Comparison of Metabolic Pathways by Considering Potential Fluxes

Paolo Baldan[1], Nicoletta Cocco[2] and Marta Simeoni[2]

[1] Dipartimento di Matematica, Università di Padova
via Trieste 63, 35121 Padova, Italy
email: baldan@math.unipd.it
[2]Dipartimento di Scienze Ambientali, Statistica e Informatica,
Università Ca' Foscari Venezia,
via Torino 155, 30172 Venezia Mestre, Italy
email: cocco@dsi.unive.it,simeoni@dsi.unive.it

Abstract. Comparison of metabolic pathways is useful in phylogenetic analysis and for understanding metabolic functions when studying diseases and in drugs engineering. In the literature many techniques have been proposed to compare metabolic pathways, but most of them focus on structural aspects, while behavioural or functional aspects are generally not considered. In this paper we propose a new method for comparing metabolic pathways of different organisms based on a similarity measure which considers both homology of reactions and functional aspects of the pathways. The latter are captured by relying on a Petri net representation of the pathways and comparing the corresponding T-invariant bases, which represent potential fluxes in the nets. The approach is implemented in a prototype tool, CoMeta, which allows us to test and validate our proposal. Some experiments with CoMeta are presented.

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 of 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 techniques have been proposed for comparing metabolic pathways of different organisms. Each approach chooses a representation of metabolic pathways which models the information of interest, proposes a similarity or a distance measure and possibly supplies a tool for performing the comparison.

Representations of metabolic pathways at different degrees of abstraction have been considered. A pathway can be simply viewed as a *set* of components of interest, which can be reactions, enzymes or chemical compounds. In other

approaches pathways are decomposed into a set of paths, leading from an initial metabolite to a final one. The most detailed representations model a metabolic pathway as a graph. Clearly, more detailed models produce more accurate comparison results, in general at the price of being more complex.

The distances in the literature generally focus on static, topological information of the pathways, disregarding the fact that they represent dynamic processes. In this paper we propose to take into account also behavioural aspects: we represent the pathways as Petri nets (PNs) and compare also aspects related to their behaviour as captured by T-invariants. Petri nets seem to be particularly natural for representing and modelling metabolic pathways (see, e.g., [8] 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 [43] 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 we use it in the comparison. Hence we propose a similarity measure between pathways which considers both homology of reactions, represented by the Sørensen index on the multisets of enzymes in the pathways, and similarity of potential fluxes in the pathways, obtained by comparing the corresponding T-invariant bases. We developed a prototype tool, CoMeta, implementing our proposal. Given a set of organisms and a set of metabolic pathways, CoMeta 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. We performed several experiments with CoMeta and, although further investigations are definitively needed, the approach appears to be promising and worth to be pursued.

The paper is organised as follows. In Section 2 we introduce metabolic pathways and give a classification of various proposals for metabolic pathways comparison. In Section 3 we show how a Petri net can model a metabolic pathway and present our proposal. In Section 4 we briefly illustrate the tool CoMeta and we present some experiments. A short conclusion follows in Section 5.

2 Comparison of Metabolic Pathways

In this section we briefly introduce metabolic pathways and classify various proposals for the comparison of metabolic pathways in the 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 for other ones.

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, the coefficients for reactants are negative, while those for products are positive. 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. In particular the *KEGG PATHWAY* database [2] (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 [23]. The pathways are manually drawn, curated and continuously updated from published materials. They are represented as maps which are linked to additional information on reactions, enzymes and genes, which may be stored in other databases. KEGG can be queried through *KGML* (KEGG Markup Language) [1], a language based on XML.

2.2 Comparison techniques for metabolic pathways

Many proposals exist in the literature for comparing metabolic pathways and whole metabolic networks in different organisms. Each proposal is based on some simplified representation of a metabolic pathway and on a related definition of similarity score (or distance measure) between two pathways. Hence we can group the various approaches in three classes, according to the structures they use for representing and comparing metabolic pathways. Such structures are:

- *Sets.* Most of the proposals in the literature represent a metabolic pathway (or the entire metabolic network) as the set of its main components, which can be reactions, enzymes or chemical compounds (see, e.g., [17, 18, 29, 22, 14, 13, 10, 48, 33]). This representation is simple and efficient and very useful when entire metabolic networks are compared. The comparison is based on suitable set operations.
- *Sequences.* A metabolic pathway is sometimes represented as a set of sequences of reactions (enzymes, compounds), i.e., pathways are decomposed into a set of selected paths leading from an initial component to a final one (see, e.g., [49, 30, 11, 27, 50]). This representation may provide more information on the original pathways, but it can be computationally more expensive. It requires methods both for identifying a suitable set of paths and for comparing them.

– *Graphs.* In several approaches, a metabolic pathway is represented as a graph (see, e.g., [20, 34, 16, 52, 28, 6, 12, 24, 31, 26, 7, 5]). This is the most informative representation in the classification, as it considers both the chemical components and their relations. A drawback can be the complexity of the comparison techniques. In fact the graph and subgraph isomorphism problems are GI-complete (graph isomorphism complete) and NP-complete, respectively. For this reason efficient heuristics are used and simplifying assumptions are introduced, which produce further approximations.

The similarity measure (or distance) and the comparison technique strictly depend on the chosen representation. When using a set-based representation, the comparison between two pathways roughly consists in determining the number of common elements. A similarity measure commonly used in this case is the *Jacard index* defined as:

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

where X and Y are the two sets to be compared. When pathways are represented by means of sequences, *alignment* techniques and *sum of scores with gap penalty* may be used as similarity measures. In the case of graph representation, more complex algorithms for *graph homeomorphism* 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 relies on a similarity measure between their components. 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 = 1) or different (similarity = 0).
– *EC hierarchy.* The similarity measure is based on comparing the unique *EC number* (Enzyme Commission number) associated to each enzyme, which represents its catalytic activity.
The EC number is a 4-level hierarchical scheme, $d_1.d_2.d_3.d_4$, developed by the International Union of Biochemistry and Molecular Biology (IUBMB) [51]. For instance, *arginase* is numbered by EC : 3.5.3.1, which indicates that the enzyme is a hydrolase ($EC : 3.*.*.*$), and acts on the "carbon nitrogen bonds, other than peptide bonds" (sub-class $EC : 3.5.*.*$) in linear amidines (sub-sub-class $EC : 3.5.3.*$). 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 their EC numbers:

$$S(e, e') = \max\{i : d_i = d'_i\}/4$$

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 is intended to correct 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. Given an enzyme class h, its information content is defined as $I(h) = -log_2 C(h)$, where $C(h)$ denotes the number of enzymes in h. The similarity between two enzymes e_i and e_j is $I(h_{ij})$, where h_{ij} is their lowest common upper class.
– *Sequence alignment.* The similarity measure is obtained by aligning the genes or the proteins corresponding to the two enzymes and by considering the resulting alignment score.

3 Behavioural Aspects in Metabolic Pathways Comparison

In this section we briefly discuss how to represent a metabolic pathway as a Petri net. Then we define a similarity measure between two metabolic pathways modelled as Petri nets, which takes into account the flows in the pathways by comparing their minimal T-invariants. Such measure is combined with a more standard one which considers homology of reactions.

3.1 Metabolic pathways as 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 [32] and [15] for surveys on PNs and their properties). A large number of tools have been developed for analysing properties of PNs. A quite comprehensive list can be found at the Petri net World site [3].

In some seminal papers Reddy et al. [37, 35, 36] and Hofestädt [21] proposed Petri nets (PNs) for representing and analysing metabolic pathways. Since then, a wide range of literature has grown on the topic [8]. 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. The number of tokens in each place indicates the amount of substance associated with that place. 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. 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. In this way, however, processes involving these substances, such as the energy balance, are not modelled.

Once metabolic pathways are represented as Petri nets, we consider their behavioural aspects as captured by the *T-invariants* (transition invariants) of the nets which, roughly, represents potential cyclic behaviours in the system. More precisely a T-invariant is a (multi)set of transitions whose execution starting from a state will bring the system back to the same state. Alternatively, the components of a T-invariant may be interpreted as the relative firing rates of transitions which occur permanently and concurrently, thus characterising a steady state. Therefore presence of T-invariants in 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.

Although space limitations prevent us from a formal presentation of nets and invariants, it is useful to recall that the set of (semi-positive) T-invariants can be characterised finitely, by resorting to its Hilbert basis [40].

Remark 1 (Unique basis). The set of T-invariants of a (finite) Petri net N admits a unique basis which is given by the collection $\mathcal{B}(N)$ of minimal T-invariants.

The above means that any T-invariant can be obtained as a linear combination (with positive integer coefficient) of minimal T-invariants. Uniqueness of the basis $\mathcal{B}(N)$ allows us to take it as a characteristic feature of the net.

The problem of determining the Hilbert basis is EXPSPACE since the size of such basis can be exponential in the size of the net. Still, in our experience, the available tools like INA [47] work fine on Petri nets arising from metabolic pathways.

In a PN model of a metabolic pathway, a minimal T-invariant corresponds to an elementary flux mode, a term introduced in [43] 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., [19, 25]) and they may provide insights into the network behaviour. 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 [38, 39, 41–44].

3.2 A combined similarity measure between pathways

Metabolic pathways are complex networks of biochemical reactions describing fluxes of substances. Such fluxes arise as the composition of elementary fluxes, i.e., cyclic fluxes which cannot be further decomposed. Most of the techniques 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 are eminently static and ignore the flow of metabolites in the pathway.

Here we propose a comparison between metabolic pathways based on the combination of two similarity scores derived from their Petri net representation. More precisely, we consider a "static" score, *R_score* (reaction score), taking into account the homology of reactions occurring in the pathways and a "behavioural" score, *I_score* (invariant score), taking into account the dynamics of the pathway as expressed by the T-invariants.

Both *R_score* and *I_score* are based on the *Sørensen index* [46] extended to multisets as below, where X_1 and X_2 are multisets and $\cap$ and $|\cdot|$ are intersection and cardinality generalised to multisets. [1]

$$S\_index(X_1, X_2) = \frac{2|X_1 \cap X_2|}{|X_1| + |X_2|}$$

Given two pathways represented as Petri nets P_1 and P_2, the *R_score* is computed by comparing their reactions. Each reaction is actually represented by the EC numbers of the associated enzymes. More precisely, if X_1 and X_2 denotes the multisets of the EC numbers in P_1 and $P2$ respectively, we define the *R_score* as

$$R\_score(X_1, X_2) = S\_index(X_1, X_2).$$

The similarity considered between enzymes is the identity, but finer similarity measures between enzymes, such as the one determined by the EC hierarchy, could be easily accommodated in this setting. We choose a multiset representation since an EC number may occur more than once in a pathway and we opted for the Sørensen index as it fits better to multisets than the Jacard index.

The distance based on reactions is then defined as follows

$$d_R(P_1, P_2) = 1 - R\_score(X_1, X_2).$$

The behavioural component of the similarity is obtained by comparing the Hilbert bases of minimal T-invariants. Each invariant is represented as a multiset of EC numbers, corresponding to the reactions occurring in the invariant, and the similarity between two invariants is given, as before, by the *S_index*. Note that when T-invariants are sets of transitions (rather than proper multisets) they can be seen as subnets of the net at hand, and the similarity between two T-invariants coincides with the *R_score* of the corresponding subnets. More generally, transitions can occur in an invariant with some multiplicity, which influences the similarity score.

A heuristic match between the two bases $\mathcal{B}(P_1)$ and $\mathcal{B}(P_2)$ is performed and the *S_index* values corresponding to the matching pairs are accumulated into I_Score(P_1, P_2) as described by the algorithm in Fig. 1.

[1] Formally, a multiset is a pair (X, m_X) where X is the *underlying set* and $m_X : X \to \mathbb{N}^+$ is the *multiplicity function*, associating to each $x \in X$ a positive natural number indicating the number of its occurrences. Then $|(X, m_X)| = \sum_{z \in X} m_X(z)$ and $(X, m_X) \cap (Y, m_Y) = (X \cap Y, m_{X \cap Y})$ where $m_{X \cap Y}(z) = min(m_X(z), m_Y(z))$ for each $z \in X \cap Y$.

```
function I_Score(P1, P2);
  input:  two metabolic pathways P1 and P2;
  output: the similarity measure between B(P1) and B(P2);
begin
  I1 = B(P1); I2 = B(P2);
  score = 0;
  card = max{|I1|, |I2|};
  while (I1 ≠ ∅ ∧ I2 ≠ ∅) do
  begin
    (X1, X2) = Find_max_Sim(I1, I2);   {Returns a pair of T-invariants, (X1, X2),
                                        in I1 × I2 such that S_index(X1, X2)
                                        is maximum}
    score = score + S_index(X1, X2);
    I1 = I1 − {X1};
    I2 = I2 − {X2};
  end;
  score = score/card;
  return score
end Compute_I_Score;
```

Fig. 1. Comparing bases of T-invariants

Again, pathways similarity based on minimal T-invariants induces a distance:

$$d_I(P_1, P_2) = 1 - I\_score(P_1, P_2)$$

The two distances are combined by taking a weighted sum as below, where $\alpha \in [0, 1]$:

$$d_D(P_1, P_2) = \alpha\ d_R(P_1, P_2) + (1 - \alpha)\ d_I(P_1, P_2)$$

The parameter α allows the analyst to move the focus between homology of reactions and similarity of functional components as represented by the T-invariants.

Two organisms O_1 and O_2 can be compared by considering n 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 consists in taking the average distance:

$$d_D(O_1, O_2) = \frac{\sum_{j=1}^{n} d_D(P_j^1, P_j^2)}{n}$$

When a pathway P_j occurs in one of the two organisms but not in the other, the corresponding pathway distance $d_D(P_j^1, P_j^2)$ in the formula above is taken to be 1.

4 Experimenting with CoMeta

In this section we briefly illustrate the tool CoMeta (Comparing METAbolic pathways) which implements our proposal, and we report on some experiments.

CoMeta is a user-friendly tool written in Java and running under Windows and Linux. Due to space limitation, we just list its main integrated functionalities:

- *Select organisms and pathways:* CoMeta proposes the lists of KEGG organisms and pathways and allows the user to select the ones to be compared. Such lists can be saved and then recovered for further processing.
- *Retrieve KEGG information:* the KEGG files corresponding to the selected organisms and pathways are automatically downloaded by CoMeta from the KEGG database [2].
- *Translate into PNs:* CoMeta automatically translates the selected organisms and pathways into corresponding Petri nets, by using an extension of the tool MPath2PN [9]. The PNML files describing the Petri nets thus obtained are available for further processing.
- *Compute T-invariants:* CoMeta uses the tool INA [47] to compute the bases of semi-positive T-invariants of the PN representations.
- *Compute Distances:* CoMeta automatically computes the reactions and invariants distances as defined in Section 3.2, and allows the user to specify the parameter α used for computing the combined distance. Distance matrices can be exported as text files. Moreover, CoMeta allows the user to inspect the details of the comparison between any pair of organisms (T-invariants bases, matches between invariants, reactions and invariants scores, etc.).
- *Show Phylogenetic trees:* the combined distance matrix may be the input of a phylogenetic tree construction method. Currently CoMeta implements the UPGMA and Neighbour Joining methods, and displays the corresponding phylogenetic trees.

4.1 Experiments

In order to validate our proposal CoMeta has been applied to many sets of organisms. We next show some interesting experiments.

Experiment 1. In the first experiment we consider the *glycolysis* pathway in five eucaryotes: *Homo sapiens* (HSA), *Rattus norvegicus* (RNO), *Meleagris gallopavo* (MGP), *Sus scrofa* (SSC), *Saccharomyces cerevisiae* (SCE).

The combined distance has been computed with the parameter α ranging in $\{0.00, 0.25, 0.50, 0.75, 1.00\}$. The corresponding phylogenetic trees built with the UPGMA method are shown in Figure 2.

The tree in Figure 2 (left) is built with $\alpha = 1$, i.e., by considering in the comparison only homology of reactions. In this case *Homo sapiens* and *Rattus norvegicus* are closely classified because they have the same *glycolysis* pathway, but *Meleagris gallopavo* is incorrectly close to them. The tree does not change for $\alpha = 0.75$. The tree in Figure 2 (right) is obtained with $\alpha = 0.5$, hence besides homology of reactions, it considers also the similarity of T-invariants

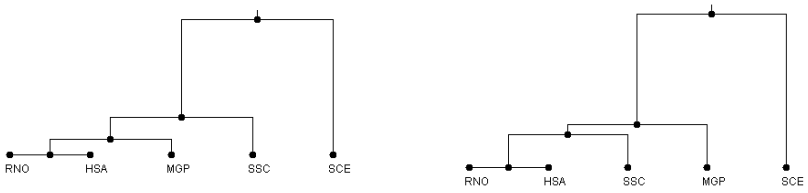

Fig. 2. UPGMA trees for Experiment 1, with $\alpha = 1$ (left) and $\alpha \leq 0.5$ (right).

(with weigth 0.5). This modifies the classification which now matches exactly the standard NCBI taxonomy [4]. With α smaller than 0.5, i.e., by increasing the relevance of the T-invariants in the computation of the distance, we obtain the same phylogenetic tree.

In this experiment the classification based on glycolysis obtained by considering only the distance on reactions does not match the NCBI taxonomy and it improves by taking into account the distance on T-invariants, i.e., the combined distance produces a better classification. This is not always true as shown by the next experiment.

Experiment 2 In this experiment we consider four eucaryotes – *Homo sapiens* (HSA) *Rattus norvegicus* (RNO) *C. elegans* (CEL) *Drosophila melanogaster* (DME) – and a bacterium – *E. coli* (ECO) – and three metabolic pathways, *glycolysis*, *pyruvate metabolism* and *purine metabolism*.

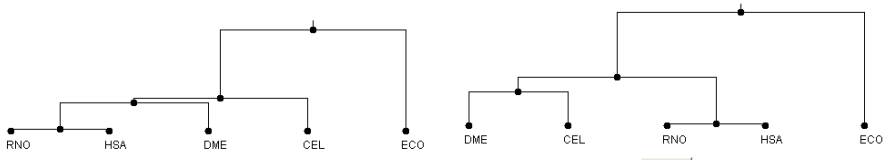

Fig. 3. UPGMA trees for experiment 2, with $\alpha = 1$ (left) and $\alpha \leq 0.75$ (right).

The results obtained with α ranging in $\{0.00, 0.25, 0.50, 0.75, 1.00\}$ are shown in Figure 3. The phylogenetic trees are built with the UPGMA method. The tree on the left of Figure 3, corresponds to $\alpha = 1$ and thus it considers only similarity of reactions in the comparison. This classification matches exactly the standard NCBI taxonomy [4] of the considered organisms. The tree in Figure 3 (right), corresponds to consider similarity both of reactions and of T-invariants, with $\alpha = 0.75$. The classification changes and it does not match any longer the standard NCBI taxonomy. This remains true by increasing the relevance of T-invariants i.e., with $\alpha = 0.50$ or smaller.

In this experiment, by considering the distance based on reactions we get a classification of the organisms matching the NCBI taxonomy. This is no longer

true when considering also T-invariants. This could be due to the fact that the reference NCBI taxonomy considers many characteristics of the organisms, not just a few metabolic functions as we do. In general, this shows that further experiments are necessary for understanding how to use our combined distance.

Experiment 3 The third experiment is conducted on a set of 16 organisms, mainly bacteria, w.r.t. the *glycolysis* pathway. It has been originally used in [20] as a test case and then considered also in [10]. The organisms and their reference NCBI taxonomy are show in Figure 4.

Focusing on an experiment already studied in the literature helps in comparing our technique with other proposals, although, as clarified below, a precise comparison is quite difficult for the variability of data sources and reference classifications.

Cod.	Organism	Reign
afu	A. fulgidus	Archea
mja	M. jannaschii	Archea
cpn	C. pneumoniae	Bacteria
mge	M. genitalum	Bacteria
mpn	M. pneumoniae	Bacteria
hin	H. influenzae	Bacteria
syn	Synechocystis	Bacteria
dra	D. radiodurans	Bacteria
mtu	M. tuberculosis	Bacteria
tpa	T. pallidum	Bacteria
bsu	B. subtilis	Bacteria
aae	A. aeolicus	Bacteria
tma	T. maritima	Bacteria
eco	E. coli	Bacteria
hpy	H. pylori	Bacteria
sce	Saccharomyces cerevisiae	Eucaryotes

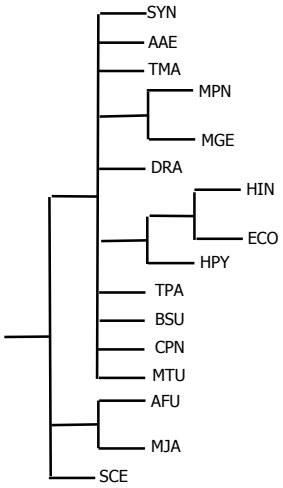

Fig. 4. Left: organisms for experiment 3. Right: reference NCBI taxonomy

As in the previous experiments, α ranges in $[0, 1]$, phylogenetic trees are built using the UPGMA method and they are compared with the reference NCBI classification of the 16 organisms. In order to perform such a comparison, following [20, 10], we used the *cousins* tool [53, 45] with threshold 2. The tool compares unordered trees with labelled leaves by counting the sets of common cousin pairs up to a certain cousin distance.[2] The outcome is reported in the table in Fig-

[2] A *cousin pair* is a triple consisting of a pair of leaves and their cousin distance: 0 if they are siblings (same parent), 0.5 if the parent of one of them is the grandparent of the other, 1 if they are cousins (same grandparent but not same parent), 1.5 if their first common ancestor is the grandparent of one of them and the great-grandparent

ure 5 (left). Our best result, 0.3163265, corresponds to the phylogenetic tree in Figure 5 (right) and to the combined distance with $\alpha \in [0.50, 0.75]$.

α	Similarity value
0.00	0.25
0.25	0.2673267
0.50	0.3163265
0.75	0.3163265
1.00	0.2621359

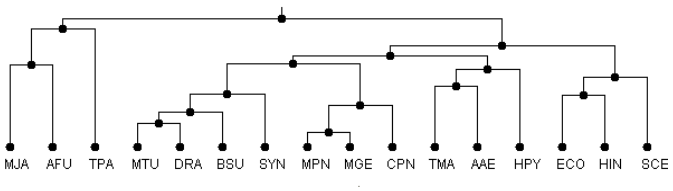

Fig. 5. Results for experiment 3. Left: similarity values computed with *cousins*. Right: UPGMA phylogenetic tree ($\alpha \in [0.50, 0.75]$).

Our results cannot be immediately compared with those in [20, 10]. In fact, the reference NCBI classification of the 16 organisms (and apparently also the corresponding KEGG data) has been changing in the meantime. Nevertheless, the experiment suggests that our technique produces results which are at least comparable with those in [20, 10].

In particular, in [20] a pathway is represented as an *enzyme graph* and a distance is defined which takes into account both the structure of the graph and the similarity between corresponding nodes. A phylogenetic tree is built with the resulting distance matrix by using the Neighbour Joining method. According to the authors, *cousins* provides a similarity value of 0.26 between their phylogenetic tree and their reference NCBI taxonomy and this outperforms the results of the phylogenies obtained by NCE, 16SrRNA and [29]. Hence our results improves those obtained in [20]. Although space limitations prevent us to report the details here, this is true also when we use Neighbour Joining trees.

Instead, in [10] a heuristic comparison algorithm is proposed which computes the intersection and symmetric difference of the sets of compounds, enzymes, and reactions in the metabolic pathways of different organisms. Their algorithm gives in output a similarity matrix which is used by a fuzzy equivalence relations-based (FER) hierarchical clustering method to compute the classification tree. The authors were not able to reproduce the experiment in [20]. In the *cousins* comparison w.r.t. the reference NCBI taxonomy their best result has a similarity value of 0.3195876, which is very close to our best result.

5 Conclusions

Biological questions related to evolution and to differences among organisms can be answered by comparing their metabolic pathways. In this paper we propose a new similarity measure for metabolic pathways which combines a similarity

of the other one, 2 if they are second cousins (same great-grandparent but not same grandparent) and so on.

based on reactions and a similarity based on behavioural aspects such as potential fluxes, which correspond to the minimal T-invariants of the Petri net representation of a pathway.

We implemented a tool, CoMeta, to experiment with our proposal. It is not easy to compare the results we obtained with those in the literature. Nevertheless experiments made with CoMeta showed that:

- Our combined measure produces valid phylogenetic classifications.
- Neither the comparison based on reactions nor the one based on T-invariants gives always correct results. The refinement due to the introduction of the behavioural measure can be useful, but further investigations are necessary to determine how to combine properly the two measures.
- Measures based on more sophisticated representations of a pathway (e.g., using graphs rather than sets, or considering also compounds besides enzymes) not necessarily give better results than our combined measure, as our third experiment shows. However also this hypothesis needs further experiments to be verified.

We are performing extensive studies on the distributions of the two proposed distances. This could reveal correlations between them, and, possibly, give insights on the ranges for the α parameter (influence of the T-invariants on the combined distance) providing the best results. We are also extending CoMeta to deal with a more refined similarity measure on EC numbers, the hierarchical similarity. We plan to add also the Tanimoto index (extended Jacard index) as an alternative to the Sørensen index. This would allow us to compare and evaluate different measures. When comparing organisms on large sets of pathways, a further extension would be to associate weights to the pathways. Weights could be chosen by the user in order to put more emphasis on some pathways of interest or could be derived on the basis of characteristics of the pathways, like their size.

Moreover, it would be very interesting to compare different organisms by considering their whole metabolic networks. This would allow one to identify more properly the T-invariants corresponding to functional units in the metabolic network. In fact, when considering single pathways some T-invariants can be not recognisable since they might be split in different pathways. However, the additional information deriving from the partitioning in well established functional pathways would be lost. Additionally, comparing full metabolic networks could be not viable from a computational point of view since in the worst case Hilbert bases can be exponential in the size of the original net.

CoMeta is part of a larger project to integrate various tools for representing and analysing metabolic pathways through Petri nets. We intend to use the distance matrices computed by CoMeta for different analyses. CoMeta is freely available at: http://www.dsi.unive.it/∼simeoni/CometaTool.tgz.

Acknowledgements. We are grateful to Paolo Besenzon and Silvio Alaimo for their contribution to the implementation of the tools used for the experiments.

References

1. Kegg Markup Language manual. http://www.genome.ad.jp/kegg/docs/xml.
2. KEGG pathway database - Kyoto University Bioinformatics Centre. http://www.genome.jp/kegg/pathway.html.
3. Petri net tools. http://www.informatik.uni-hamburg.de/TGI/PetriNets/tools.
4. Taxonomy - site guide - NCBI. http://www.ncbi.nlm.nih.gov/guide/taxonomy/.
5. F. Ay, M. Dang, and T. Kahveci. Metabolic network alignment in large scale by network compression. *BMC Bioinformatics*, 13 (Suppl 3), 2012.
6. F. Ay, T. Kahveci, and V. de Crecy-Lagard. Consistent alignment of metabolic pathways without abstraction. In *Int. Conf. on Computational Systems Bioinformatics (CSB)*, pages 237–248. 2008.
7. F. Ay, M. Kellis, and T. Kahveci. SubMAP: Aligning metabolic pathways with subnetwork mappings. *Journal of Computational Biology*, 18(3):219–235, 2011.
8. P. Baldan, N. Cocco, A. Marin, and M Simeoni. Petri nets for modelling metabolic pathways: a survey. *Natural Computing*, 9(4):955–989, 2010.
9. P. Baldan, N. Cocco, F. De Nes, M. Llabrés Segura, and M. Simeoni. MPath2PN - Translating metabolic pathways into Petri nets. In M. Heiner and H. Matsuno, editors, *BioPPN2011 Int. Workshop on Biological Processes and Petri Nets*, CEUR Workshop Proceedings, pages 102–116, 2011.
10. 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. of the 11th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, 2006.
11. M. Chen and R. Hofestadt. Web-based information retrieval system for the prediction of metabolic pathways. *IEEE Trans. on NanoBioscience*, 3(3):192–199, 2004.
12. Q. Cheng, R. Harrison, and A. Zelikovsky. MetNetAligner: a web service tool for metabolic network alignments. *Bioinformatics*, 25(15):1989–1990, 2009.
13. JC. 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.
14. O. Ebenhöh, T. Handorf, and R. Heinrich. A cross species comparison of metabolic network functions. *Genome Informatics*, 16(1):203–213, 2005.
15. J. Esparza and M. Nielsen. Decidability issues for Petri Nets - a survey. *Journal Inform. Process. Cybernet. EIK*, 30(3):143–160, 1994.
16. C.V. Forst, C. Flamm, I. L. Hofacker, and P. F. Stadler. Algebraic comparison of metabolic networks, phylogenetic inference, and metabolic innovation. *BMC Bioinformatics*, 7(67), 2006.
17. C.V. Forst and K. Schulten. Evolution of metabolism: a new method for the comparison of metabolic pathways using genomics information. *Journal of Computational Biology*, 6(3/4):343–360, 1999.
18. C.V. Forst and K. Schulten. Phylogenetic analysis of metabolic pathways. *Journal of Molecular Evolution*, 52(16):471–489, 2001.
19. M. Heiner and I. Koch. Petri Net Based Model Validation in Systems Biology. In *Petri Nets and Other Models of Concurrency - ICATPN 2004*, volume 3099 of *LNCS*, pages 216–237. Springer, 2004.
20. M. Heymans and A. M. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19(1):i138–i146, 2003.

21. R. Hofestädt. A Petri net application of metbolic processes. *Journal of System Analysis, Modelling and Simulation*, 16:113–122, 1994.
22. S. H. Hong, T. Y. Kim, and S. Y. Lee. Phylogenetic analysis based on genome-scale metabolic pathway reaction content. *Appl. Microbiol. Biotechnology*, 65:203–210, 2004.
23. M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi. KEGG for linking genomes to life and the environment. *Nuc. Acids Research*, pages 480–484, 2008.
24. G. W. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1), 2009.
25. I. Koch and M. Heiner. Petri nets. In B. H. Junker and F. Schreiber, editors, *Analysis of Biological Networks*, Book Series in Bioinformatics, pages 139–179. Wiley & Sons, 2008.
26. O. Kuchaiev, T. Milenkovic, V. Memisevic, W. Hayes, , and N. Przulj. Topological network alignment uncovers biological function and phylogeny. *J. R. Soc. Interface*, 7(50):1341–1354, 2010.
27. Y. Li, D. de Ridder, M.J.L. de Groot, and M.J.T. Reinders. Metabolic pathway alignment between species using a comprehensive and flexible similarity measure. *BMC Systems Biology*, 2008.
28. Z. Li, S. Zhang, Y. Wang, X.S. Zhang, and L. Chen. Alignment of molecular networks by integer quadratic programming. *Bioinformatics*, 23(13):1631–1639, 2007.
29. S. Liao, L.and Kim and J.F. Tomb. Genome comparisons based on profiles of metabolic pathways. In *Proc. of the 6th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES 02)*, pages 469–476, 2002.
30. E. Lo, T. Yamada, M. Tanaka, M. Hattori, S. Goto, C. Chang, and M. Kanehisa. A method for customized cross-species metabolic pathway comparison. In *Proc. of Genome Informatics 2004*. GIW 2004 Poster Abstract: P068, 2004.
31. A. Mithani, G.M. Preston, and J. Hein. Rahnuma: Hypergraph based tool for metabolic pathway prediction and network comparison. *Bioinformatics*, 25(14):1831–1832, 2009.
32. T. Murata. Petri Nets: Properties, Analysis, and Applications. *Proceedings of IEEE*, 77(4):541–580, 1989.
33. S. Oehm, D. Gilbert, A. Tauch, J. Stoye, and A. Goessmann. Comparative Pathway Analyzer - a web server for comparative analysis, clustering and visualization of metabolic networks in multiple organisms. *Nuc. Acids Research*, 36:433–437, 2008.
34. R.Y. Pinter, O. Rokhlenko, E. Yeger-Lotem, and M. Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16):3401–3408, 2005.
35. V. N. Reddy. Modeling Biological Pathways: A Discrete Event Systems Approach. Master's thesis, The Universisty of Maryland, M.S. 94-4, 1994.
36. V. N. Reddy, M.N. Liebman, and M.L. Mavrovouniotis. Qualitative Analysis of Biochemical Reaction Systems. *Comput. Biol. Med.*, 26(1):9–24, 1996.
37. V. N. Reddy, M. L. Mavrovouniotis, and M. N. Liebman. Petri net representations in metabolic pathways. In *ISMB93: First Int. Conf. on Intelligent Systems for Molecular Biology*, pages 328–336. AAAI press, 1993.
38. C. H. Schilling, D. Letscherer, and B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *Journal of Theoretical Biology*, 203:229–248, 2000.
39. C. H. Schilling, S. Schuster, B. O. Palsson, and R. Heinrich. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, 15:296–303, 1999.

40. A. Schrijver. *Theory of linear and integer programming.* Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1999.
41. S. Schuster, T. Dandekar, and D. A. Fell. Detection of elementary flux modes in biochemical networks: a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnology*, 17(March):53–60, 1999.
42. S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathway useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18(March):326–332, 2000.
43. S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2:165–182, 1994.
44. S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, 18(2):351–361, 2002.
45. D. Shasha, J. T. L. Wang, and S. Zhang. Unordered tree mining with applications to phylogeny. In *20th Int. Conf. on data engineering*, pages 708–719. IEEE Computer Society, 2004.
46. T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on danish commons. *Biologiske Skrifter / Kongelige Danske Videnskabernes Selskabg*, 5(4):1–34, 1948.
47. P.H. Starke and S. Roch. The Integrated Net Analyzer. *Humbolt University Berlin*, 1999. www.informatik.hu-berlin.de/ starke/ina.html.
48. Y. Tohsato. A method for species comparison of metabolic networks using reaction profile. *Inf. and Media Technology*, 2(1):109–114, 2007.
49. Y. Tohsato, H. Matsuda, and A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. In *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, pages 376–383, 2000.
50. Y. Tohsato and Y. Nishimura. Metabolic pathway alignment based on similarity between chemical structures. *Inf. and Media Technology*, 3(1):191–200, 2008.
51. E. C. Webb. *Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes.* San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press, 1992.
52. S. Wernicke and F. Rasche. Simple and fast alignment of metabolic pathways by exploiting local diversity. *Bioinformatics*, 23(15):1978–1985, 2007.
53. K. Zhang, J.T.L. Wang, and D. Shasha. On the editing distance between undirected acyclic graphs. *Int. Journal of Foundations of Computer Science*, 3(1):43–57, 1996.