity in Partial EC Numbers. *Nucleic Acids Research*, 33(13):4035–4039, 2005.

[54] R. S. Gupta. Protein Phylogeneies and Signature Sequences: a Reappraisal of Evolutionary Relationships Among Archaebacteria, Eubacteria, and Eukaryotes. *Microbiology and Molecular Biology Reviews*, 62(4):1435–1491, 1998.

[55] P. R. Halmos. *Naive Set Theory.* Springer-Verlag, New York NY, USA, 1960.

[56] J. T. Hancock. *Cell Signaling.* Oxford University Press, New York NY, USA, 2005.

[57] Z. Harris. Distributional Structure. In J. J. Katz, editor, *The Philosophy of Linguistics*, pages 26–47. Oxford University Press, 1985.

[58] K. Hayashi et al. Highly Accurate Genome Sequences of *E. coli* K-12 Strains MG1655 and W3110. *Molecular Systems Biology*, 2(2006.0007), 2006.

[59] S. Henikoff and J. G. Henikoff. Amino Acid Substitution Matrices from Protein Blocks. *Proceedings of the National Academy of Sciences, USA*, 89(22):10915–10919, 1992.

[60] K. Henze and W. Martin. Essence of Mitochondria. *Nature*, 426(6963):127–128, 2003.

[61] M. Heymans and A. K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. Technical Report 2002-33, available at `http://cs.ucsb.edu/research /tech_reports/reports/2002-33.pdf`.

[62] M. Heymans and A. K. Singh. Deriving Phylogenetic Trees from the Similarity Analysis of Metabolic Pathways. *Bioinformatics*, 19(Suppl. 1):i138–i146, 2003.

[63] D. G. Higgins and P. M. Sharp. CLUSTAL: a Package for Performing Multiple Sequence Alignemnt on a Microcomputer. *Gene*, 73(1):237–244, 1988.

[64] M. T. G. Holden et al. Complete Genomes of Two Clinical *Staphylococcus aureus* Strains: Evidence for the Rapid Evolution of Virulence and Drug Resistance. *Proceedings of the National Academy of Sciences, USA*, 101(26):9786–9791, 2004.

[65] S. E. Humphries and S. Malcolm, editors. *From Genotype to Phenotype.* βios Scientific Publishers, Oxford, UK, 1994.

[66] R. Jain, M. C. Rivera, and J. A. Lake. Horizontal Gene Transfer Among Genomes: the Complexity Hypothesis. *Proceedings of the National Academy of Sciences, USA*, 96(7):3801–3806, 1999.

[67] W. James. *Some Problems of Phylosophy: A Beginning of an Introduction to Phylosophy.* University of Nebraska Press, 1996.

[68] T. Jech. *Set Theory.* Springer Verlag, Berlin, Germany, 2006.

[69] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and Centrality in Protein Networks. *Nature*, 411(6833):41–42, 2001.

[70] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From Genomics to Chemical Genomics: New Developments in KEGG. *Nucleic Acids Research*, 34(D):354–357, 2006.

[71] M.-C. King and A. C. Wilson. Evolution at Two Levels in Humans and Chimpanzees. *Science*, 188(4184):107–116, 1975.

[72] D. E. Knuth. *Selected Papers in Computer Science.* Center for the Study of Language and Information, Stanford CA, USA, 1996.

[73] D. E. Knuth. *The Art of Computer Science, Volume 3: Sorting and Searching.* Addison-Wesley, Reading MA, USA, 1998.

[74] J. Köbler, U. Schöning, and J. Torán. *The Graph Isomorphism Problem: Its Structural Complexity.* Birkhäuser, Boston MA, USA, 1999.

[75] J. O. Korbel, B. Snel, M. A. Huynen, and P. Bork. SHOT: a Web Server for the Construction of Genome Phylogenies. *Trends in Genetics*, 18(3):158–162, 2002.

[76] M. Koyutürk, Y. Kim, S. Subramaniam, W. Szpankowski, and A. Grama. Detecting Conserved Interaction Patterns in Biological Networks. *Journal of Computational Biology*, 13(7):1299–1322, 2006.

[77] T. Krell, J. R. Coggins, and A. J. Lapthorn. The Three-Dimensional Structure of Shikimate Kinase. *Journal of Molecular Biology*, 278(5):983–997, 1998.

[78] R. M. Kuffner et al. PathDB. `http://www.ncgr.org/pathdb/`.

[79] C. Kurland, B. Canback, and O. G. Berg. Horizontal Gene Transfer: a Critical View. *Proceedings of the National Academy of Sciences, USA*, 100(17):9658–9662, 2003.

[80] M. Kuroda et al. Whole Genome Sequencing of Meticillin-Resistant *Straphylococcus aureus*. *The Lancet*, 357(9264):1225–1240, 2001.

[81] D. S. Latchman. *Gene Regulation: a Eukaryotic Perspective.* Taylor & Francis, New York NY, USA, 2005.

[82] J. G. Lawrence and J. R. Roth. Selfish Operons: Horizontal Gene Transfer May Drive the Evolution of Gene Clusters. *Genetics*, 143(4):1843–1860, 1996.

[83] A. Lazcano and S. L. Miller. The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time. *Cell*, 85(6):793–798, 1996.

[84] A. Lazcano and S. L. Miller. On the origin of metabolic pathways. *Journal of Molecular Evolution*, 49(4):424–431, 1999.

[85] A. L. Lehninger, editor. *Principles of Biochemistry.* Worth Publishers, Inc., New York NY, USA, 1982.

[86] U. Leser. A Query Language for Biological Networks. *Bioinformatics*, 21(Supl. 2):ii33–ii39, 2005.

[87] M. Levy and S. L. Miller. The Stability of the RNA bases: Implications for the Origin of Life. *Proceedings of the National Academy of Sciences, USA*, 95(14):7933–7938, 1998.

[88] L. Liao, S. Kim, and J.-F. Tomb. Genome Comparisons Based on Profiles of Metabolic Pathways. In *Proc. 6th Int. Conf. Knowledge-Based Intelligent Information and Engineering Systems*, pages 469–476, 2002.

[89] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. A Tool for Multiple Sequence Alignment. *Proceedings of the National Academy of Sciences, USA*, 86(12):4412–4415, 1989.

[90] D. J. Lipman and W. Pearson. Rapid and Sensitive Protein Similarity Searches. *Science*, 227(4693):1435–1441, 1985.

[91] J. Lomax. Get Ready to GO! A Biologist's Guide to the Gene Ontology. *Briefings in Bioinformatics*, 6(3):298–304, 2005.

[92] W. Ludwig and K.-H. Schleifer. Phylogeny of *Bacteria* beyond the 16s rRNA Standard. *ASM News*, 65(11):752–757, 1999.

[93] L. Luo, F. Ji, and H. Li. Fuzzy classification of nucleotide sequences and bacterial evolution. *Bulletin of Mathematical Biology*, 57(4):527–537, 1995.

[94] H. Ma and A.-P. Zeng. Reconstruction of Metabolic Networks from Genome Data and Analysis of their Global Structure for Various Organisms. *Bioinformatics*, 19(2):270–277, 2003.

[95] W. Martin, T. Rujan, E. Richly, A. Hansen, S. Cornelsen, T. Lins, D. Leister, B. Stoebe, M. Hasegawa, and D. Penny. Evolutionary Analysis of *Arabidopsis*, Cyanobacterial, and Chloroplast Genomes reveals Plastid Phylogeny and Thousands of Cyanobacterial Genes in the Nucleus. *Proceedings of the National Academy of Sciences, USA*, 99(19):12246–12251, 2002.

[96] E. Mayr. Two Empires or Three? *Proceedings of the National Academy of Sciences, USA*, 95(17):9720–9723, 1998.

[97] P. A. McNamara, editor. *Trends in RNA Research.* Nova Science Publishers, Inc., New York NY, USA, 2006.

[98] C. D. Michener and R. R. Sokal. A Quantitative Approach to a Problem in Classification. *Evolution*, 11(2):130–162, 1957.

[99] G. Min-Oo, A. Fortin, M.-F. Tam, A. Nantel, M. M. Stevenson, and P. Gros. Pyruvate Kinase Deficiency in Mice Protects against Malaria. *Nature Genetics*, 35(4):357–362, 2003.

[100] C. Moore. *Daniel H. Burnham. Architecht, Planner of Cities.* Houghton Mifflin, Boston MA, USA, 1921.

[101] J. M. Mordeson, D. S. Malik, and S.-C. Cheng. *Fuzzy Mathematics in Medicine*, volume 55 of *Studies in Fuzziness and Soft Computing.* Springer-Verlag, New York NY, USA, 2000.

[102] E. J. Muñoz-Elías and J. D. McKinney. *M. tuberculosis* Isocitrate Lyases 1 and 2 are Jointly Required for *in vivo* Growth and Virulence. *Nature Medicine*, 11(6):638–644, 2005.

[103] E. Murphy and M. Hellerstein. Is In Vivo Nuclear Magnetic Resonance Spectroscopy Currently a Quantitative Method for Whole-Body Carbohydrate Metabolism? *Nutrition Reviews*, 58(10):304–314, 2000.

[104] J. Nathan. *Mishima: A Biography.* Da Capo Press, 2000.

[105] S. B. Needleman and C. D. Wunsch. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology*, 48(3):443–453, 1970.

[106] M. Nei. Phylogenetic Analysis in Molecular Evolutionary Genetics. *Annual Reviews of Genetics*, 30:371–403, 1996.

[107] C. B. Newgard. While Tinkering With the β-Cell...Metabolic Regulatory Mechanisms and New Therapeutic Strategies. *Diabetes*, 51(11):3141–3150, 2002.

[108] H. Ogata, W. Fujibuchi, H. Bono, S. Goto, and M. Kanehisa. Analysis of Binary Relations and Hierarchies of Enzymes in the Metabolic Pathways. *Genome Informatics*, 7:128–136, 1996.

[109] H. Ogata, W. Fujibuchi, S. Goto, and M. Kanehisa. A Heuristic Graph Comparison Algorithm and its Application to Detect Functionally Related Enzyme Clusters. *Nucleic Acids Research*, 28(20):4021–4028, 2000.

[110] R. Overbeek et al. WIT: Integrated System for High-Throughput Genome Sequence Analysis and Metabolic Reconstruction. *Nucleic Acids Research*, 28(1):123–125, 2000.

[111] E. Pennisi. Is it Time to Uproot the Tree of Life? *Science*, 284(5418):1305–1307, 1999.

[112] D. Penny and A. Poole. The Nature of the Last Universal Common Ancestor. *Current Opinion in Genetics & Development*, 9(6):672–677, 1999.

[113] T. D. Pham, D. Beck, and D. I. Crane. Fuzzy Clustering of Stochastic Models for Molecular Phylogenetics. *WSEAS Trans. Mathematics and Computers in Biology and Biomedicine*, 1(2):87–92, 2005.

[114] H. Philippe and P. Forterre. The Rooting of the Universal Tree of Life is Not Reliable. *Journal of Molecular Evolution*, 49(4):5097–523, 1999.

[115] R. Y. Pinter, O. Rokhlenko, D. Tsur, and M. Ziv-Ukelson. Approximate Labelled Subtree Homeomorphism. In *Proc. 15th Ann. Symp. of Combinatorial Pattern Matching, LNCS 3109*, volume 3109 of *Lecture Notes in Computer Science*, pages 59–73. Springer-Verlag, 2004.

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

[117] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI Reference Sequence (RefSeq): a Curated Non-Redundant Sequence Database of Genomes, Transcripts and Proteins. *Nucleic Acids Research*, 33 (Database issue):D501–D504, 2005.

[118] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg. Metabolic Pathway Analysis Web Service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, 21(7):1189–1193, 2005.

[119] M. Riddley. The Search for LUCA. *Natural History*, 11:82–85, 2000.

[120] S. C. Rison and J. M. Thornton. Pathway Evolution, Structurally Speaking. *Current Opinion in Structural Biology*, 12(3):374–382, 2002.

[121] N. Saitou and M. Nei. The Neighbor-joining Method: a New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.

[122] L. Schröder, M.-A. Weber, M. Ulrich, and J. U. Regula. Metabolic Imaging of Atrophic Muscle Tissue using Appropriate Markers in ^1H and ^{31}P NMR Spectroscopy. *Neuroradiology*, 48(11):809–816, 2006.

[123] S. Schuster, D. A. Fell, and T. Dandekar. A General Definition of Metabolic Pathways Useful for Systematic Organization and Analysis of Complex Metabolic Networks. *Nature Biotechnology*, 18(3):326–332, 2000.

[124] T. Shlomi, D. Segal, E. Ruppin, and R. Sharan. QPath: a Method for Querying Pathways in a Protein-Protein Interaction Network. *BMC Bioinformatics*, 7(199), 2006.

[125] G. Stamatoyannopoulos, S. Chen, and M. Fukui. Livel Alcohol Dehydrogenase in Japanese: High Population Frequency of Atypical Form and its Possible Role in Alcohol Sensitivity. *American Journal of Human Genetics*, 27(6):789–796, 1975.

[126] J. Stelling et al. Metabolic Network Structure Determines Key Aspects of Functionality and Regulation. *Nature*, 420(6912):190–193, 2002.

[127] M. Syvanen. Horizontal Gene Transfer: Evidence and Possible Consequences. *Annual Reviews of Genetics*, 28:237–261, 1994.

[128] R. L. Tatusov et al. The COG Database: New Developments in Phylogenetic Classification of Proteins from Complete Genomes. *Nucleic Acids Research*, 29:22–28, 2001.

[129] T. A. Titus et al. The Fanconi Anemia Gene is Conserved from Zebrafish to Human. *Gene*, 371(2):211–223, 2006.

[130] Y. Tohsato, H. Matsuda, and A. Hashimoto. A Multiple Alignment Algorithm for Metabolic Pathway Analysis using Enzyme Hierarchy. In *Proc. 8th Int. Conf. Intelligent Systems for Molecular Biology*, pages 376–383, 2000.

[131] G. Valiente. *Algorithms on Trees and Graphs.* Springer-Verlag, 2002.

[132] G. Valiente. Constrained Tree Inclusion. *Journal of Discrete Algorithms*, 3(2–4):431–447, 2005.

[133] D. Voet, J. G. Voet, and C. W. Pratt, editors. *Fundamentals of Biochemistry.* John Wiley & Sons, Inc., New York NY, USA, 1999.

[134] C. Walsh. *Antibiotics: Actions, Origins, Resistance.* ASM Press, Washington DC, USA, 2003.

[135] Z. Wang et al. Exploring Photosynthesis Evolution by Comparative Analysis of Metabolic Networks between Chloroplasts and Photosynthetic Bacteria. *BMC Genomics*, 7:100, 2006.

[136] D. J. Watts and S. H. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393(6684):440–442, 1998.

[137] E. C. Webb, editor. *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, San Diego CA, USA, 1993.

[138] D. L. Wheeler et al. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 28(1):10–14, 2000.

[139] R. H. Whittaker. New Concepts of Kingdoms or Organisms. *Science*, 163(3683):150–160, 1969.

[140] U. Wittig and A. D. Beuckelaer. Analysis and Comparison of Metabolic Pathway Databases. *Briefings in Bioinformatics*, 2(2):126–142, 2001.

[141] C. R. Woese. The Universal Ancestor. *Proceedings of the National Academy of Sciences, USA*, 95(12):6854–6859, 1998.

[142] C. R. Woese and G. E. Fox. Phylogenetic Structure of the Prokaryotic Domain: the Primary Kingdoms. *Proceedings of the National Academy of Sciences, USA*, 74(11):5088–5090, 1977.

[143] C. R. Woese, O. Kandler, and M. L. Wheelis. Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya. *Proceedings of the National Academy of Sciences, USA*, 87(12):4576–4579, 1990.

[144] S. R. y Cajal. *Los tónicos de la voluntad: reglas y consejos sobre investigación científica.* Gadir, 2006.

[145] T. Yamada, S. Goto, and M. Kanehisa. Extraction of Phylogenetic Network Modules from Prokaryote Metabolic Pathways. *Genome Informatics*, 15(1):249–258, 2004.

[146] T. Yamada, S. Goto, and M. Kanehisa. Extraction of Phylogenetic Network Modules from the Metabolic Network. *BMC Bioinformatics*, 7:130, 2006.

[147] M. Yčas. On Earlier States of the Biochemical System. *Journal of Theoretical Biology*, 44(1):145–160, 1974.

[148] L. A. Zadeh. Similarity Relations and Fuzzy Orderings. *Information Sciences*, 3(1):177–206, 1971.

[149] A. Zanella and P. Bianchi. Red Cell Pyruvate Kinase Deficiency: From Genetics to Clinical Manifestations. *Best Practice & Research Clinical Haematology*, 13(1):57–81, 2000.

[150] A. Zaslaver et al. Just-in-time Transcription Program in Metabolic Pathways. *Nature Genetics*, 36(5):486–491, 2004.

[151] K. Zhang, J. T.-L. Wang, and D. Shasha. On the Editing Distance between Undirected Acyclic Graphs. *International Journal of Foundations of Computer Science*, 7(1):43–57, 1996.

[152] O. Zhaxybayeva and J. P. Gogarten. Cladogenesis, Coalescence and the Evolution of the Three Domains of Life. *TRENDS in Genetics*, 20(4):182–187, 2004.

[153] D. Zhu and Z. S. Qin. Structural Comparison of Metabolic Pathways in Selected Single Cell Organisms. *BMC Bioinformatics*, 6:8, 2005.

Publications

[1] José C. Clemente, Kenji Satou and Gabriel Valiente. Phylogenetic Reconstruction from Non-Genomic Data. *Bioinformatics*, 23(2):e110–e115. Oxford University Press, January 2007.

[2] José C. Clemente. Pathway Alignment through Metabolic Context Similarity. *JAIST Research Report*, KS-RR-2007-001. January 2007.

[3] José C. Clemente, Kenji Satou and Gabriel Valiente. Finding Conserved and Non-Conserved Reactions Using a Metabolic Pathway Alignment Algorithm. *Genome Informatics*, 17(2):46–56. Universal Academy Press, Tokyo, Japan, December 2006.

[4] Jaume Casasnovas, José C. Clemente, Joe Miró-Julia, Francesc Rosselló, Kenji Satou and Gabriel Valiente. Fuzzy Clustering Improves Phylogenetic Relationships Reconstruction from Metabolic Pathways. In *Proc. 11th Int. Conf. Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Editions EDK, pp. 2807–2814, July 2006.

[5] José C. Clemente, Kenji Satou and Gabriel Valiente. Reconstruction of Phylogenetic Relationships from Metabolic Pathways based on the Enzyme Hierarchy and the Gene Ontology. *Genome Informatics*:16(2):45–55. Universal Academy Press, Tokyo, Japan, December 2005.

[6] Tho Hoan Pham, José C. Clemente, Kenji Satou and Tu Bao Ho. Computational Discovery of Transcriptional Regulatory Rules. *Bioinformatics*, 21 (supl. 2):ii101–ii107. Oxford University Press, September 2005.

[7] Tho Hoan Pham, José C. Clemente, Kenji Satou and Tu Bao Ho. Rule Evaluation Heuristics for Knowledge Discovery. In *Proc. Int. Workshop on Knowledge Discovery and Data Management in Biomedical Science*, pp. 29–44, May 2005.