Comparison of Metabolic Pathways by Considering Potential Fluxes
Comparison of Metabolic Pathways by Considering Potential Fluxes

Paolo Baldan¹, Paolo Besenzon², Nicoletta Cocco² and Marta Simeoni²

¹ Dipartimento di Matematica Pura e Applicata, Università di Padova
via Trieste 63, 35121 Padova, Italy
email: baldan@math.unipd.it

² Dipartimento di Informatica, Università Ca’ Foscari di Venezia,
via Torino 155, 30172 Venezia Mestre, Italy
email: cocco@dsi.unive.it, paolobese@alice.it, simeoni@dsi.unive.it

Abstract. Comparison of metabolic pathways is useful in phylogenetic analysis and for understanding metabolic functions. In this paper we first briefly survey and classify the proposals in the literature, thus observing that behavioural aspects are little considered. Then we propose a new method for comparing pathways of different organisms based on a simple similarity measure which captures both homology of reactions and functional aspects of metabolic pathways. The method requires to represent the pathways as Petri nets and to compare their T-invariant bases, which represent possible fluxes in the nets. We developed a prototype tool, CoMeta: given a set of organisms and a set of metabolic pathways, it automatically gets the corresponding data from the KEGG database, builds the corresponding Petri nets, computes the T-invariants and the similarity measures and shows the results of the comparison among organisms as a phylogenetic tree. To validate our proposal, CoMeta has been applied to different sets of organisms.

1 Introduction

The life of an organism depends on its metabolism, the chemical system which generates the essential components - amino acids, sugars, lipids and nucleic acids - and the energy necessary to synthesise and use them. Subsystems of metabolism dealing with some specific function are called metabolic pathways. Comparing metabolic pathways among different species yields interesting information on their evolution and it may help in understanding metabolic functions. This is important for metabolic engineering and for studying diseases and drugs design.

In the recent literature many proposals can be found of techniques for comparing metabolic pathways of different organisms. Each approach is based on a representation of metabolic pathways which models some information of interest, abstracting out from the others. Then it proposes a similarity or a distance measure between pathway representations and often it supplies a tool for performing the comparison.

Different representation of metabolic pathways have been proposed in the literature. The more abstract view consists in considering just a set of components
of interest in the pathways, which can be reactions, enzymes or chemical compounds. In other approaches the pathways are decomposed into a set of paths, going from an initial metabolite to a final one. The most natural and detailed representation is obtained by modelling a metabolic pathway as a graph, showing also the relations among components intervening in the pathway. Clearly, the more abstract is the view, the simpler is the comparison defined between two pathways.

Almost invariably, the distances considered in the literature focus on static, topological information of the involved networks, disregarding the fact that pathways are intended to represent dynamic processes. In this paper we propose to take into account also behavioural aspects in the definition of the distance between pathways. Specifically, we represent the pathways as Petri nets and consider, in the comparison, some aspects related to their behaviour as captured by T-invariants. Petri nets (PNs) seem to be particularly natural for representing and modelling metabolic pathways (see, e.g., [15] and references therein). The graphical representations used by biologists for metabolic pathways and the ones used in PNs are similar; the stoichiometric matrix of a metabolic pathway is analogous to the incidence matrix of a PN; the flux modes and the conservation relations for metabolites correspond to specific properties of PNs. In particular minimal (semi-positive) T-invariants correspond to elementary flux modes [58] of a metabolic pathway, i.e., minimal sets of reactions that can operate at a steady state. The space of semi-positive T-invariants has a unique basis of minimal T-invariants which is characteristic of the net and it can be used in the comparison. Hence we propose a similarity measure between pathways which considers homology of reactions, through the Sørensen index, but also potential fluxes in the pathways, through the comparison of their T-invariant bases. We developed a prototype tool, CoMeta, in order to validate and to apply our proposal.

The paper is organised as follows. Section 2 introduces metabolic pathways and it surveys the literature by illustrating and classifying the main proposals for metabolic pathway comparison. Section 3 gives a short introduction to Petri nets for representing a metabolic pathway. In Section 4 we present our proposal, namely a similarity measure between two metabolic pathways which represents both homology of reactions and functional behaviour of the nets. We briefly illustrate the tool CoMeta together with some experiments performed for validating the proposal. A short conclusion follows in Section 5.

2 Metabolic pathways and their Comparison

In this section we briefly introduce metabolic pathways and we give a short survey of the main proposals for metabolic pathway comparison in the recent literature.

2.1 Metabolic pathways

Biologists usually represent a metabolic pathway as a network of chemical reactions, catalysed by one or more enzymes, where some molecules (reactants or
substrates) are transformed into others (products). Enzymes are not consumed in a reaction, even if they are necessary and used while the reaction takes place. The product of a reaction is the substrate of the next one.

To characterise a metabolic pathway, it is necessary to identify its components (namely the reactions, enzymes, reactants and products) and their relations. Quantitative relations can be represented through a stoichiometric matrix, where rows represent molecular species and columns represent reactions. An element of the matrix, a stoichiometric coefficient \( n_{ij} \), represents the degree to which the \( i \)-th chemical species participates in the \( j \)-th reaction. By convention, reactants have negative coefficient, while products have positive ones. The kinetic of a pathway is determined by the rate associated to each reaction. It is represented by a rate equation, which depends on the concentrations of the reactants and on a reaction rate coefficient (or rate constant) which includes all the other parameters (except for concentrations) affecting the rate.

Information on metabolic pathways are collected in databases. The KEGG PATHWAY database [5] (KEGG stands for Kyoto Encyclopedia of Genes and Genomes) contains the main known metabolic, regulatory and genetic pathways for different species. It integrates genomic, chemical and systemic functional information [32]. The pathways are manually drawn, curated and continuously updated from published materials. They are represented as maps which are connected to additional information describing reactions, proteins and genes (they may be stored also in other databases). KEGG can be queried through a language based on XML [3], called KGML (KEGG Markup Language) [4]. Another important repository is the BioModels Database related to the SBML.org site [10]. It allows biologists to store, search and retrieve published mathematical models of biological interest. The models are coded in SBML (Systems Biology Markup Language) [10], a language based on XML. There are also other free access databases containing information on metabolic pathways, among others MetaCyc [6, 16], which is part of BioCyc Database Collection [1]. It is maintained by exploiting the scientific experimental literature.

2.2 Comparison techniques for metabolic pathways

Many proposals exist for comparing different organisms through the analysis of their metabolic pathways. Since a metabolic pathway is a complex system, each proposal is based on some simplified representation of its network of chemical reactions. This representation leads to the definition of a numerical score of similarity, or to a distance measure, between two pathways.

Although the literature on the subject is quite vast and a comprehensive review is out of the scope of this paper, in the following we organise the various approaches in three classes, according to the structures used for representing and comparing metabolic pathways. Such structures can be:

- **sets**, this is the simplest choice. A pathway is represented as the set of its main components, which can be reactions, enzymes or chemical compounds.
sequences, this is a more detailed representation. Sequences of reactions (enzymes, compounds) are considered, that is a pathway is decomposed into a set of selected paths from an initial component to a final one.

graphs, this is the most concrete representation. The pathway is represented as some kind of graph, which modelling the components of metabolic pathways and the relations among them.

The similarity measure (or distance) and the comparison technique defined between two pathways strictly depend on the representation choice. When a set-based representation is adopted, the comparison between the two pathways normally consists in determining the number of common elements. A typical similarity measure used in this case is the so called Jacard index. Let $X$ and $Y$ be the two sets to be compared, then the Jacard index is defined as:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

When pathways are represented by means of sequences, alignment techniques and sum of scores with gap penalty may be used as similarity measure. In the case of graph representation, more complex algorithms for graph homomorphism or graph isomorphism are used and some approximations are introduced to reduce the computational costs.

In any case the definition of a similarity measure between two metabolic pathways is based on a similarity measure between components. In particular, reactions are generally identified with the enzymes which catalyse them, and the most used similarity measures between two reactions/enzymes are based on:

- **Identity.** The simplest similarity measure is just a boolean value: two enzymes can either be identical (similarity 0) or different (similarity 1).
- **EC hierarchy.** The similarity measure is based on comparing the unique EC (Enzyme Commission) number associated to each enzyme, which represents its catalytic activity.

The EC number is a 4-level hierarchical scheme, $d_1.d_2.d_3.d_4$, that has been developed by the International Union of Biochemistry and Molecular Biology (IUBMB) [68]. For instance, arginase is numbered by EC3.5.3.1, which indicates that the enzyme is a hydrolase ($EC3.\ast.\ast.\ast$), and acts on the “carbonnitrogen bonds, other than peptide bonds” (sub-class $EC3.5.\ast.\ast$) in linear amidines (sub-sub-class $EC3.5.3.\ast$). Enzymes with similar EC classifications are functional homologues, but do not necessarily have similar amino acid sequences.

Given two enzymes $e = d_1.d_2.d_3.d_4$ and $e' = d'_1.d'_2.d'_3.d'_4$, their similarity $S(e, e')$ depends on the length of the common prefix of the corresponding codes:

$$S(e, e') = \max\{i : d_i = d'_i\}/4$$
For instance, the similarity between arginase \((e = 3.5.3.1)\) and creatinase \((e' = 3.5.3.3)\) is 0.75.

- **Information content.** The similarity measure is based on the EC numbers of enzymes together with the information content of the numbering scheme. This aims at correcting the large deviation in the distribution in the enzyme hierarchy. For example, the enzymes in the class 1.1.1 range from \(EC1.1.1.1\) to \(EC1.1.1.254\), whereas there is a single enzyme in the class 5.3.4, i.e., \(EC5.3.4.1\). Let \(C(h)\) denote the number of enzymes in an enzyme class \(h\). Then the information content of \(h\) is defined as

\[
I(h) = -\log_2 C(h).
\]

Given two enzymes \(e_i\) and \(e_j\), if \(h_{ij}\) is their lowest common upper class, then \(I(h_{ij})\) expresses the similarity between \(e_i\) and \(e_j\). In words, the larger is the number of enzymes in the smallest subtree that contains both \(e_i\) and \(e_j\), the smaller is their similarity.

- **Sequence alignment.** The similarity measure is obtained by aligning the genes or the proteins corresponding to the enzymes and by considering the resulting alignment score. Both the alignment techniques and the score can be those generally used for aligning biosequences.

Next, we discuss some specific approaches to the comparison of metabolic pathways organised along the classification outlined above. We do not consider approaches mainly intended for regulatory networks or protein interaction networks such as \([45, 33, 67, 35]\). An interesting survey on such approaches is given in \([60]\).

### 2.3 Comparison techniques based on set representation

Most of the proposals in the literature for comparing metabolic pathways abstractly represent a chemical network as a set of components. This representation is simple and efficient, which is quite useful when entire metabolic networks are compared. With this representation, comparison between metabolic pathways employs set operations.

In \([24, 25]\) the distance between two pathways, with identical topology, is defined as the weighted sum of the distances between corresponding components, i.e. substrates and enzymes, in the pathways. The distance between two corresponding components is obtained by pairwise aligning their gene sequences. The weights distinguish ortholog from paralog genes. In case of structural differences in the pathways to be compared, they can be dealt with by analysing their incidence matrices: when some functional role is missing in one pathway, a gap penalty is considered in the distance. A threshold is introduced to stop the comparison when the structural difference is too high. To illustrate the method, two electron transport pathways, the Krebs citric acid cycle and the tryptophan biosynthesis pathway have been analysed. The data are mainly taken from the WIT-system \([46]\).
In [38] the authors propose to compare metabolic networks of different organisms for studying evolution. A metabolic network is represented by a binary profile showing presence/absence of each metabolic pathway. A similarity measure between two binary profiles is computed by considering a scoring scheme for match and mismatch of bit pairs and a heuristic method to take into account the hierarchical relationships among pathways in the metabolic network. The proposal is validated by considering organisms from the three reigns, Archaea, Bacteria and Eukarya, and by deriving their phylogenetic tree from the proposed similarity measure. The data are taken from the WIT-system.

In [31] a metabolic network is divided into 64 metabolic pathways on the basis of metabolic map classifications. The j-th pathway is represented by the pathway content $p_{ij}$:

$$p_{ij} = 100 \times \frac{r_{ij}}{R_j}$$

where $r_{ij}$ is the number of reactions in the j-th pathway in i-th strain, and $R_j$ is the number of non-duplicate reactions involved in the j-th pathway collectively found in all the organisms examined. Thus multiple isozymes catalysing the same reaction are counted only once, and multifunctional enzymes are counted as many times as they catalyse different reactions. An organism is represented by a profile of its pathway contents. The Pearson correlation coefficient is used to compute the similarity measure $D(i, j)$ between two organisms represented by the profiles $p_{i1}, \ldots, p_{in}$ and $p_{j1}, \ldots, p_{jn}$:

$$D(i, j) = \frac{1}{n} \sum_{k=1}^{n} \left( \frac{p_{ik} - \mu_i}{\sigma_i} \right) \left( \frac{p_{jk} - \mu_j}{\sigma_j} \right)$$

where $\mu_i, \mu_j$ and $\sigma_i, \sigma_j$ are the average values and the standard deviation values respectively of $p_{i1}, \ldots, p_{in}$ and $p_{j1}, \ldots, p_{jn}$. A hierarchical clustering of the profiles is performed and the results are visualised as phylogenetic trees. The data on metabolic networks for the analysed organisms are taken from KEGG and from MetaCyc databases.

In [20] a cross species comparison is performed by comparing the metabolic networks of two different organisms. The authors consider two different approaches. In the first approach a metabolic network $N$ is represented as the set of its reactions and the distance between two networks $N$ and $N'$ is defined as

$$d_S(N, N') = |N \cup N'| - |N \cap N'|.$$ 

In the second approach functional properties of the network $N$ are considered such as its scope, namely the set of metabolites which can be synthesised from seed compounds, i.e., chemical substances provided as external resources. These methods are applied to the investigation of the metabolic networks of 178 species. The data sets for the experimentations are taken from the KEGG database.
In [18] a metabolic pathway is represented by the sets of its compounds, enzymes and reactions. Metabolites which are cofactors in many reactions, such as \( H_2O, ATP, NAD^+ \) etc., are discarded. The similarity relation considered between two compounds is the identity, while for enzymes three different similarity measures are considered: the hierarchical similarity (1), the information content similarity (2) and the gene ontology similarity. The gene ontology similarity is based on the minimum distance between the Gene Ontology (GO) [12] concepts representing the two enzymes. A heuristic algorithm, based on the component distances, is proposed for computing the distance between two metabolic pathways. It considers the intersection and the symmetric difference of the sets of compounds, enzymes and reactions and each non-common compound, enzyme and reaction is mapped on the most similar corresponding one in the other metabolic pathway. The proposal is validated on the \textit{glycolysis} pathway of 73 organisms, whose data are taken from KEGG, and the produced phylogeny is compared with the NCBI taxonomy [70].

In [65] metabolic networks of various species are compared by representing them as bit strings, where each bit indicates the presence/absence in the network of a specific enzymatic reaction. As a similarity measure among networks, the author uses the Tanimoto coefficient (extended Jaccard coefficient), that is the ratio between the number of common reactions in both reaction profiles and the number of reactions which are not common, i.e.,

\[
T(X, Y) = \frac{N_z}{N_x + N_y - N_z}
\]

where \( X \) and \( Y \) are two organisms, \( N_x \) and \( N_y \) are the number of 1 bits in their reaction profiles and \( N_z \) is the number of common reactions in both reaction profiles. The index \( T(X, Y) \) ranges over \([0, 1]\), and it is closer to 1 when the similarity between the two reaction profiles is higher. Multiple isozymes catalysing the same reaction are counted only once, multifunctional enzymes are counted as many times as they catalyse different reactions. The method is applied to a set of 33 representative organisms, taken from the MetaCyc database, to build a phylogenetic tree.

In [44] the Comparative Pathway Analyser (CPA) is presented, which is a web server supporting the comparative visualisation of metabolic networks of different organisms and their differential reaction content. Pathway visualisation and definition is based on KEGG’s data.

2.4 Comparison techniques based on sequence representation

Some proposals in the literature represent a metabolic pathway as a set of sequences of reactions. In this case it is necessary to have both a method for identifying the paths into the network of chemical reactions and a method to compare the selected sets of paths.
In [66] a multiple alignment based on reaction similarity is proposed for the comparative analysis of metabolic pathways. By manual pre-processing, a set of non-branching paths that exhibit reaction similarities is extracted from a set of metabolic pathways. Each path is represented by the sequence of the corresponding enzymes. Then a multiple alignment procedure is applied, based on the global pairwise alignment algorithm of Needleman and Wunsch [43] and the greedy approach of Feng and Doolittle [22]. Given \( n \) patterns, the procedure chooses all pairs of patterns, executes the corresponding alignments, and selects the pair producing the pattern with the largest information content. By removing such a pair of aligned patterns and adding the resulting one, a set of \( n-1 \) patterns is obtained. The procedure is iterated until a single pattern remains, which is the solution. If the procedure cannot obtain a set of patterns whose information content is larger than the information content of the previous set of patterns, it returns the previous set of patterns as a solution, and it stops. The multiple alignment can be used for finding common patterns among pathways or as a guide for building phylogenetic trees of the different organisms.

Enzymes are represented by their EC numbers and the information content of a pattern is based on the enzyme hierarchy, as explained below.

Let \( S = \{s_1, \ldots, s_n\} \) be a set of paths and, for \( i \in [1, n] \), let \( N(s_i) \) be the number of enzymes which appear in the path \( s_i \in S \) and \( o(e, s_i) \) the number of occurrences of the enzyme \( e \) in \( s_i \). Then the occurrence probability of \( e \) in \( S \) is defined as

\[
p(e) = \frac{\sum_{i \in [1, n]} o(e, s_i)}{\sum_{i \in [1, n]} N(s_i)}
\]

The enzyme hierarchy of EC numbers is extended with a new top class, \([*]\), which represents arbitrary enzymes. Each element of the hierarchy is called enzyme class (for example \([2.2.2.1]\), \([2.2]\) and \([*]\)).

The information content for an enzyme class \( h \), which contains \( C(h) \) enzymes, is defined as

\[
I(h) = -\log_2 C(h) - \log_2 p(h)
\]

where \( p(h) \) is the occurrence probability of \( h \) in \( S \), defined as the sum the occurrence probabilities of the enzymes in \( h \). For two enzymes \( e_i, e_j \), if their common upper class is \( h_{ij} \), then \( I(h_{ij}) \) expresses the similarity between \( e_i \) and \( e_j \).

The method is applied to the sugar, DNA and amino acid metabolic pathways extracted from the KEGG database.

In [39] metabolic information of two different species \( A \) and \( B \) are compared by considering the set of all the paths in their metabolic networks between chosen starting and ending points \( C \) and \( D \). A similarity measure between the two sets of paths is defined as

\[
\frac{|Path^A_{C-D} \cap Path^B_{C-D}|}{|Path^A_{C-D} \cup Path^B_{C-D}|}
\]

where \( Path^X_{C-D} \) is the set of all the paths from \( C \) to \( D \) in organism \( X \). The proposed method compares two paths as strings by using the exact match
with two different criteria: the first one considers only compounds in the path, the second one considers both compounds and reactions. The method is applied to *Escherichia coli*, *Saccaromyces cerevisiae* and *Salmonella typhi*, whose data are taken from the KEGG database.

In [17] PathAligner, a tool for comparative analysis of metabolic pathways is presented. The authors represent a metabolic pathway as a sequence of enzymatic reactions that describe the biochemical conversion of a given reactant into an end product. The comparison between two pathways is obtained by pairwise aligning the corresponding sequences of enzymes. The alignment algorithm does not compute the edit distance but it hierarchically determines a correspondence between enzymes in the two pathways from both ends, splitting the alignment into sub-alignments accordingly. The score of the alignment is based on a hierarchical similarity score between enzymes derived from their EC numbers.

In [36] M-PAS, a framework for identifying and ranking conserved metabolic pathways in the metabolic networks of different organisms, is proposed. Two metabolic pathways are aligned into a conserved pathway only if their individual reactions transform common substrates into common products at each step. Each such a pair of matching reactions is called a building block (BB). Six types of building blocks have been identified to allow for some variation. These building blocks are assembled into pathways of a specified length, taking reaction directions into account. Finally, the similarity score for each aligned pathway is computed. The similarity score considers all the components (substrates, products and enzymes) and it can be tuned depending on the user’s biological interests. M-PAS can find only linear pathways which are strictly similar. By combining these linear pathways it is possible to reconstruct some tree-like subnets and cycles, but not all network structures can be captured. The alignment and search can be conducted in the whole metabolic network, hence M-PAS may relate reactions in different KEGG maps which are elusive by looking only at subnetworks. The authors apply M-PAS to the comparative analysis of the metabolic pathways of *Escherichia coli* and *Saccaromyces cerevisiae* taken from the KEGG database.

### 2.5 Comparison techniques based on a graph representation

In several approaches, the network of chemical reactions is represented as a graph, encoding the network components and their relations. The drawback of a more informative representation is a greater complexity in the comparison. In fact the graph and subgraph isomorphism problems are GI-complete (graph isomorphism complete) and NP-complete, respectively. For this reason efficient heuristics are needed to solve these problems in a reasonable time and simplifying assumptions are introduced which produce further approximations.

In [29] a metabolic pathway is represented as an enzyme graph, which is the graph resulting from a pathway by removing metabolites and substrates and keeping only enzymes/reactions in the pathway and their connections. For
the comparison, the authors combine a similarity measure between enzymes with a similarity measure between enzyme graphs. The similarity score between enzymes can be a similarity score between the DNA sequences of their genes, the sequences of their proteins, or their EC numbers. The similarity score between two graphs $G_1$ and $G_2$ is computed in four steps:

- the similarity score between every pair of nodes $(a, b)$ is computed, where $a \in G_1$ and $b \in G_2$;
- a bipartite graph is built by using the similarity scores and a maximal weight matching of this bipartite graph is determined;
- a similarity measure between every pair of matched nodes is recomputed;
- the final score is computed by summing up the similarity of the matched nodes and by normalising this sum.

The method is implemented and applied to the *citric acid cycle* and the *glycolysis* and to the *carbohydrate and lipid metabolic pathways* of different organisms to obtain phylogenies. The data sets are taken from the KEGG database.

In [47] MetaPathHunter, a tool for pathways homology search, is presented. Given a pathway query and a collection of pathways, the tool returns all the pathways in the collection which are similar to the query, and it computes for each of them a similarity score and a measure of statistical significance. Pathways are represented as graphs, where the nodes correspond to enzymes and an edge connects two nodes if for the corresponding enzymes the product of one is a substrate of the other. In order to master the complexity, only tree-shaped graphs are considered. This choice allows one to deal with the parts of a pathway which supply certain products. The search for similar trees is based on subtree homeomorphism, since a single reaction in one pathway may replace a few consecutive reactions in another pathway. The similarity score among trees combines topological and enzyme similarities. Given two labeled homeomorphic trees $T_1$ and $T_2$, the similarity score is

$$
\delta(|T_2| - |T_1|) + \sum_{(u,v) \in M} \Delta(u, v)
$$

where $\delta$ denotes a penalty for deleting a node from a tree and $M$ is the node mapping in the homeomorphism. $\Delta(u, v)$ is the distance between corresponding nodes, based on functional homology of enzymes, i.e. enzymes $u$ and $v$ are represented by their EC number and their distance is determined by the information content $I(h_{uv})$ of their common upper class, as in [66]. The tool is applied to the metabolic networks of *Escherichia coli* and *Saccaromyces cerevisiae* for studying their similarity and differences and the data are taken from the EcoCyc[2] database.

In [23] the authors represent metabolic networks as directed hypergraphs, where nodes are metabolites and hyperedges are enzymes/reactions. They introduce union, intersection and difference operations to form an algebra of such
hypergraphs and propose the symmetric difference as a distance measure. They implement their method and apply it to procaryotes to obtain phylogenies. The claim is that their method identifies subnetworks which correspond to innovations. The data sets are taken from the KEGG database.

In [69] MetaPath is proposed as a tool for homology search: given a pattern network, it searches for similar subnetworks in a given host network. A metabolic pathway is represented as a graph, where nodes are metabolites and arcs are enzymes/reactions. The algorithm, which determines if two subgraphs are homeomorphic, is based on an exhaustive search for similar paths and it is incremental. To improve its efficiency, a local diversity threshold is introduced in order to cut most of the extensions of the paths to be explored. Local diversity is based on gaps and enzymes similarity (enzyme similarity is based on the EC numbers and it consists of the information content of their common upper class, as in [66]). The tool is applied to the metabolic pathways of five different organisms, namely Bacillus subtilis, E.coli, Homo sapiens, Saccharomyces cerevisiae and Thermus thermophilus, taken from the BioCyc database. The metabolites H2O, ATP and ADP are excluded from all pathways. The authors claim that the tool can be used for pathways comparison and enzyme classification.

In [37] an efficient algorithm for aligning two molecular networks is proposed. The algorithm, based on integer quadratic programming, identifies similar parts in the networks maximising a similarity score. The score depends both on similarity of pair of nodes (biomolecules) and on similarity between pair of edges (interactions between the molecules) with a parameter which can move the emphasis on node or edge matching score. In particular a metabolic pathway is represented as a directed graph, nodes correspond to enzymes and edges connect two enzymes if the product of one is the substrate of the other. The similarity measure considered between two enzymes is the one given by the EC hierarchy. The method is applied to compare all pairs of pathways of the organisms Escherichia coli (113 pathways) and Saccharomyces cerevisiae (151 pathways), as in [47]. The data are extracted from EcoCyc and SGD [9].

In [14] a method is proposed to align two metabolic pathways, represented as directed graphs whose nodes are reactions, compounds and enzymes. The method aligns reactions, compounds and enzymes, by solving three separate, but consistent, eigenvalue problems through power method. Beside functional homology of the pathways, the method considers also their topology, by taking into account the reachability sets of the aligned entities. The extracted mappings of entities is reported as an alignment together with a similarity score between the aligned pathways. The similarity score between two pathways $P$ and $P'$ is defined on the basis of a mapping $\phi = [\phi_R, \phi_C, \phi_E]$ between reactions, compounds and enzymes of $P$ and $P'$, as

$$\text{Sim}_{P_\phi}(P, P') = \frac{\beta}{|\phi_C|} \sum_{(C_i, C'_j) \in \phi_C} \text{Sim}_C(C_i, C'_j) + \frac{(1-\beta)}{|\phi_E|} \sum_{(E_i, E'_j) \in \phi_E} \text{Sim}_E(E_i, E'_j)$$
where $|\phi_C|$ and $|\phi_E|$ denote the cardinality of the corresponding mappings and $\beta \in [0, 1]$ is a parameter to adjust the relative influence of compounds and enzymes on the alignment score. Note that $\phi_R$ is extracted first, then $\phi_C$ and $\phi_E$ have to be consistent with it. For $SimC$ there are two possibilities: it can be either a boolean value indicating the equality of compounds, or the SIMCOMP compound similarity score [27]. Also $SimE$ has two possible definitions, either based on the EC number hierarchy or on the information content. Finally $SimP_\phi$ ranges in $[0, 1]$ and a bigger score implies a better alignment between pathways. The unexpectedness of the resulting alignment is given by calculating its z-score.

The method is applied to pathways taken from the KEGG database to determine functionally similar entities in different organisms, such as alternative enzymes or alternative paths.

In [41] Rahnuma, a tool for the analysis and comparison of metabolic networks is proposed. A metabolic network is represented as a hypergraph, where nodes are compounds and hyperhedges are the reactions connecting the compounds. Rahnuma computes all the paths between individual metabolites or a group of metabolites using depth first traversal of the hypergraph representing the metabolic network. Rahnuma translates KEGG data into hypergraphs, and supplies three main modules, namely Network Analysis, Pathway Analysis and Comparative Analysis. The user can also specify a list of metabolites to be ignored, such as the ubiquitous metabolites $O_2$, $H_2O$ and $CO_2$.

3 Representing metabolic pathways with Petri nets

PNs are a well known formalism introduced in computer science for modelling (discrete) concurrent systems. PNs have a sound theory and many applications both in computer science and in real life systems (see [42] and [21] for surveys on PNs and their properties). A PN model can be decomposed to master complexity and it enables a large number of different analyses on a network. In some seminal papers Reddy et al. [52, 50, 51] and Hofestädt [30] proposed Petri nets (PNs) for representing and analysing metabolic pathways. Since then, a wide range of literature has grown on the topic [15]. In this section we give a short overview of the basic model of Petri net used for the representation of a metabolic pathway.

3.1 Basic Petri nets

A (finite marked) Petri net (PN) is a tuple $N = (P, T, W, M_0)$ where:

- $P = \{p_1, \ldots, p_n\}$ is the set of places;
- $T = \{t_1, \ldots, t_m\}$ is the set of transitions;
- $W : ((P \times T) \cup (T \times P)) \rightarrow \mathbb{N}$ is the weight function; \(^1\)
  when $W(x, y) = k$, $k > 0$, the net includes an arc from $x$ to $y$ with weight $k$;
- $M_0$ is an $n$-dimensional vector of non-negative integers which represents the initial marking of the net.

\(^1\) We denote by $\mathbb{N}$ the set of natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$
PNs admit a simple graphical representation, where places are drawn as circles, transitions as rectangles, the presence of an arc \((p, t)\) or \((t, p)\) between a transition \(t\) and a place \(p\), is represented by a directed arc, labelled with the corresponding weight \(W(p, t)\) or \(W(t, p)\), respectively. When the weight is 1 the label is usually omitted. A marking \(M\) (or state) is an \(n\)-dimensional vector of non-negative integers which represents the amount of tokens in the places of the net \(N\). \(M\) is graphically represented by inserting in each place \(p_i\) a corresponding number \(m_i\) of black circles representing tokens. A simple example of PN is given in Fig. 1.

Fig. 1. A simple Petri net.

The input bag or precondition of the transition \(t\) is the \(n\)-dimensional vector of non-negative integers \(\bullet t = (i_1, \ldots, i_n)\), where \(i_j = W(p_j, t)\) for any \(j \in \{1, \ldots, n\}\). Dually, the output bag or post-condition of the transition \(t\) is an \(n\)-dimensional vector of non-negative integers \(t^\bullet = (o_1, \ldots, o_n)\), where \(o_j = W(t, p_j)\) for any \(j \in \{1, \ldots, n\}\). Intuitively, \(\bullet t\) indicates, for any place of the net, the number of tokens needed to enable transition \(t\). The firing of \(t\) removes such tokens and generates new ones, as indicated by \(t^\bullet\). For instance, in Fig. 1, we have \(\bullet t_1 = (1, 0, 0, 0, 0)\) while \(t_1^\bullet = (0, 1, 1, 0, 0)\). Similarly, \(\bullet t_5 = (0, 0, 0, 2, 1)\) and \(t_5^\bullet = (1, 0, 0, 0, 0)\).

A transition \(t\) with input bag \(\bullet t\) is enabled by the marking \(M = (m_1, \ldots, m_n)\), if \(\bullet t \leq M\). In this case \(t\) can fire and, as a consequence, the net marking changes from \(M\) to a new marking \(M'\) defined as follows:

\[
M' = M - \bullet t + t^\bullet,
\]

and we write \(M \xrightarrow{t} M'\).

Note that a marking can enable more than one transition. In this situation one of the enabled transitions is non-deterministically chosen and fired. It can also happen that some of the enabled transitions compete for a token; in this case
they are in conflict and the execution of one of them can prevent the other ones from firing. For instance, consider the PN in Fig. 1. Its marking \( M = (1, 0, 0, 0, 0) \) enables transitions \( t_1 \) and \( t_2 \), which are in conflict. After firing e.g., transition \( t_1 \), the new marking is \( M' = (0, 1, 1, 0, 0) \) and \( t_2 \) is no longer enabled.

The incidence matrix of a PN \( N \), denoted by \( A_N \), is the \( n \times m \) matrix which has a row for each place, and a column for each transition. The column associated to transition \( t \) is the vector \( t^T - \star t^T \), which represents the marking change due to the firing of \( t \).

The reachability set (or state space) of the net \( N \) is the set of all the markings of the net which are reachable by a firing sequence from the initial marking \( M_0 \), and it is denoted by \( R(N, M_0) \).

A common kind of structural analysis based on the incidence matrix aims to determine the so-called invariants of the net.

In this paper we are interested in T-invariants (transition invariant) of a net \( N \), which are \( m \)-dimensional vectors in which each component represents the number of times that a transition should fire to take the net from a state \( M \) back to \( M \) itself.

**Definition 1 (T-invariant).** A semi-positive T-invariant, in the following simply T-invariant, of a net \( N \) is defined as a solution of the following equation:

\[
A_N \cdot X = 0, \quad \text{where} \quad X = (x_1, \ldots, x_m)^T \quad \text{and} \quad x_i \in \mathbb{N}, \quad \text{for} \quad i \in \{1, \ldots, m\}.
\]

A T-invariant \( X \neq 0 \) indicates that the system can cycle on a state \( M \) enabling the cycle. The presence of T-invariants in a PN model of a metabolic pathway is biologically of great interest as it can reveal the presence of steady states, in which concentrations of substances have reached a possibly dynamic equilibrium.

**Definition 2 (support).** The support of a T-invariant \( X = (x_1, \ldots, x_m)^T \) is the set of transitions corresponding to non-zero coefficients \( \text{supp}(X) = \{t_i \mid x_i > 0\} \). The support of a T-invariant is minimal if there is no other T-invariant \( X' \) whose support is strictly included in that of \( X \), i.e., such that \( \text{supp}(X') \subset \text{supp}(X) \).

For any minimal support there exists a unique T-invariant \( X \) with that support and such that for any other T-invariant \( X' \), \( X' \leq X \) implies \( X = X' \).

**Definition 3 (minimal T-invariants).** The T-invariants with minimal support are called minimal support T-invariants, or simply minimal T-invariants. The set of minimal T-invariants of a net \( N \) is denoted by \( B(N) \).

In a PN model of a metabolic pathway, a minimal T-invariant corresponds to an elementary flux mode, a term introduced in [58] to refer to a minimal set of reactions that can operate at a steady state. It can be interpreted as a minimal self-sufficient subsystem which is associated to a function. Minimal T-invariants are important in model validation techniques (see, e.g., [28, 34]) and they may provide insights into the network behaviour.
The set of invariants, although clearly infinite, can be characterised finitely in a nice way, by resorting to its Hilbert basis [63].

**Proposition 4 (Unique basis).** The set of T-invariants of a net $N$ admits a unique basis which is given by the set $B(N)$ of minimal T-invariants, i.e., any T-invariant can be obtained as a linear combination of minimal T-invariants.

The set of minimal T-invariants of a net, $B(N)$, thus represents a characteristic of $N$ related to its dynamics, which can be used in the definition of a distance.

A large number of tools have been developed for analysing properties of PNs, including T-invariants analysis. A quite comprehensive list can be found at the Petri net World site [7].

### 3.2 Petri net representation of a metabolic pathway

The structural representation of a metabolic pathway by means of a PN can be derived by exploiting the natural correspondence between PNs and biochemical networks. In fact places are associated with molecular species, such as metabolites, proteins or enzymes; transitions correspond to chemical reactions; input places represent the substrate or reactants; output places represent reaction products. The incidence matrix of the PN is identical to the stoichiometric matrix of the system of chemical reactions. As a classical example (see, e.g., [42]), reaction $2H + O \rightarrow H_2O$ will lead to a PN with three places, corresponding to the substances $H$, $O$ and $H_2O$, and one transition, which consumes two tokens from place $H$ and one from place $O$ and produces one token in place $H_2O$. The number of tokens in each place indicates the amount of substance associated with that place. It may represent either the number of molecules expressed in moles or the level of concentration, suitably discretised by introducing a concept of concentration level [26]. Once we have a qualitative model, quantitative data can be added to refine the representation of the behaviour of the pathway. In particular, extended PNs may have an associated transition rate which depends on the kinetic law of the corresponding reaction.

![Fig. 2. Two biochemical reactions in the KEGG glycolysis pathway.](image)

As an example, consider the two reactions in Fig. 2, which appear in the glycolysis pathway, as given in KEGG database. By clicking on enzyme 3.1.3.11
in the KEGG map, it is possible to see the reaction associated with the enzyme, namely:

\[ \text{D-fructose 1,6-bisphosphate} + \text{H}_2\text{O} = \text{D-fructose 6-phosphate} + \text{phosphate} \]

where the substrates are D-fructose 1,6–bisphosphate and water, and the products are D-fructose 6-phosphate and phosphate. Note that KEGG always uses the equal sign in reaction formulae even though the reaction is irreversible. The direction of a reaction is indicated by an arrow in the KEGG diagram.

The other reaction in the same figure, catalysed by the enzyme 2.7.1.11, is:

\[ \text{ATP} + \text{D-fructose 6-phosphate} = \text{ADP} + \text{D-fructose 1,6-bisphosphate} \]

where the substrates are ATP and D-fructose 6-phosphate, and the products are ADP and D-fructose 1,6-bisphosphate. Also this reaction is irreversible.

If we represent each component of the substrate, each enzyme and each product of the reaction as a place of a Petri net, and chemical reactions by transitions, we obtain the PN of Fig. 3(a).

![Fig. 3. PN associated with the biochemical reactions of Fig. 2.](image)

Observe that in Fig. 3(a), the fact that an enzyme is taken and released by the corresponding reaction is represented graphically by connecting the enzyme place to the transition corresponding to the associated reaction with a double arrow. Often enzyme places are omitted from the representation, which is appropriate as long as their concentrations do not change.

Note that in the KEGG map of Fig. 2, substances such as \( \text{H}_2\text{O} \), phosphate, ADP and ATP are not shown. They are ubiquitous molecules and their concentrations, which are assumed to be constant, are taken into account by the constants of the reaction rates.

By omitting places corresponding to enzymes and ubiquitous substances the net of Fig. 3(a) can be simplified, thus resulting in the PN of Fig. 3(b). In
general, large and complex networks can be greatly simplified by avoiding an explicit representation of enzymes and by assuming that ubiquitous substances are in a constant amount. On the other hand, as an obvious drawback, processes involving these substances, such as the energy balance, are not modelled.

We developed a tool, PANGEA [48, 62, 19], to automatically translate a metabolic pathway into its Petri net representation. This permits to apply the analysis techniques which are available for PNs also to metabolic pathways. PANGEA is now in a prototype version which allows one to get the biological data about a metabolic pathway from different databases (KEGG and BioModels databases) with different formats (KGML and SBML) and to automatically translate them into a PN representation in the formats accepted by a set of analysis tools for PNs (PIPE2 [8], Snoopy [11], INA [61] and TimeNET [13]).

4 Considering behavioural aspects in metabolic pathways comparison

In this section we define a similarity measure between two metabolic pathways which takes into account the flows in the pathways, by comparing the minimal T-invariants of their Petri net representations. Such measure is combined with more standard similarity measure which considers homology of reactions. We also briefly illustrate a tool which implements our proposal and we report on a few experiments for its validation.

4.1 Similarity measure between pathways based on T-invariants

Metabolic pathways are complex networks of biochemical reactions describing fluxes of substances. Such fluxes can be characterised by decomposing them into elementary fluxes, that is cyclic fluxes which cannot be further decomposed. Most of the techniques for pathways comparison, briefly illustrated in Section 2, compare pathways on the basis of homology of their reactions, that is they determine a point to point functional correspondence. Some proposals consider also the topology of the network, but still most techniques we are aware of are eminently static, i.e., they disregard the fact that the network is intended to represent a dynamic system and they ignore, e.g., the flow of metabolites in the pathway.

The analysis of fluxes is essential in studying a pathway behaviour and it is worthy of note that such analysis can be conducted by means of structural analysis methods, not requiring kinetic information which are generally difficult to determine. By assuming both the fluxes and the pool sizes constants, with some further simplifying assumption, the stoichiometry of the network restricts the space of all possible net fluxes to a rather small linear subspace. Such subspace can be analysed in order to capture possible behaviours of the pathway and its functional subunits [54–59].

We propose a comparison technique for pathways which combines homology of enzymes/reactions with a measure of the similarity of flows in the pathways.
More precisely, we model metabolic pathways as Petri nets and, in the comparison, we consider the similarity of the corresponding T-invariants, which correspond to fluxes in the pathways. Hence our similarity score has two components: one for taking into account the homology between the reactions in the two pathways, the other for the similarities among their elementary fluxes as captured by the minimal T-invariants.

Let \( P_1 \) and \( P_2 \) be two metabolic pathways. In principle, any of the distances discussed in the previous section can be “refined” by taking into account the T-invariants. For simplicity, in this paper we stick to a distance between pathways based on their representation as sets of reactions/enzymes, as in the proposals illustrated in Section 2.3. Since the same reaction may occur more than once in a pathway, we actually consider a multiset representation. As a similarity index on multisets we choose the Sørensen index \([64]\), which extends the Jacard index to multisets:

\[
S\_\text{score}(X_1, X_2) = \frac{2|X_1 \cap X_2|}{|X_1| + |X_2|}
\]

where \( X_1 \) and \( X_2 \) are two multisets, \( \cap \) is the intersection and \( |\cdot| \) the cardinality extended in the obvious way to multisets.

Then, if \( R_1 \) and \( R_2 \) are the multisets of reactions of pathways \( P_1 \) and \( P_2 \), the homology of reactions the two pathways is taken into account by considering a distance based on the Sørensen index:

\[
d_S(P_1, P_2) = 1 - S\_\text{score}(R_1, R_2)
\]

Reactions are represented by their EC numbers and the similarity considered between them is just the identity.

Behavioural aspects of the pathway are included in the distance by focusing on the bases of minimal T-invariants \( \mathcal{B}(P_i) \) of the two metabolic pathways. Each invariant is seen as a multiset of reactions and the similarity index between two invariants is computed, as before, by the Sørensen index. Then the contribution to the distance provided by the T-invariants is computed by performing a sort of best match between the two bases \( \mathcal{B}(P_1) \) and \( \mathcal{B}(P_2) \). More precisely we apply the procedure in Fig. 4.

As before, the similarity index induces a distance between two metabolic pathways \( P_1 \) and \( P_2 \) based on their minimal T-invariants:

\[
d_I(P_1, P_2) = 1 - I\_\text{score}(P_1, P_2)
\]

The two distances based on homology of reactions and on the comparison of the T-invariants are then combined into a weighted sum:

\[
d_D(P_1, P_2) = \alpha d_S(P_1, P_2) + \beta d_I(P_1, P_2)
\]

where \( \alpha + \beta = 1 \). The weights \( \alpha \) and \( \beta \) allow the analyst to move the focus between homology of reactions and similarity of functional components as represented by the T-invariants.
function I\_SCORE(P\_1, P\_2);
    input: two metabolic pathways P\_1 and P\_2;
    output: the similarity measure between B(P\_1) and B(P\_2);
begin
    I\_1 = B(P\_1); I\_2 = B(P\_2);
    score = 0;
    card = max\{|I\_1|, |I\_2|\};
    while (I\_1 \neq \emptyset \land I\_2 \neq \emptyset) do
        begin
            (X\_1, X\_2) = FIND\_MAX\_SIM(I\_1, I\_2);
            \{Find a pair in I\_1 \times I\_2 with maximum similarity\}
            score = score + S\_score(X\_1, X\_2); \{S\_score is the Sørensen index\}
            I\_1 = I\_1 - \{X\_1\};
            I\_2 = I\_2 - \{X\_2\};
        end;
    score = score / card;
end Compute\_I\_SCORE;

Fig. 4. Comparing bases of T-invariants

Two organisms O\_1 and O\_2 can be compared, more generally, by considering their similarity on \(n\) chosen metabolic pathways \(P\_1, \ldots, P\_n\). In this case the distances between the two organisms with respect to the various metabolic pathways \(P\_j, j \in [1, n]\), need to be combined. The simplest solution, currently adopted in the tool, consists in taking the average distance (although refinements based on weights could be appropriate to allow the user to put more emphasis on some pathways of interest):

\[
d_D(O\_1, O\_2) = \frac{\sum_1^n d_D(P\_1^1, P\_2^2)}{n}
\]

4.2 The Tool

We realised a tool, CoMETA (COmparing METAbolic pathways), for validating our proposal. For a specified set of organisms and subset of metabolic pathways, CoMETA computes a distance between pairs of organisms by comparing the chosen pathways. The distance is computed as discussed in the previous section, by combining a similarity score based on homology of reactions and one based on T-invariants. The present prototype implementation of CoMETA computes a homology score corresponding to the Sørensen index on multisets of reactions, as defined in Section 4.1, but due to its modular structure, other homology scores can be easily included.

19
CoMeta is user-friendly: it has a graphical user interface; it automatically gets the data from KEGG database, it builds and analyses the corresponding Petri nets by making use of suitable tools and it shows the results of the comparison of different organisms as a phylogenetic tree built either with the UPGMA method [40] or with the Neighbour Joining [53] method.

CoMeta offers a set of integrated functionalities:

- it automatically gets the information on the specified organisms and metabolic pathways from the KEGG database [5];
- it builds the corresponding Petri nets by using the tool PANGEA [49, 62, 19] for translating metabolic data into a Petri net model;
- it computes the basis of semi-positive T-invariants by using the tool INA [61] for Petri nets analysis;
- for each pair of organisms, $O_i, O_j$, it computes $S_{score}$, the Sørensen index on multisets of reactions, and the $I_{score}$;
- for each pair of organisms and for each similarity score, it computes a distance table:
  \[d_S(O_i, O_j) = 1 - S_{score}(O_i, O_j),\]
  \[d_I(O_i, O_j) = 1 - I_{score}(O_i, O_j);\]
and a combined distance table, allowing the user to specify the weights $\alpha$ and $\beta$:
  \[d_D(O_i, O_j) = \alpha d_S(O_i, O_j) + \beta d_I(O_i, O_j);\]
- it builds a phylogenetic tree corresponding to such a table, with the phylogenetic method specified by the user (either the UPGMA method or the Neighbour Joining method), in order to visualise the distances among the various organisms.

CoMeta has been developed in Java, version 6.20, in order to be compatible with PANGEA and INA and it uses Java components (e.g. JPanel, JFrame, JOptionPane) for the graphical interface.

CoMeta offers some useful options to analyse the comparison results:

- the possibility to save organisms and metabolic pathways for further processing;
- the possibility to visualise multisets of reactions and T-invariant bases associated to metabolic pathways;
- the possibility to visualise the scores, the distances and the associations between T-invariants.

### 4.3 Validation

CoMeta has been applied to different sets of organisms in order to validate our proposal.
<table>
<thead>
<tr>
<th>Cod.</th>
<th>Organism</th>
<th>Reign</th>
</tr>
</thead>
<tbody>
<tr>
<td>HSA</td>
<td>Homo sapiens</td>
<td>Animals</td>
</tr>
<tr>
<td>MCC</td>
<td>Macaca mulatta</td>
<td>Animals</td>
</tr>
<tr>
<td>GGA</td>
<td>Gallus gallus</td>
<td>Animals</td>
</tr>
<tr>
<td>SBI</td>
<td>Sorghum bicolor</td>
<td>Plants</td>
</tr>
<tr>
<td>ANI</td>
<td>Aspergillus nidulans</td>
<td>Fungi</td>
</tr>
<tr>
<td>AFU</td>
<td>Archaeoglobus fulgidus</td>
<td>Archea</td>
</tr>
<tr>
<td>PHO</td>
<td>Pyrococcus horikoshii</td>
<td>Archea</td>
</tr>
<tr>
<td>TGA</td>
<td>Thermococcus gammatolerans</td>
<td>Archea</td>
</tr>
<tr>
<td>MGE</td>
<td>Mycoplasma genitalium</td>
<td>Bacteria</td>
</tr>
<tr>
<td>MPN</td>
<td>Mycoplasma pneumoniae</td>
<td>Bacteria</td>
</tr>
<tr>
<td>HIN</td>
<td>Haemophilus influenzae Rd KW20</td>
<td>Bacteria</td>
</tr>
</tbody>
</table>

Fig. 5. Table of organisms for experiment 1

**Experiment 1** The first experiment considers the *glycolysis* pathway of 11 different organisms, distributed among the kingdoms of Animals, Plants, Fungi, Archaea and Bacteria.

The similarity measure combines the Sørensen index on multiset of reactions $d_S$, with weight $\alpha = 0.75$, and the invariant score $d_I$, with weight $\beta = 0.25$. CoMETa correctly classifies the organisms as shown by the phylogenetic tree in Figure 5 which is built with the UPGMA method.

**Experiment 2** In a second experiment we consider the *glycolysis* pathway in 25 different organisms belonging to Animals, Bacteria and Archaea. The Animals belong to the phyla Vertebrates and Arthropoda. The Bacteria belong to the phyla Gammaproteobacteria, Cyanobacteria, Alphaproteobacteria, Betaproteobacteria and Tenericutes. The Archaea belong to the phyla Euryarchaeota e Crenarchaeota.

The similarity measure combines the Sørensen index on multiset of reactions, with weight $\alpha = 0.75$, and the invariant score, with weight $\beta = 0.25$. CoMETa correctly classifies the organisms as shown by the phylogenetic tree in Figure 7 which is built with the UPGMA method.

**Experiment 3** In a third experiment we consider the *glycolysis* and the *Krebs Cycle* (TCA) pathways in 6 different animals.

In order to show the influence of the T-invariants comparison, the weight $\alpha$ takes the values \{0.00, 0.25, 0.50, 0.75, 1.00\} in the combined distance. The effect of varying the weight of T-invariants is shown by the phylogenetic trees in Figures 9, 10, 11 and 12, which are built with the UPGMA method.

The first tree, in Figure 9, corresponds to consider only homology of reactions in the comparison. In this case *Homo sapiens, Macaca mulatta, Pan troglodytes* and *Equus caballus* are closely classified.
The second tree, in Figure 10, corresponds to consider homology of reactions and T-invariants with the same weight. The classification improves with respect to the previous one, since *Homo sapiens*, *Macaca mulatta* and *Pan troglodytes* are still closely classified, but *Equus caballus* is now rather distinct from them. On the other hand *Sus scrofa* and *Gallus gallus* are closely related, which somehow surprising.
When considering only T-invariants, the classification hardly distinguishes among very similar organisms such as *Homo sapiens*, *Macaca mulatta* and *Pan troglodytes*, see Figure 11.

The best classification is obtained in Figure 12, with $\alpha = 0.25$, which corresponds to consider mainly homology of reactions and a minor contribution of T-invariants.

This experiment in particular shows that the combined distance we propose produces valid phylogenetic classifications and that the refinement obtained by considering T-invariants can be useful. It also shows that the distance based on T-invariants alone may be misleading. This can be understood observing that two pathways may have similar invariants but significantly differ in the other reactions. Thus homology of reactions seems to be a component which is complementary and necessary to be considered in the comparison.
5 Conclusions

Biological questions related to evolution and to differences among organisms can be answered by comparing their metabolic pathways. In this paper, after providing a brief survey of the main proposals in the literature for comparing metabolic pathways, we have proposed a new similarity measure which combines homology of reactions and behavioural aspects of metabolic pathways. Such a measure considers potential fluxes in a pathway, which correspond to the minimal $T$-invariants of the Petri net representation of a pathway. Our similarity measure
between two metabolic pathways is defined on the Hilbert basis of semi-positive T-invariants of their Petri net representations and it can be usefully combined with other similarity measures representing homology of reactions. In particular we choose a simple similarity measure, given by the Sørensen index, based on identity of reactions. We implemented a user friendly tool, CoMETA, to experiment with this combined similarity measure. The user chooses a set of organisms
and a set of metabolic pathways, CoMETA automatically gets the corresponding data from the KEGG database, it builds the corresponding Petri nets and computes the T-invariant bases, it computes the homology and the behavioural similarity measures and combines them as required by the user. CoMETA shows the results of the comparison among organisms as a phylogenetic tree built either with the UPGMA method or with the Neighbour Joining method.

We made some experiments with CoMETA showing that:

− the combined measure we propose produces valid phylogenetic classifications;
− the refinement due to the introduction of the behavioural measure is useful;
− homology of reactions seems to be a property which is complementary and necessary to be considered in the comparison.

We expect that considering measures based on more sophisticated representations of a pathway (e.g., graphs rather than sets) would not add much to our combined measure, since some structural aspects are somehow captured by behavioural similarity. However this hypothesis needs further experiments to be
verified. As a future development we plan to extend CoMETA to deal with other similarity measures in the literature. This would allow us both to verify the previous hypothesis and to compare and evaluate the different existing proposals.

References


