Comparing Metabolic Pathways through Potential Fluxes: a Selectively Open Approach

Paolo Baldan[1], Martina Bocci[2], 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: [martina.bocci, cocco,simeoni]@unive.it

Abstract. In our previous work we developed CoMeta, a tool for comparing metabolic pathways of different organisms, using the KEGG database as data source. The similarity measure adopted combines homology of reactions and functional aspects of the pathways. The latter are captured by T-invariants in the Petri net representation, which correspond to potential fluxes in the pathways. A Petri net can model a metabolic pathway of an organism either in isolation, focussing on its internal behaviour (isolated net), or as an interactive subsystem of the full metabolic network (open net). Modelling a pathway as an isolated net normally works fine for comparison purposes, but unsatisfactory results can arise as it supplies a partial view on internal fluxes. A representation as an open net makes additional information available, but the choice of the interactions of the pathway with the environment is non-trivial. Considering all possible interactions with the environment (an information automatically retrieved from KEGG) is not appropriate. Some interactions may add noise to the model, the size of invariants bases grows up to an order of magnitude and the comparison results might be less precise than with the isolated representation. Here we propose an extension of CoMeta which allows the user to select which metabolites should be considered as interactions of interest, discriminating between input and output metabolites. We illustrate some experiments which show the advantages of this more flexible approach. Our experience suggests that in general a good choice is to take as open metabolites those which are the input and output compounds for the pathway.

1 Introduction

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.

G. Balbo and M. Heiner (Eds.): BioPPN 2013, a satellite event of PETRI NETS 2013,
CEUR Workshop Proceedings Vol. 988, 2013.

In [9, 6] we proposed to represent pathways as Petri nets (PNs) and compare them by considering static aspects, provided by the reactions, and information on the behaviour, as captured by the T-invariant bases of the corresponding Petri net models. Petri nets seem to be particularly natural for modelling metabolic pathways (see, e.g., [7] 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 [21] 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 used it in the comparison.

We developed CoMeta, a tool implementing our proposal. Given a set of organisms and a set of metabolic pathways, CoMeta automatically gets the corresponding data from the KEGG database [2], builds the corresponding Petri nets, computes the T-invariants and the similarity measure, and shows the results of the comparison among organisms as a phylogenetic tree.

The prototype version of CoMeta presented in [9] produced *isolated* PN models. In an *isolated model* the connections of the metabolic pathway with the environment are not represented. The potential fluxes which can be observed and compared with thus only the internal ones. According to our experiments, isolated net models normally work fine, but in some cases they may lead to unsatisfactory results. This happens when internal fluxes do not sufficiently characterise the behaviour of the net, for example when a pathway has very few internal cycles. In this case neglecting the interactions with the enviroment becomes problematic.

In [6] an extended version of CoMeta is proposed which gives the choice of producing either *isolated* or *open* PN models. In an *open model*, in order to express the interaction of the pathway with the environment, some compounds are represented as open places, i.e. places where the environment can freely put/remove substances through corresponding input/output transitions. Open places may be both the compounds which link the pathway to the rest of the metabolic network and the compounds which are only substrates or only products (the sources and the sinks of the net). In [6], by choosing an open model representation, all the compounds shared with the rest of the network and all the sources and sinks of a pathway are automatically modelled as open places. But further experiments show that this choice might lead to unsatisfactory results: some interactions reported in KEGG may be not precise or they may introduce noise in the comparison. Moreover, the added open places determine a growth in the size of invariants bases up to an order of magnitude, with consequences on the effieciency of the comparison.

The new version of CoMeta presented in this paper extends the previous ones with the possibility, for the user, to selectively open the model. The set of potentially open compounds is proposed to the user, who can decide, on

the basis of her/his knowledge, which should be the open input and output metabolites. Then, in addition to the internal fluxes, potential fluxes involving the chosen input/output metabolites will be considered in the comparison. From our experience it is always convenient to open the sources and the sinks in the networks. Hence currently in CoMeta this is the default choice proposed to the user, which is however free to add or remove metabolites as open places.

The paper is organised as follows. In Section 2 we show how a Petri net can model a metabolic pathway in the isolated and open approach. In Section 3 we briefly illustrate the new version of CoMeta and in Section 4 we present some experiments with it. A short conclusion follows in Section 5.

2 Petri net representation of a metabolic pathway

PNs are a well known formalism originally 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 [16] and [10] 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 Nets World* site [4].

Starting with [19, 14], Petri nets have been used as a model for representing and analysing metabolic pathways. A large body of literature exists on the topic (see, e.g, [7] for a survey). The structural representation of a metabolic pathway by means of a PN can be obtained 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. In particular, extended PNs can be enriched with a transition rate which depends on the kinetic law of the corresponding reaction.

When metabolic pathways are represented as Petri nets, we may consider their behavioural aspects as captured by the T-invariants (transition invariants) of the nets which, roughly, represent potential cyclic behaviours in the system. More precisely a T-invariant is a multiset of transitions whose execution starting from a state will bring the system back to the same 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.

The set of semi-positive T-invariants of a finite PN N admits a finitary representation by means of the so-called Hilbert basis [20], denoted $\mathcal{B}(N)$, which consists of the set of minimal T-invariants. Any T-invariant can be obtained as a linear combination (with positive integer coefficient) of elements of the basis.

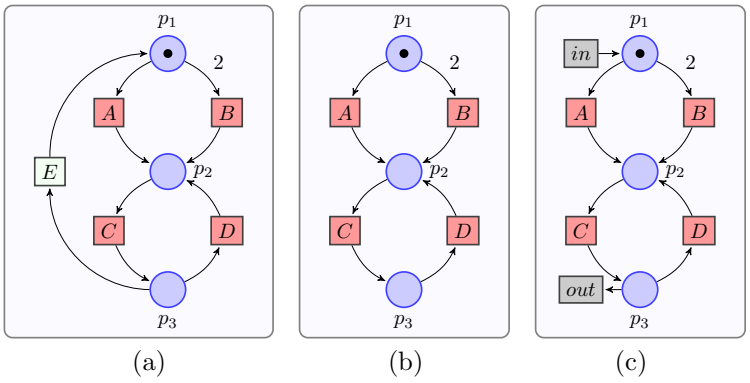


Fig. 1. A net system.

Uniqueness of the basis $\mathcal{B}(N)$ allows us to take it as a characteristic feature of the net.

In a PN model of a metabolic pathway, a minimal T-invariant corresponds to an elementary flux mode, a term introduced in [21] 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 have been used in Systems Biology as fundamental tool in model validation techniques (see, e.g., [13, 15]) and in analysis and decomposition techniques (see, e.g., [12, 11]).

The Petri nets corresponding to the metabolic pathways of an organism are subnets of a larger net representing its full metabolic network. They can be considered as *isolated* subnets, by ignoring their interactions with the environment, or as *open* subnets, i.e., interactive subsystems which exchange compounds with the environment. This is obtained by taking their input/output metabolites as open places, where the environment can freely put/remove substances. The minimal T-invariants of these subnets have a clear relation with (minimal) T-invariants of the full network. It can be easily seen that modelling the pathway as an isolated subsystem guarantees correctness: minimal T-invariants of the pathway are minimal T-invariants of the full network, but they capture only internal fluxes. If, instead, we consider the pathway as an open subsystem, then we get completeness: any invariant of the full network, once projected onto the pathway, is an invariant of the open pathway. The converse does not hold, i.e., there may be invariants of the open pathway which do not correspond to invariants of the full network. Hence, in the open approach, we may loose correctness, but, still, as shown in [18], minimal T-invariants of the full network can be obtained compositionally from those of the subnetworks.

As an example, consider the simple Petri net in Fig. 1(a). It has two minimal invariants, namely $I_1 = \{A, C, E\}$ and $I_2 = \{C, D\}$. Note that $\{B, C, E\}$ is not an invariant, since B requires two tokens in p_1. Assume that the subnet

of interest consists of the dark red transitions A, B, C, D (with their pre- and post-places, i.e., p_1, p_2 and p_3). The isolated representation of this subnet is given in Fig. 1(b). It is obtained by just removing transition E. Note that invariant I_2 is still there, while I_1 is lost. In the open representation of Fig. 1(c), places p_1 and p_2 are opened in input and output, respectively, meaning that the environment can put and remove arbitrarily many tokens in such places. This is represented by inserting the transitions *in* and *out*. As a consequence, there are three invariants in the open subnet: I_2, which was already in the original net, $\{in, A, C, out\}$, which is the projection of I_1 over the subnet, and $\{2{\cdot}in, B, C, out\}$ which, instead, does not correspond to any invariant of the original net.

The present version of CoMeta allows the user to choose either the isolated or the open view, and, in the latter case, to finely tune the representation of the compounds on which the interaction with the environment takes place.

3 The tool CoMeta

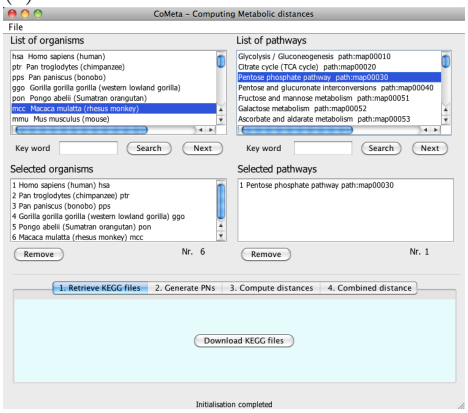
CoMeta, (COmparing METAbolic pathways) is a tool for comparing metabolic pathways in different organisms relying on their PN representation. The comparison is based on the combination of two distances, a "static" one, d_R, taking into account the reactions in the pathways and a "behavioural" one, d_I, taking into account potential fluxes in the pathways at steady state, as expressed by the T-invariants of the corresponding PNs. Given two pathways represented as PNs, P_1 and P_2, each distance is derived from a corresponding similarity score: $d_X(P_1, P_2) = 1 - score_X(P_1, P_2)$, with $X \in \{R, I\}$.

When computing d_R, $score_R(P_1, P_2)$ represents the similarity between the reactions in P_1 and the ones in P_2. Each reaction is represented by the enzymes which catalyse it and, in turn, each enzyme is identified by its EC number [26]. The similarity between enzymes is simply the identity, but finer similarity measures between enzymes could be easily accommodated in our setting. Concretely, $score_R(P_1, P_2)$ is a similarity index between the multisets of the EC numbers associated to the reactions in P_1 and P_2, respectively. The present version of CoMeta offers the choice between the Sørensen [24] and the Tanimoto [25] index extended to multisets.

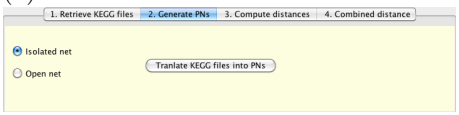
When computing d_I, the sets of minimal T-invariants (Hilbert bases) $\mathcal{B}(P_1)$ and $\mathcal{B}(P_2)$ of the two nets are compared. Each invariant is represented as a multiset of EC numbers, corresponding to the reactions in the invariant, and the similarity between two invariants is given, as before, by a similarity index. The similarity score is computed through a heuristic match between the two Hilbert bases and it represents the similarity of the matching pairs. In CoMeta the two distances may be combined: $d_C(P_1, P_2) = \alpha\ d_R(P_1, P_2) + (1-\alpha)\ d_I(P_1, P_2)$, with $\alpha \in [0, 1]$, to move the focus between reactions and functional components, and two organisms can be compared on n metabolic pathways $P_1, \ldots, P_n$ by considering their average distance on the n pathways. More details on the distances may be found in [6] where a prototype version of CoMeta was presented.

CoMeta is a user-friendly tool written in Java and running under Linux and Mac. CoMeta offers a set of integrated functionalities through a graphical user interface shown in Figure 2(a). In the upper part of the window the desired KEGG organisms and pathways can be selected. In the lower part a tabbed panel offers the commands to be performed. The first tab of the panel is shown in the main window, while the others are shown in Figure 2(b), 2(c), and 2(d), respectively. The main functionalities of the tool are the following ones:

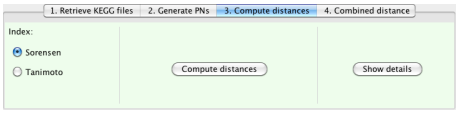
(a) CoMeta main window



(b) Second tab: Generate PNs



(c) Third tab: Compute Distances



(d) Fourth tab: Combined Distance

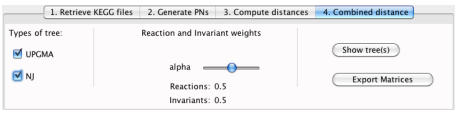


Fig. 2. The CoMeta graphical user interface

- *Select organisms and pathways* (Figure 2(a)): CoMeta proposes the lists of all KEGG organisms and pathways and allows the user to select the ones to be compared by double-clicking them.
- *Retrieve KEGG information:* CoMeta automatically downloads from the KEGG database the selected organisms and pathways.
- *Translate into PNs* (Figure 2(b)): CoMeta translates the selected organisms and pathways into corresponding PNs by using the tool *MPath2PN* [8]. *MPath2PN* produces a translation enzyme-based and without ubiquitous substances from KGML (KEGG Markup Language) [1] to PNML [3], a standard format for PNs tools. CoMeta produces the stoichiometric matrix of the net in a text file.
  Currently Cometa offers the possibility of representing the pathways either as an isolated or as an open subnet of the full metabolism. The user can model the pathway as an isolated subsystem and focus only on internal fluxes, or he can consider an open net and choose the open places among the compounds shared with the rest of the network and the compounds which are sources/sinks for the pathway. To assist the user, the tool proposes a canonical choice of open places, namely the sources and the sinks of the net, but this choice can be modified by adding and removing places in the

list of potential open places of the specific pathway and organism. Figure 3 shows the selectively open window for the organism *Vitis vinifera* wrt. the *Sulfate metabolism* pathway. Note that the first three checkbox columns in the window specify which metabolites link to other pathways and which ones are sources or sinks. By clicking on the checkboxes in the two rightmost columns the user can select to open in input or in output any metabolite in the pathway. The canonical choice, sources in input and sinks in output, automatically selected and proposed to the user, is shown in Figure 3.

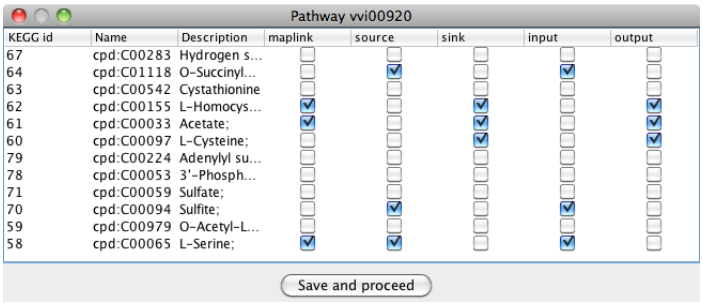


Fig. 3. The selectively open window with the canonical choice for *Vitis vinifera* wrt. *Sulfate metabolism*

- *Compute Distances* (Figure 2(c)): d_R and d_I are computed as previously described. The user can select either the Sørensen or the Tanimoto index. For computing Hilbert bases CoMeta resorts to 4ti2 [5], an efficient tool offered in a software package for solving algebraic, geometric and combinatorial problems on linear spaces. The details of the comparison between any pair of organisms (T-invariants bases, invariants matches, reactions and invariants scores, etc.) can be displayed to be analysed by the user.
- *Show Phylogenetic trees* (Figure 2(d)): CoMeta computes d_C, the distance which combines d_R and d_I according to a weight parameter α specified by the user. Such a distance may be used to produce and visualise corresponding phylogenetic trees. The user can specify the method for the generation of the phylogenetic trees. Currently CoMeta offers the UPGMA [23, 22] and Neighbour Joining methods [17, 22]. The matrices for d_R, d_I and d_C can be exported as text files for further analyses.

4 Experiments

In this section we discuss some experiments performed with CoMeta in order to illustrate how the choice of the isolated or the open approach may affect the results of the comparison of metabolic pathways. We consider a small group of organisms and analyse d_I, the distance based on T-invariants, with the Sørensen

index, when modelling the pathways as isolated, fully open and selectively open PNs. By fully open we mean a model in which all potentially open places, namely the ones linking the pathway to the rest of the metabolic network and the ones which are only substrates (sources) or only products (sinks), are indeed open. In the selectively open approach we use the default choice, i.e., we open in input source places by adding an input transition to each source, and we open in output sink places by adding an output transition to each sink.

The following experiments have a common feature: the pathways of the organisms we compare have few reversible reactions and few internal cycles. As a consequence, the comparison of the isolated PN models is not very detailed because of the small number of internal T-invariants and it produces only a rough classification of the organisms. On the other hand, by considering fully open models the information on the links among pathways given by KEGG become mostly relevant in the comparison, even if they are imprecise or not so important for distinguihing the specific functionalities. The classification of the organisms results distorted. In such cases, the selectively open approach seems to give the best results in the comparison, in fact it permits to add relevant information to the pathway model without overweighting boundary information. The resulting classification is more precise than in the other two approaches. In particular the canonical choice, i.e. opening in input source places and in output sink places, can be used with good results when no special knowledge on the pathway boundary is available.

4.1 Sulfur metabolism pathway

The *Sulfur metabolism* pathway describes the sulfur metabolism, including reduction and fixation processes. Sulfur enters in the composition of proteins (amino acids cysteine and methionine) and from the catabolism of these amino acids, it is liberated in the form of hydrogen sulphide (H_2S). Bacteria in soils and waters oxidise hydrogen sulfide in various steps, to its highest oxidation state − sulphate (SO_4^{2-}). Algae, Plants and Bacteria are capable to take the sulfur as sulphate and to process it to the most reduced form (sulfide) for incorporation into amino acids (cysteine and methionine). Animals are not able to synthesise methionine which is an essential amino acid to be assumed with the diet.

We report here on two different experiments performed with *Sulfur metabolism*. The first experiment aims at evaluating the ability of d_I to discriminate among very different groups of organisms. The second experiment aims at checking whether d_I is able to identify fine-grained differences among organisms.

First experiment. For this experiment we consider Archaea, Bacteria, Fungi, Plants (that are able to utilise sulfur as sulphate), Birds and Mammals (Animals, other than ruminants, take up sulfate only in reduced form in amino acids). The organisms are therefore expected to show similarity within each group and strong dissimilarity between groups. The list of selected organisms is shown in the following table.

Code	Organism	Reign
hsa	*Homo sapiens*	*Mammals*
ecb	*Equus caballus*	*Mammals*
gga	*Gallus gallus*	*Birds*
tgu	*Taeniopygia guttataa*	*Birds*
ath	*Arabidopsis thaliana*	*Plants*
osa	*Oryza sativa japonica*	*Plants*
bdi	*Brachypodium distachyon*	*Plants*
nfi	*Neosartorya fischeri*	*Fungi*
ang	*Aspergillus niger*	*Fungi*
cpw	*Coccidioides posadasii*	*Fungi*
cow	*Caldicellulosiruptor owensensis*	*Bacteria*
toc	*Thermosediminibacter oceani*	*Bacteria*
hsl	*Halobacterium salinarum R1*	*Archaea*
hvo	*Haloferax volcanii*	*Archaea*
pto	*Picrophilus torridus*	*Archaea*

By using CoMeta, we compute d_I, the distance based on T-invariants, for the isolated, selectively open and fully open approaches. The *Sulfur metabolism*

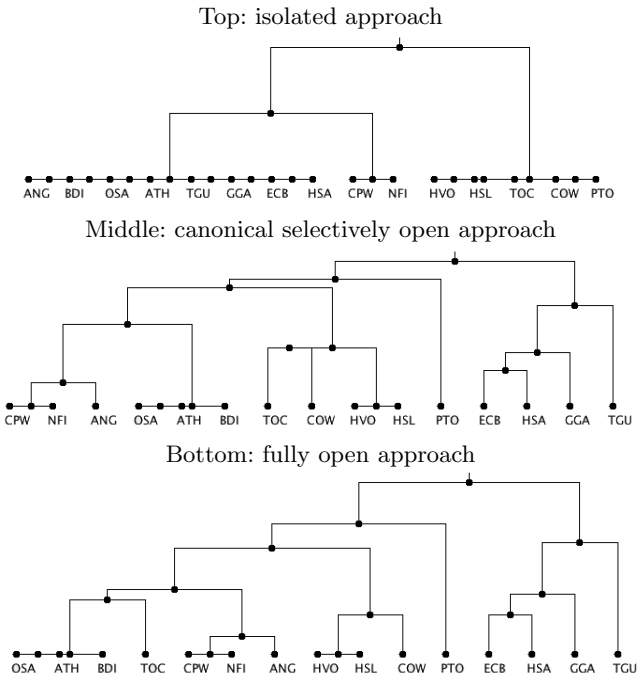


Fig. 4. First experiment: UPGMA trees based on d_I wrt. *Sulfur metabolism*

pathway has very small PN models. Depending on the organism, in the isolated models there are at most 9 enzymes/reactions and at most 1 T-invariant, in

the fully open models at most 19 enzymes/reactions and 6 invariants and in the canonical selectively open models at most 15 enzymes/reactions and 5 invariants. The selection is among 13 compounds at most.

Figure 4 shows the UPGMA tree corresponding to d_I in the isolated (top tree), selectively open (middle tree) and fully open (bottom tree) approaches. Note that in the isolated approach (top tree) Archaea and Bacteria are first discriminated from Fungi, Plants and Animals. At a second level, two out of three Fungi species are discriminated from Plants and Animals. No other grouping is evident and the classification is rather coarse. The selectively open approach (middle tree) discriminates at a first level the Animals (Mammals and Birds) from other groups. At a lower classification level all the three Fungi species are grouped together, as well as the three Plants species and four out of five Archaea and Bacteria species. Only the Archaea *pto* is not corectly grouped with the other Archaea (*hsl* and *hvo*) and Bacteria (*cow*, *toc*). Finally, with the fully open approach (bottom tree) the Archaea and Bacteria groups are not so well defined as in the previous case, i.e. *toc* and *cow* are not grouped together. Hence, in this experiment, the selectively open approach seems to better identify and group together organisms according to major taxonomic groups.

Second experiment We consider the *Sulfur metabolism* and the organisms in the following table.

Code	Organism	Reign
pae	*Pseudomonas aeruginosa PAO1*	*Bacteria*
pfo	*Pseudomonas fluorescens Pf0-1*	*Bacteria*
tin	*Thiomonas intermedia*	*Bacteria*
tcx	*Thiomicrospira crunogena*	*Bacteria*
cpr	*Clostridium perfringens SM101*	*Bacteria*
cst	*Clostridium stricklandii*	*Bacteria*
ddn	*Desulfovibrio desulfuricans ND132*	*Bacteria*
vvi	*Vitis vinifera*	*Plants*
zma	*Zea mays*	*Plants*

For this experiment we select within the Bacteria Reign some species having different sulfur metabolism and playing different roles within the bio-geo-chemical cycle of sulfur. The organisms *pae* and *pfo* are capable to oxidize elemental sulfur to sulfate. Sulfate can be assimilated by Plants and by Bacteria, such as the Clostridium species considered in the experiment. The sulfate-reducing bacteria, as *ddn*, are able to reduce sulfate to sulphide and responsible of bio-corrosion. On the contrary, *tin* and *tcx* oxidize sulphide back to sulfur. By using CoMeta, we compute d_I, the distance based on T-invariants. Figure 5 shows the UPGMA tree corresponding to d_I in the isolated (top tree), selectively open (middle tree) and fully open (bottom tree) approaches. Note that with the isolated approach (top tree) Plants, which assimilate sulfur as sulphate from the soil, are classified together with Bacteria (Pseudomonas genus) which are capable of oxidising sulfur to sulphate. On the right side of the tree, decomposing Bacteria (*cst* and

cpr) are grouped together with other oxidising Bacteria (*tcx* and *tin*). The selectively open approach (middle tree) provides a better classification, with all the sulphide/sulfur oxidising Bacteria grouped together (*tcx* and *tin*; *pfo* and *pae*). *Desulfovibrio desulfuricans* (*ddn*), a sulfate-reducing Bacteria is also well discriminated, as well as the two Plant species (*zma* and *vvi*), which assimilate sulphate, and the two decomposing Bacteria species (*cpr* and *cst*). The fully open approach (bottom tree) does not provide a well defined grouping of organisms as in the selectively open approach (e.g., *cst* and *cpr* are not grouped together; *pae* is erroneously grouped with *ddn*). In this experiment the canonical selectively open approach shows the ability to distinguish organisms at a fine classification level. In this case it is able to discriminate organisms belonging to the Bacteria Reign, having different ecological roles within the biological sulfur cycle.

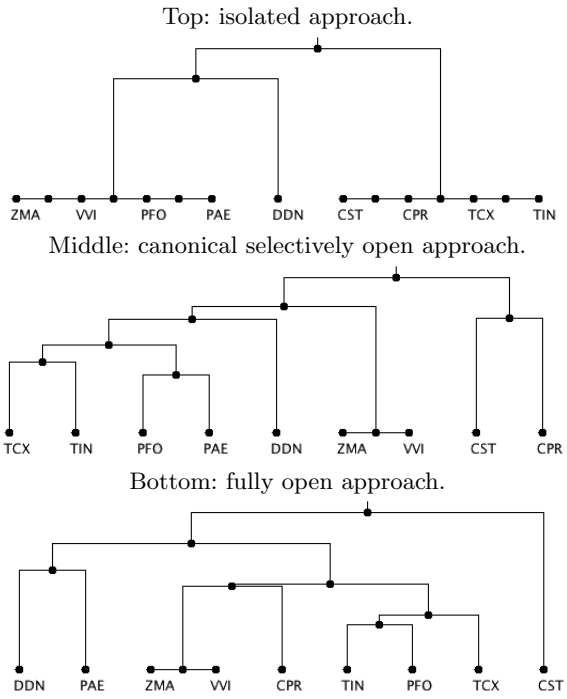


Fig. 5. Second experiment: UPGMA trees based on d_I wrt. *Sulfur metabolism*

Note that, by considering together all the organisms of the two experiments on the *Sulfur metabolism* the classification follows the same pattern: the more detailed classification is obtained by using the canonical selectively open approach, the isolated approach produces a rudimentary classification and the fully open approach does not produce a well defined classification. We preferred to anal-

yse the two groups separately in order to be able to show more precisely the granularity of the obtained classifications.

4.2 Carbon fixation pathway

In this experiment we consider the pathway *Carbon fixation in photosynthetic organisms*. This cycle consists of a series of reactions that lead to the biosynthesis of carbohydrates in the so called "dark phase" of photosynthesis. In most photosynthetic organisms this cycle is denominated *Calvin cycle* or *reductive pentose-phosphates cycle*. Some plants, in relation to environmental adaptations, exhibit specific variants of this cycle (C4 plants, CAM plants). A peculiarity of this pathway is that it is mainly composed by irreversible reactions.

We consider the organisms in Fig. 6 and we compute d_I, with the Sørensen index, for the isolated, selectively open and fully open approaches. Depending on the organism, the PN models contain at most 35 enzimes/reactions and 5 invariants in the isolated case, at most 52 enzimes/reactions and 42 invariants in the open case and at most 41 enzimes/reactions and 9 invariants in the canonical selectively open. In the selectively open approach the choice is among at most 34 compounds.

Code	Organism	Reign
gmx	*Glycine max*	*Plants, Eudicots*
pop	*Populus trichocarpa*	*Plants, Eudicots*
vvi	*Vitis vinifera*	*Plants, Eudicots*
osa	*Oryza sativa japonica*	*Plants, Monocots*
zma	*Zea mays*	*Plants, Monocots*
bdi	*Brachypodium distachyon*	*Plants, Monocots*
cre	*Chlamydomonas reinhardtii*	*Plants, green algae*
vcn	*Volvox carteri f. nagariensis*	*Plants, green algae*
npu	*Nostoc punctiforme*	*Bacteria*
acy	*Anabaena cilindrica*	*Bacteria*
oni	*Oscillatoria nigro-viridis*	*Bacteria*
mar	*Microcystis aeruginosa*	*Bacteria*

Fig. 6. Organisms for the experiment on the pathway *Carbon fixation in photosynthetic organisms*

Figure 7 shows the UPGMA tree corresponding to d_I in the isolated (top tree), canonical selectively open (middle tree) and fully open (bottom tree) approaches. Note that the isolated approach produces a rough classification, separating the bacteria from the other organisms. The selectively open approach permits the discrimination among the photosynthetic bacteria. The organism *vcn* is isolated due to its very simplified cycle, with a reduced number of intermediate products and enzymes involved. It is well-known that this genus is a very ancient group of organisms, originated from unicellular organisms. The

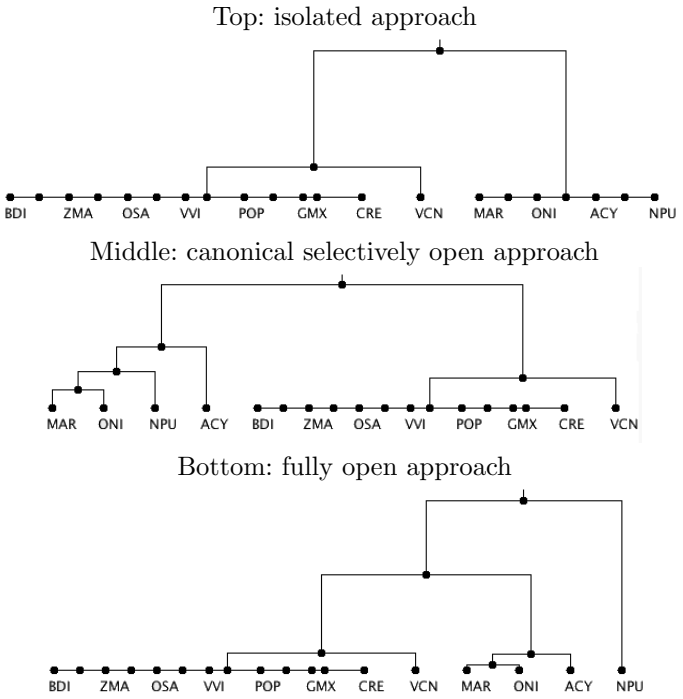


Fig. 7. Third experiment: UPGMA trees based on d_I wrt. *Carbon fixation in photosynthetic organisms*

fully open approach gives similar results to the selectively open one, but it does not group all the photosynthetic Bacteria together (*npu* is separated from all other organisms).

5 Conclusions

Metabolic pathways are subsystems of the full metabolic network. When constructing a model of a pathway this fact has to be taken into account and it requires some choices: the pathway can be represented in isolation or as a subsystem of the full network, interacting with its environment through some common compounds. In [9, 6] we proposed to use Petri net models for pathway comparisons based on reaction homology and functional aspects as captured by T-invariants. In this paper we consider the two modelling alternatives, isolated or open to any interaction with the environment, and conclude that neither of them is definitively better than the other. An isolated PN model guarantees correctness, namely minimal T-invariants of the pathway are minimal T-invariants of the full network, and it works well in most cases, but it captures only internal fluxes. Sometimes this is not sufficient to characterise the potential behaviours of the pathway, for example when a pathway has few internal cycles. In a fully

open PN model all potentially open places, namely the ones linking the pathway to the rest of the metabolic network and the ones which are sources or sinks, are indeed open. This approach may loose the correctness of T-invariants and in general it increases the size of the model without guaranteeing a better characterisation of the potential behaviours of the pathway. The information on the links among pathways becomes very relevant, even when they are not so important for distinguishing the functionalities associated to the pathway and, unfortunately, link informations is sometimes imprecise in KEGG. The most useful approach seems to be an intermediate one, in which a pathway is considered as an open subsystem, but the compounds on which the interaction takes place can be selected by the user. Our experience suggests a canonical choice of the open places which seems to produce the best results even in the absence of specific knowledge on a pathway, i.e., opening in input source places and opening in output sink places. We presented an extension of CoMeta, a tool for comparing metabolic pathways in different organisms, implementing our proposal. The tool retrieves the information about each selected pathway from the KEGG database, it determines the compounds which may potentially interact with the environment of the pathway and it offers to the user the possibility to select the interactions of interest, discriminating between input and output metabolites. CoMeta proposes the canonical choice in the selection (sources open in input and sinks open in output), but the user can freely add and delete open compounds. We presented some experiments which, although the work is still in a preliminary stage, suggest the appropriateness of this approach.

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 Markup Language. http://www.pnml.org.
4. Petri net tools. http://www.informatik.uni-hamburg.de/TGI/PetriNets/tools.
5. 4ti2 team. 4ti2—a software package for algebraic, geometric and combinatorial problems on linear spaces. Available at www.4ti2.de.
6. P. Baldan, N. Cocco, F. Giummolé, and M Simeoni. Comparing Metabolic Pathways through Reactions and Potential Fluxes. In W. van der Aalst and A. Yakovlev, editors, *Special Issue of ToPNoC(Workshops and Tutorials)*, LNCS. Springer, 2013. Accepted for publication.
7. P. Baldan, N. Cocco, A. Marin, and M Simeoni. Petri nets for modelling metabolic pathways: a survey. *Natural Computing*, 9(4):955–989, 2010.
8. 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.
9. P. Baldan, N. Cocco, and M Simeoni. Comparison of metabolic pathways by considering potential fluxes. In M. Heiner and R. Hofestädt, editors, *BioPPN2012 - 3rd International Workshop on Biological Processes and Petri Nets, satellite event*

of Petri Nets 2012, Hamburg, Germany, June 25, 2012, CEUR Workshop Proceedings - Vol-852, pages 2–17. ceur-ws.org, 2012. ISSN 1613-0073, online: http://ceur-ws.org/Vol-852.
10. J. Esparza and M. Nielsen. Decidability issues for Petri Nets - a survey. *Journal Inform. Process. Cybernet. EIK*, 30(3):143–160, 1994.
11. E. Grafahrend-Belau, F. Schreiber, M. Heiner, A. Sackmann, B. H. Junker, S. Grunwald, A. Speer, K. Winder, and I. Koch. Modularization of biochemical networks based on classification of Petri net t-invariants. *BMC Bioinformatics*, 9(90), 2008.
12. S. Hardy and P.N. Robillard. Petri net-based method for the analysis of the dynamics of signal propagation in signaling pathways. *Bioinformatics*, 24(2):209–217, 2008.
13. 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.
14. R. Hofestädt. A Petri net application of metbolic processes. *Journal of System Analysis, Modelling and Simulation*, 16:113–122, 1994.
15. 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.
16. T. Murata. Petri Nets: Properties, Analysis, and Applications. *Proceedings of IEEE*, 77(4):541–580, 1989.
17. Saitou N. and Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4):406–425, 1987.
18. M. Pedersen. Compositional definitions of minimal flows in Petri nets. In M. Heiner and A. M. Uhrmacher, editors, *Proceeedings of CMSB'08*, volume 5307 of *Lecture Notes in Computer Science*, pages 288–307. Springer, 2008.
19. 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.
20. A. Schrijver. *Theory of linear and integer programming.* Wiley-Interscience series in discrete mathematics and optimization. Wiley, 1999.
21. S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *Journal of Biological Systems*, 2:165–182, 1994.
22. P. Sestoft. Programs for biosequence analysis. http://www.itu.dk/people/sestoft/bsa.html.
23. R. Sokal and Michener C. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:14091438, 1958.
24. 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.
25. T.T. Tanimoto. Technical report, IBM Internal Report, November 17 1957.
26. 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.