A Taxonomy of Causality-Based Biological Properties *

C. Bodei[1], A. Bracciali[1], D. Chiarugi[2], and R. Gori[1]

1 : Dipartimento di Informatica, Università di Pisa, Italy
{chiara,braccia,gori}@di.unipi.it
2 : Dipartimento di Scienze Matematiche e Informatiche, Università di Siena, Italy
chiarugi3@unisi.it

We formally characterize a set of causality-based properties of metabolic networks. This set of properties aims at making precise several notions on the production of metabolites, which are familiar in the biologists' terminology. From a theoretical point of view, biochemical reactions are abstractly represented as causal implications and the produced metabolites as causal consequences of the implication representing the corresponding reaction. The fact that a reactant is produced is represented by means of the chain of reactions that have made it exist. Such representation abstracts away from quantities, stoichiometric and thermodynamic parameters and constitutes the basis for the characterization of our properties. Moreover, we propose an effective method for verifying our properties based on an abstract model of system dynamics. This consists of a new abstract semantics for the system seen as a concurrent network and expressed using the Chemical Ground Form [6] calculus. We illustrate an application of this framework to a portion of a real metabolic pathway.

1 Introduction

Understanding the relationships amongst the elements of biological interaction networks is a relevant problem in Systems Biology. In the words of [23], "diagrams of interconnections represent a sort of static roadmaps, but what we really seek to know are the traffic patterns, why such patterns emerge, and how we can control them". Formal descriptions of interconnections and methodologies for performing traffic simulations *in silico* can orientate *in vitro* experimentation.

We focus here on metabolic networks, i.e. the set of the cellular biochemical pathways involved in energy management and in the synthesis of structural components. Biochemical pathways are typically composed of chains of enzymatically catalyzed chemical reactions and are interconnected in a complex way. This makes difficult to understand the overall emerging behaviour of a network, starting from the detailed knowledge of the single reactions.

An interesting issue is the identification of the parts of a network whose integrity is crucial for certain functionalities. These "hot points" represent candidate drug targets for repressing undesired metabolic functions involved in pathological states, such as infectious diseases and cancer [9, 15]. Several properties characterizing different aspects of the network functionalities have been introduced in the biological literature, often with slightly different versions for the same property. What formal methods can offer is a way to make precise and classify properties, too often expressed only at an intuitive level.

Since, broadly speaking, *causality* plays a key role in finding chains of reactions that connect the parts of a network, we base our understanding of properties in terms of causality relations. Following the approach in [5], in order to give a formal characterization of causality-based properties, we interpret

*This work has been partially supported by the British-Crui Partnership Programme 2009.

E. Merelli and P. Quaglia (Eds.)
From Biology To Concurrency and back 2010 (FBTC'10)
EPTCS 19, 2010, pp. 116–133, doi:10.4204/EPTCS.19.8

biochemical reactions as "logical consequences", where the source metabolites *can* cause, i.e. produce the target ones. Furthermore, we adopt the notion of *explanation* of a certain metabolite. Given a set of reactions and initial conditions, an explanation represents the chain of reactions, causally dependent one from the another, that leads to the metabolite. Our approach therefore models the biochemical dynamics, capturing causal dependencies, while abstracting away from other aspects, like quantities, stoichiometric and thermodynamic parameters.

On top of the causality notion, we formalize several properties from a potentially longer list. Beyond the relevance of their biological meaning, these properties show how the few and simple ingredients we propose are sufficiently expressive to make precise several common notions, often intuitively used in biology. Specifically, the set of properties we present concerns the role and the relations of metabolites and reactions within a metabolic network.

We propose an effective method for verifying the formalized causal properties, based on the construction *once for all* of an abstract representation of the dynamics of the biological system. The system is specified as a concurrent network in terms of the Chemical Ground Form calculus [7]. We opted for the CGC for its extreme simplicity and well established theories and techniques, while it is, at the same time, sufficiently concrete to capture our properties. For our verification purposes, we have defined a slightly different semantics from the one in [7, 10]. It is worth pointing out that our choice mainly strives for simplicity. Other specification languages suitable for biological networks, e.g. [29, 34, 33, 8, 21], could have been adopted as well, some perhaps even more expressive, but generally requiring higher costs for model construction and verification procedures.

Overall, we are interested in efficiently evaluating the impact of changes on working hypotheses, such as the variation of the initial conditions and of the sets of reactions, according to a *what-if* strategy. The method we propose is meant to be exploited as a sort of preliminary *in silico* screening, aiming at determining the most promising experiments to be carried out *in vitro*. Finally, we believe that our framework should be palatable to biologists, since it is very close to the biochemical intuition of causality and to the spirit of many informal notions currently in use.

Related Work. Due to recent progress of wet-lab techniques, many metabolic networks are structurally well characterized and can be reconstructed for many organisms up to the genome-scale level (see e.g. [30]). However, approaches grounded on dynamical modeling, e.g. Metabolic Control Analysis or Metabolic Flux Analysis (see [16]), may encounter difficulties, mainly because part of the needed kinetic parameters are not known. In contrast, structure oriented analysis only requires information about the topology of the investigated networks, which is often known. Even though this kind of approach may not provide a detailed knowledge of the dynamics underlying the target phenomenon, it allows key properties of metabolic networks to be addressed, as demonstrated by the plethora of works in the literature. We mention here [41] and [40], where the authors propose to exploit "elementary modes" or "extreme pathways" to perform pathway analysis and to assess structural properties, such as structural robustness and redundancy. In [42], a method is presented that relies on the network structure for predicting robustness in gene regulation networks, while [2] reviews a group of works in which recurring patterns of interaction in biochemical networks (a.k.a. network motifs) are identified and related to specific behaviors or network robustness. In [26] a novel method is used to target those nodes whose deletion causes the failure of certain network functionalities.

Process algebras have been often used to abstractly model biological systems as concurrent systems, e.g. [37, 39, 38, 12, 13, 6, 36]. Closer to our approach is the work presented in [11], where the authors apply a causal semantics of the π-calculus [27] in order to describe biochemical processes. We use instead CGF, with a simpler semantics, but suitable for establishing the causal dependencies of interest.

Our results are close to those obtained by applying Control Flow Analysis (CFA), a quite efficient static technique, to process calculi used for modeling biological systems, e.g. [28, 29, 34, 33, 4]. In all the cases, an over-approximation of the behaviour of a system is offered. In particular, the analyses presented in [34, 33] capture causality information relevant for interpreting biological phenomena, and the authors propose a formalization of properties, like in our approach. Temporal and causal properties are also addressed in [21], where an Abstract Interpretation Analysis for systems specified in the BioAmbients calculus is used to model the quantities of molecules involved in interactions.

Our approach also shares some similarities with BIOCHAM [8] and the Pathway Logic [14]. BIOCHAM is based on the Biochemical Abstract Machine, which offers a formal modeling environment for biochemical processes and qualitative descriptions of these processes. BIOCHAM is based on a rule-centered language for specifying biochemical systems and, differently from our approach, it provides tools for querying temporal properties expressed in the Computation Tree Logic. Pathway Logic uses rewriting logic for modeling biological pathways and for enabling the symbolic analysis on them. In a way similar to ours, biochemical reactions are rendered in terms of rules acting on molecules. Both these approaches allow biochemical networks to be specified at a high level of abstraction. However, some expressible features, e.g. the distinction among different classes of molecules or reactions, have appeared too detailed for the aim of tracking causality and for our quest for a skeletal language for characterizing causality-based relevant properties.

As discussed, several of the mentioned approaches may provide more detailed models and properties than ours, however they generally require computationally expensive verification techniques. Our proposal combines the formalization of properties with a light-weight, approximate in some regards, computational machinery.

Synopsis. The metabolic network model is illustrated in §2, properties are formalized in §3 and the process-algebraic computational framework is introduced in §4. An example is discussed in §5.

2 A formal model of metabolic networks

We give an abstract representation of metabolic networks and of the corresponding biochemical reactions. More precisely, we abstract away from quantities, stoichiometric proportions, kinetic or thermodynamic parameters, that are involved in reactions, e.g. consider a standard biochemical reaction like:

$$aA + bB \quad \rightarrow^{r} \quad cC + dD \tag{1}$$

where A, B, C and D are the species involved, a, b, c and d are the corresponding stoichiometric coefficients, and r represents the rate at which reactants become products. We abstractly represent (1) as:

$$A \circ B \quad \rightarrow \quad C \circ D \tag{2}$$

We focus on the fact that the presence of both A and B represents the *possibility* for C and D to be *produced* or *caused*. Furthermore, we abstract from the dynamic evolution of the network, implicitly assuming that reactants are never consumed, that it is also also an abstraction over their quantities. As a consequence (2) reads as $A \circ B \rightarrow A \circ B \circ C \circ D$. Our model gives therefore an over-approximation of the set of the actual pathways, possibly including some pathways that could be actually prevented, for instance, by the lack of a suitable quantity of reactants or by an inadequate temperature.

For easing the computational machinery, we further decompose any rule causing more than one metabolite into a set of rules with only one caused metabolite each, e.g. rule (2) becomes $A \circ B \rightarrow C$ (3)

plus $A \circ B \to D$ (4). This transformation does not impact on causality: the set of metabolites producible by the original rule can be still produced by applying the new rules, as the premises are the same.

Finally, following [20], the unlikelihood of reactions involving more than two species, leads us to address only reactions with two causing metabolites at most. By summarizing, in the reactions we consider, either two molecules produce a new molecule, or a molecules degrades to another one.

Definition 2.1 (Rules) *Given a finite set of* metabolites **M**, *ranged over by over by A, A_i, B, C, D..., a* rule *is either in the form (1) $A_1 \circ A_2 \to C$, or (2) $A \to C$*

The description of causal relations within a metabolic network can be obtained by defining a set of reaction rules R that describe how new metabolites can be produced, and a set S of metabolites, initially present in the network solution, which can be seen as premise-less rules.

Definition 2.2 (m_network and Initial Solution) *An* m_network *R is a finite set of rules with non-empty premises. An* initial solution *S is a finite set of premise-less rules in the form $\to A$.*

The fact that a metabolite is caused by a network is made precise by means of the following definition relating the metabolite to a chain of reactions that produce it.

Definition 2.3 (Explanation) *$\mathscr{E}_{S,R}(C)$ is an* explanation *for $C \in \mathbf{M}$ with respect to S and R if either*

- *$C \in S$ and $\mathscr{E}_{S,R}(C) = C[\,]$, or*
- *$A_1 \circ A_2 \to C = r \in R$, $\exists \mathscr{E}_{S,R}(A_1), \mathscr{E}_{S,R}(A_2)$ and $\mathscr{E}_{S,R}(C) = C_r[\mathscr{E}_{S,R}(A_1), \mathscr{E}_{S,R}(A_2)]$, or*
- *$A \to C = r \in R$, $\exists \mathscr{E}_{S,R}(A)$ and $\mathscr{E}_{S,R}(C) = C_r[\mathscr{E}_{S,R}(A)]$.*

Note that a metabolite can be initially present in the solution or be produced anew from the network, or both. These cases can be distinguished by the structure of the relative explanations. For simplicity, hereafter in the following definitions we only report the case of rules in the form $A_1 \circ A_2 \to C$, by leaving out the simpler case of rules is in the form $A \to C$, where $\mathscr{E}_{S,R}(C) = C_r[\mathscr{E}_{S,R}(A)]$. For observing the explanation structure, we resort to the following auxiliary definition.

Definition 2.4 *Given an explanation $\mathscr{E}_{S,R}(C)$,*

- *the* set of metabolites *required for C, written $\mathscr{M}(\mathscr{E}_{S,R}(C))$, is defined as follows:*

$$\mathscr{M}(\mathscr{E}_{S,R}(C)) = \begin{cases} C & \text{if } \mathscr{E}_{S,R}(C) = C[\,] \\ \{A_1, A_2\} \cup \mathscr{M}(\mathscr{E}_{S,R}(A_1)) \cup \mathscr{M}(\mathscr{E}_{S,R}(A_2)) & \text{if } \mathscr{E}_{S,R}(C) = C_r[\mathscr{E}_{S,R}(A_1), \mathscr{E}_{S,R}(A_2)] \end{cases}$$

- *the* set of reactions *required for C, written $\mathscr{R}(\mathscr{E}_{S,R}(C))$, is defined as follows:*

$$\mathscr{R}(\mathscr{E}_{S,R}(C)) = \begin{cases} C & \text{if } \mathscr{E}_{S,R}(C) = C[\,] \\ \{r\} \cup \mathscr{R}(\mathscr{E}_{S,R}(A_1)) \cup \mathscr{R}(\mathscr{E}_{S,R}(A_2)) & \text{if } \mathscr{E}_{S,R}(C) = C_r[\mathscr{E}_{S,R}(A_1), \mathscr{E}_{S,R}(A_2)] \end{cases}$$

Of course, given S and R, there might be more explanation for the same metabolite C that corresponds to different ways to produce it. In turn, the explanation of another metabolite D that uses the metabolite C more than once, could include different explanations for C at different points. For the sake of simplicity, we assume to use only one explanation for each metabolite inside another explanation. For this reason, we introduce the notion of a *uniform explanation*.

Definition 2.5 (Uniform Explanation) *An explanation $\mathscr{E}_{S,R}(C)$ for $C \in \mathbf{M}$ w.r.t. S and R is a* uniform explanation *(written $\mathscr{E}^U_{S,R}(C)$) if it is an explanation for $C \in \mathbf{M}$ and $\forall D \in \mathscr{M}(\mathscr{E}^U_{S,R}(C))$, if $\mathscr{E}_{S,R}(D)$ and $\mathscr{E}'_{S,R}(D)$ occur in $\mathscr{E}^U_{S,R}(C)$, then $\mathscr{E}_{S,R}(D) = \mathscr{E}'_{S,R}(D)$, i.e. $\mathscr{E}^U_{S,R}(C)$ does not contain two different explanation for the same metabolite D.*

Since we are going to observe the whole set of explanations for each metabolite, in order to characterize our properties, we do not loose generality, by restricting ourselves to uniform explanations. From now on, we will then use only uniform explanations and therefore we will omit the superscript U. The following result relates general and uniform explanations.

Theorem 2.6 *Given S and R, we have that* $\exists\mathscr{E}_{S,R}(C)$ *iff* $\exists\mathscr{E}^{U}_{S,R}(C)$ *for all* $C \in \mathbf{M}$.

3 Causality-based properties

Several properties regarding metabolic networks, which are widely accepted at an informal level, can be made precise within our framework. Distinguishably, reasoning in terms of explanations adds an extra level of detail to the definition of the properties of interest, as well as having an explicit characterization of the network environment allows us to take into consideration the different conditions under which a network may work. We present properties that can be interpreted in terms of our notions of causality and explanations and that, given the abstraction of our model, are qualitative properties. We group them in properties about reactions and about networks. The first ones allow us to interpret the results of perturbative experiments, due to variations of the initial solution S or of the rules in R, while network properties have to do with robustness.

Reaction properties Often, the rules defining the reactions of a metabolic network correspond to enzymes that catalyze such reactions or to genes that code for such enzymes or for the proteins involved in reactions. Rules are hence the main object when studying a network behavior and it is quite natural trying to characterize their role in the production of metabolites. The next definition states when a rule has to be considered *essential* for the production of a given metabolite. A rule is essential if it is not dispensable, i.e. the network, deprived of it (e.g. by knocking-out the corresponding gene), is not able to produce the metabolite. Generally, in the biological literature, the notion of essentiality has been expressed informally and often referred to the elusive notion of viability of an organism, e.g. [19].

Definition 3.1 (Essentiality) *Given R, a rule* $r \in R$ *is* essential *in S for the metabolite C iff* $\exists\ \mathscr{E}_{S,R}(C)$ *and* $\nexists\ \mathscr{E}_{S,R\setminus r}(C)$.

Note that if a rule r is essential in S for the metabolite C, then all the explanations of C use r. From a biological point of view, it can be significant to distinguish amongst two degrees of essentiality. In the first case, essential rules correspond to those reactions whose essentiality holds only in a given solution S. Characterizing these "hot points" in a biochemical network operating in a given solution, can be useful when the studied networks are typically resident in a well defined environment. This is the case, e.g. , of drug development for cancer therapy, as malignant cells typically live in human blood or inter-cellular matrix. Essential rules in the metabolic network of malignant cells represent potential targets for anti-cancer drugs designed for disrupting that network. Since cancerous cells always act in a unique environment, it is important to identify their "weak points" always considering an initial solution S resembling the composition of human blood or inter-cellular matrix. In contrast, when the target system is an organism capable of living in various environments (such as a bacteria), identifying a stronger kind of essentiality, where a rule is essential for all possible solutions S, turns out to be a better choice in order to find "universal" targets for inhibiting the production of a given metabolite. Note that verifying this second kind of essentiality for a certain metabolite C is straightforward, because it simple amounts to verifying whether there is only one rule (not having C in the premise) for producing C.

Also relationships between rules have been traditionally explored, as has been done with the notion of mutual essentiality, e.g. [46]. We say that two rules are *mutually essential* for C, when their individual exclusion does not prevent the production of C, i.e. neither of the two rules is essential, but their simultaneous exclusion does. Detecting mutually essential reactions can be useful, again, in drug research for identifying multiple targets for drugs against parts of a network that represent functional alternatives for the production of a given metabolite.

Definition 3.2 (Mutual essentiality) *Given R, the rules $r1, r2 \in R$ are* mutually essential *in S for the metabolite C iff $\exists\, \mathscr{E}_{S,R\setminus r1}(C)$ and $\exists\, \mathscr{E}_{S,R\setminus r2}(C)$, while $\nexists\, \mathscr{E}_{S,R\setminus\{r1,r2\}}(C)$.*

Moreover, we establish that two explanations for a metabolite C are *vicarious* when they use different sets of rules, thus representing two different ways of producing C.

Definition 3.3 (Vicariate) *Given R, and S, and a metabolite C, an explanation $\mathscr{E}_{S,R}(C)$ is* vicarious *of $\mathscr{E}'_{S,R}(C)$ iff $\mathscr{R}(\mathscr{E}_{S,R}(C)) \neq \mathscr{R}(\mathscr{E}'_{S,R}(C))$.*

This property is related to the previous one, e.g. if two rules r_1 and r_2 are mutually essential for C, then $\mathscr{E}_{S,R\setminus r1}(C)$ is vicarious of $\mathscr{E}_{S,R\setminus r2}(C)$.

Furthermore, we investigate the order in which different metabolites are produced, and in particular we determine whether the production of a metabolite is a necessary condition (i.e. it is a *checkpoint*) for the production of another one.

Definition 3.4 (Checkpoint) *Given R and an initial solution S, B is* necessary *for C iff for all explanations of C $\mathscr{E}_{S,R}(C)$, $B \in \mathscr{M}(\mathscr{E}_{S,R}(C))$.*

Identifying checkpoints offers some insights on the structure of metabolic networks. From a topological point of view, checkpoint elements can be related to "bottlenecks" in molecular interaction networks [47]. As shown in [47] these elements, due to their strategical position in the network, are candidate for being essential as well as the reactions through which they are produced.

Similarly to the previous property, one can be interested in the order between rules and whether the application of some rules of R it is a necessary condition for the application of other rules.

Definition 3.5 (Causality) *Given R including r_1 and r_2, and an initial solution S, let the metabolite C be the conclusion of rule $r_2 \in R$. The rule r_1* causes *r_2 ($r_1 \sqsubseteq r_2$) iff for all explanations $\mathscr{E}_{S,R}(C) = C_{r_2}[\mathscr{E}_{S,R}(A_1), \mathscr{E}_{S,R}(A_2)]$, either $r_1 \in \mathscr{R}(\mathscr{E}_{S,R}(A_1))$ or $r_1 \in \mathscr{R}(\mathscr{E}_{S,R}(A_2))$.*

Note that if $r_1 \sqsubseteq r_2$ then the metabolite produced by rule r_1, say A, is necessary for the metabolite produced by rule r_2, say B, while if A is necessary for B it can be the case that $r_1 \not\sqsubseteq r_2$. Also this property can be exploited (eventually synergically with the checkpoint property) to gain topological insights concerning the investigated network. For instance if a rule r causes a group of other rules it is possible to say that r acts as a bottleneck.

The next property is useful to reason on which metabolites can be omitted from the initial solution, without compromising the initial capability of the system to produce metabolites in many different ways. Roughly speaking a metabolite can be omitted from the initial solution because it not necessary in the production of a given C or it is necessary but the system is already able to produce it.

Identifying these metabolites can aid in metabolic engineering [45], e.g. when for optimizing resources usage is requested to characterize the minimal environment needed for a bioreactor. Note that the so-called conditional mutants differ from the *wild type* (i.e. the microorganism possessing the genome commonly found in nature) only for the minimal environment needed for their viability. The genome of conditional mutants do not code for an enzyme essential for its life and their survival is conditioned by the presence in S of the metabolite produced by the missing reaction.

Definition 3.6 (Redundancy) *Given R and a metabolite C, an initial solution S is* redundant *for $C \notin S$ iff there exists at least a metabolite $B \in S$ s.t. for all $\mathscr{E}_{S,R}(C)$ for C, there exists $\mathscr{E}_{S\setminus\{B\},R}(C)$ such that $\mathscr{R}(\mathscr{E}_{S,R}(C)) \subseteq \mathscr{R}(\mathscr{E}_{S\setminus\{B\},R}(C))$.*

According to this definition, a given initial solution S can be redundant for a metabolite C in the *wild-type* but not redundant for the same metabolite in the conditional mutant. Of course, the previous property can be weakened obtaining another property that checks whether R is still able to produce C after the exclusion of the reagent B from the initial solution. In other words, it addresses the impact of the exclusion of some metabolites from the initial solution, offering straightforward applications to the resource optimization problem described above.

Definition 3.7 (Exclusion) *Given R, a metabolite $B \in S$* cannot be excluded *for the production of metabolite $C \notin S$ iff $\nexists\, \mathscr{E}_{S\setminus\{B\},R}(C)$.*

Note that B cannot be excluded for the production of C if and only if for all explanations $\mathscr{E}_{S,R}(C)$, $B[]$ does occur in $\mathscr{E}_{S,R}(C)$. The previous property is in some way related to the checkpoint property expressed before: if B belongs to S and is not necessary for the production of C, then we can harmlessly exclude it from the initial solution. However, in general, the two properties do not coincide (see Ex. 4.13).

Metabolic network properties Informally, robustness can be defined as the capability of a whole network of resisting to damages. In the biological literature there is not a common agreement on what "robustness" exactly means [25]. One of the most used definition says: "robustness is a property that allows a system to maintain its functions against internal and external perturbations." ([24]). A similar definition [44] is also widely used: "robustness, the ability to maintain performance in the face of perturbations and uncertainty, is a long-recognized key property of living systems". Both definitions, however, result to be not well assessed and therefore open to different possible interpretations. Moreover, the notion of robustness is related to the maintenance of a "function" or of "performance". Both these concepts subsume quantitative issues and their exact meaning change depending on the work considered. The uncertainty in definitions makes difficult both evaluating robustness effectively and comparing different networks addressing this property. A more reliable way to assess this notion consists in considering the qualitative features of the network in hand rather than its quantitative throughput [1, 43]. The notion of network robustness can be linked to the overall error tolerance, seen as the capability of carrying information in spite of local failures which, in turn, depend critically on the topology of network wiring [1]. In our framework this corresponds to evaluate the resistance to failures in terms of the maintenance of the capability of producing a given metabolite. This paradigm-shift allows us to propose the following formal notions of robustness.

Definition 3.8 (Strong Robustness) *Given two m_networks R_1 and R_2, R_2 is* strongly more robust than *R_1 in S for C, written $R_1 \ll_{S,C} R_2$ iff*

$$\exists\, \mathscr{E}_{S,R_1}(C) \quad \Rightarrow \quad \exists\, \mathscr{E}_{S,R_2}(C) \quad \wedge \quad \mathscr{R}(\mathscr{E}_{S,R_1}(C)) = \mathscr{R}(\mathscr{E}_{S,R_2}(C)).$$

This is quite a strong requirement, accounting to say that all the rules used for producing P in R_1, are present in R_2 and can be used as well. Of course, R_2 may also allow more explanations, using different rules. This consideration leads us to formulate a weaker property, by requiring that R_2 is able to produce the same metabolite, without constraints on the rules to be used.

Definition 3.9 (Weak Robustness) *Given two m_networks R_1 and R_2, R_2 is* weakly more robust than *R_1 in S for C, written $R_1 \prec_{S,C} R_2$ iff*

$$\exists\, \mathscr{E}_{S,R_1}(C) \quad \Rightarrow \quad \exists\, \mathscr{E}_{S,R_2}(C)$$

R	$::= 0 \mid A = M, R$	Reagents Environment	(empty, or a reagent and Reagents Env)
M	$::= 0 \mid \pi^{\lambda}.S + M$	Molecule	(empty, or an interaction and Molecule)
S	$::= 0 \mid A \mid S$	Solution	(empty, or a variable and Solution)
π	$::= a \mid \overline{a} \mid \tau$	Basic Action	(input, output, delay)
CGF	$::= (R, S)$	Chemical Ground Form	(reagent environment with initial Solution)

Table 1: Syntax of simplified CGF

4 Verification Methodology

Our methodology is based on the construction of an abstract model of the biological system. The model is obtained by a new abstract semantics of the system, interpreted as a concurrency network and expressed using the Chemical Ground Form (CGF) [7] calculus. We exploit the notion of path for verifying our properties. The CGF is a fragment of the stochastic π-calculus [35, 32]. Since we abstract away from quantities, we resort to a simplified version of CGF, in which stochastic features, like action rates, are discarded. In particular, we abstract away from the version of CGF, presented in [10, 22], because we represent solutions as sets of reagents rather than multisets. The syntax of CGF is defined in Tab. 1. We consider a set of *Names* (ranged over by $a, b, c, \ldots$), a set of *labels* $\mathscr{L}$ (ranged over by $\lambda, \mu \ldots$), and a set *Mol* (ranged over by *A*,*B*,....) of *variables* (the reagents). A CGF specification is composed by a (finite) list of reagent definitions $A_i = M_i$, where A_i is a variable that stands for the name of a chemical species and M_i is a molecule that describes the interaction capabilities of the corresponding species. The environment R defines the reagents of a solution S. A *molecule* M may do nothing, or may change after a delay (e.g. because of a molecular decay) or may interact with other reagents. A standard notation is adopted: τ represents a delay; a and $\overline{a}$ model interaction over a shared channel a (the input and output, respectively). Together with the reagents definition, a CGF includes a solution S, that represents the initial conditions and is described by a parallel composition of variables, i.e. a finite list of reagents. This maps onto the initial solution from Def. 2.2.

In order to distinguish the actions that participate to a move, we label them. In a CGF (R, S), R is *well-labeled*, if basic action labels are all distinct. We assume to have well-defined reagents environment and, given R, to have a definition for each variable A in R or in S. Moreover, given a label $\lambda \in \mathscr{L}$, we use the notation $R.A.\lambda$ to indicate the process $\pi^{\lambda}.S$, provided that $A = \ldots + \pi^{\lambda}.S + \ldots$ is the definition of A occurring in R. Finally, given a CGF (R, S), we denote with $\top$ the set of all molecules occurring in either S or in the rules of R. In the following, we will use *Sol* for the domain of sets of reagents. There is one transition rule for delay actions and one for synchronizations. Transition are of the form

$$S \xrightarrow{\Theta, \hat{S}, X} S' \text{ with } S, S', \hat{S} \in Sol,\ X \in Mol,\ \Theta \in \widehat{\mathscr{L}} = \mathscr{L} \cup (\mathscr{L} \times \mathscr{L})$$

- Θ reports the label(s) of the basic action(s), which participate to the move,
- $\hat{S} \subseteq S_0$ reports the subset of the reagents of the initial solution directly involved in the current move
- X reports the *unique* reagent produced by the move (i.e. by the corresponding reaction).

The Rule (**Delay**) models the move of a process $\tau^{\lambda}.Q$ appearing in the definition of a reagent A. The transition records the label λ, the singleton A if A belongs to the initial solution S_0 and the reagent produced by the reaction. The Rule (**Sync**) models the synchronization between two processes $a^{\lambda}.Q_1$

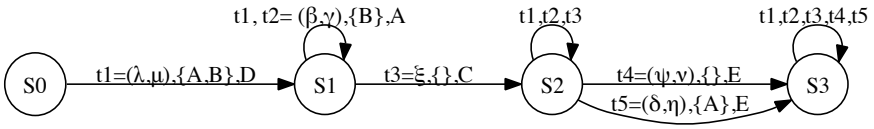

Figure 1: LTS Graph of Ex. 4.1, where $S_0 = \{A,B\}$, $S_1 = S_0 \cup \{D\}$, $S_2 = S_1 \cup \{C\}$, $S_3 = S_2 \cup \{E\}$

and $\overline{a}^{\mu}.0$ occurring in the definition of A and B, resp. The transition records the label pair (λ,μ), together with the part of S_0 used in the rule (i.e. $\{A,B\} \cap S_0$) and the possibly new reagent produced by the reaction.

$$(\textbf{Delay}) \quad \frac{R.A.\lambda = \tau^{\lambda}.Q}{S \xrightarrow{\lambda,\{A\}\cap S_0,Q} S \cup \{Q\}} \qquad (\textbf{Sync}) \quad \frac{R.A.\lambda = a^{\lambda}.Q \qquad R.B.\mu = \overline{a}^{\mu}.0}{S \xrightarrow{(\lambda,\mu),\{A,B\}\cap S_0,Q} S \cup \{Q\}}$$

We denote with $Tr((R,S_0)) = (\mathscr{S},\rightarrow,S_0,R)$ the *labeled transition system* (LTS), obtained, starting from the initial state $S_0 \in \mathscr{S}$, w.r.t. to environment R, and with $Gr((R,S_0))$ the corresponding graph. Since environments are well-labeled, different transitions leaving from the same state carry distinct labels.

For simplicity, we identify each reaction by a label $\Theta \in \widehat{\mathscr{L}}$. In our model, a reaction $r : A \circ B \rightarrow D$ can be identified by (λ,μ) and rendered by the following reagent definitions:

$$A = a^{\lambda}.D$$
$$B = \overline{a}^{\mu}.0$$

Given an initial solution $S_0 = \{A,B\}$, the system may perform the transition $\{A,B\} \xrightarrow{(\lambda,\mu),\{A,B\},D} \{A,B,D\}$, since $\hat{S} = \{A,B\} \cap \{A,B\}$ and $S' = S_0 \cup \{D\}$. If A and B are involved in other reactions, other actions can be added in their specifications, as in the example below, where we illustrate our approach on a toy reaction network.

Example 4.1 *Consider the initial solution $S_0 = \{A,B\}$ and an m-network R, consisting of the rules reported below, on the left-hand side.*

$$\begin{array}{lrcl} (\lambda,\mu) & A \circ B & \rightarrow & D \\ (\delta,\eta) & A \circ C & \rightarrow & E \\ (\beta,\gamma) & D \circ B & \rightarrow & A \\ \xi & D & \rightarrow & C \\ (\psi,\nu) & D \circ C & \rightarrow & E \end{array} \qquad \begin{array}{l} A = a^{\lambda}.D + \overline{b}^{\eta}.0 \\ B = \overline{a}^{\mu}.0 + d^{\beta}.A \\ D = \tau^{\xi}.C + \overline{c}^{\nu}.0 + \overline{d}^{\gamma}.0 \\ C = b^{\delta}.E + c^{\psi}.E \end{array}$$

The corresponding CGF specification is above on the right, while the corresponding graph is in Fig.1. For simplicity, in the presence of multiple self-loops, we collapse the self-loop arcs in a single one.

- *Starting from $S_0 = \{A,B\}$, the only possible transition (here called t_1) is the one that uses rule (λ,μ) and leads to the state S_1 containing D. After this,*
- *either we can fire transition t_3 (rule ξ), that leads to S_2, that includes C;*
- *or, we can fire transition t_2 (rule (β,γ)) that leads to S_1, where A is already present.*
- *From S_2, both transitions t_4 (rule (δ,η)) and t_5 (rule (ψ,ν)) are possible, lead to S_3 and produce E.*
- *Intuitively, we can observe that some transitions,* cause *some others: t_1 causes t_2, t_3 t_4 and t_5, t_3 causes both t_4 and t_5, while t_4 and t_5 are independent from each other.*

• *We have two paths reaching a state that includes E:* $S_0 \xrightarrow{t_1} S_1 \xrightarrow{t_3} S_2 \xrightarrow{t_4} S_3$ *that corresponds to the explanation* $E_{(\delta,\eta)}[A[],C_\xi[D_{(\lambda,\mu)}[A[],B[]]]]$, *and* $S_0 \xrightarrow{t_1} S_1 \xrightarrow{t_3} S_2 \xrightarrow{t_5} S_3$, *that corresponds to the explanation* $E_{(\psi,\nu)}[D_{(\lambda,\mu)}[A[],B[]],C_\xi[D_{(\lambda,\mu)}[A[],B[]]]]$.
• *Establishing which metabolites in* S_0 *are used in each transition, can be useful to investigate their impact on the production of the other metabolites. For instance, A is necessary for the production of C and E, because both the states including C and E are reached, using* t_1 *(rule* (λ,μ)*), that requires A.*

In the following, we are going to make precise the notions only informally introduced in Ex. 4.1. Note that self-loops can correspond either to the application of a rule already applied or to the application of a rule, that has not already been applied, but that produces a metabolite already present. Self-loops of the first kind do not add any useful information from a causality point of view, in terms of the properties introduced in the previous sections. Self-loops of the second kind can be instead useful, especially if we are interested in checking the possibility of the system to produce a certain metabolite even if it is already present in the initial solution. We first focus on the computation paths not including self-loops at all, that we call *causally* relevant paths. This notion is used to verify many properties of the previous section.

Definition 4.2 (χ-path) *A path p in* $Gr((R,S_0))$ *is a* causally relevant *path (χ-path) if*

$$p = S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots S_{m-1} \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m \text{ and } S_i \neq S_{i-1} \text{ for all } i \in [1,m]$$

We say that p leads to *C if* $C = X_{m-1}$ *(i.e. if* S_m *is the first state including C).*

Theorem 4.3 (Correspondence) *Given a* χ*-path p in* $Gr((R,S_0))$ *that* leads to *C*

$$p = S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots . S_{m-1} \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m$$

let $tr_p()$ *be the function, which for a given path p and reagent B is defined as follows:*

$$tr_p(B) \;=\; \begin{cases} B_\Theta[tr_p(A_1),tr_p(A_2)] & \text{if} \quad \exists\, i \in [1,m].\ X_{i-1} = B,\ \text{and } \Theta : A_1 \circ A_2 \to B \\ & \qquad \text{is the rule applied in the transition } t_i, \\ B[\,] & \text{if} \quad B \in S_0. \end{cases}$$

We obtain an explanation $\mathscr{E}_{S_0,R}(C)$ *for C, as* $tr_p(C)$*, which uses the same rules of p.*

Moreover, given an explanation $\mathscr{E}_{S_0,R}(P)$ we can construct a set of corresponding paths, starting from the subset of the initial solution used in the explanation, i.e. all the metabolites A such that $A[]$ occurs in $\mathscr{E}_{S_0,R}(P)$. We then proceed by exploring the explanation structure from innermost outermost and therefore firing the transitions corresponding to the rules used in the explanation. Note that, serializing the possible parallelism of the explanation can give rise to a set of paths rather than to a unique path. For instance, if we start from an explanation $C_{\Theta_1}[A_{\Theta_2}[B[],D[]],F_{\Theta_3}[E[],G[]]]$, corresponding to the application of rules $\Theta_1 = A \circ F \to C$, $\Theta_2 = B \circ D \to A$ and $\Theta_3 = E \circ G \to F$, then we have two corresponding paths, where the order in which the transitions occur is different. Note that paths obtained by $\mathscr{E}_{S_0,R}(P)$ are χ-paths.

Definition 4.4 (ρ-path) *A path p in* $Gr((R,S_0))$ *is a* relevant *path (ρ-path) if*

$$p = S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots S_{m-1} \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m \text{ and } \forall\, j \in [1,m]\ S_j = S_{j-1} \Rightarrow$$

(i) $X_j \in S_0$ *(the produced metabolite was already in* S_0*),*
(ii) $\{X_j\} \cap (\bigcup_{0 \le i < j} X_i) = \emptyset$ *(the produced metabolite was never produced before),*
(iii) $\{X_j\} \cap (\bigcup_{0 \le i < j} \hat{S}_i) = \emptyset$ *(the produced metabolite was never required before).*

We say that p leads to *C if* $C = X_{m-1}$.

Intuitively, conditions (i)-(iii) will aid us to determine which metabolites could harmlessly excluded from the initial solution, identifying the metabolites that the system itself is able to produce before they are required. Note that ρ-paths as well as χ-paths are always finite: by definition, a self-loop transition can be included in a ρ-path at most once. In particular, each χ-path is also a ρ-path.

We are now ready to characterize all the properties introduced in §3, in terms of χ- and ρ-paths. A rule Θ is *essential* for the metabolite C if every χ-path leading to C includes Θ, while the rules Θ_1 and Θ_2 are *mutually essential* for C if every χ-path leading to C includes at least one of the two rules.

Theorem 4.5 (Essentiality and Mutual Essentiality)

- *A rule Θ, is* essential *in S_0 for a reagent $C \notin S_0$ iff $\forall$ χ-path $S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m$ in $Gr((R,S_0))$, leading to C, there exists at least an $i \in [0,m-1] : \Theta_i = \Theta$.*
- *Two rules Θ_1 and Θ_2 are* mutually essential *in S_0 for a reagent $C \notin S_0$ iff*
  - *neither Θ_1 nor Θ_2 are essential in S_0 for C, and*
  - *$\forall$ χ-path $S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m$ in $Gr((R,S_0))$, leading to C, there exists at least an $i \in [0,m-1]$ s.t. $\Theta_i = \Theta_1$ or $\Theta_i = \Theta_2$.*

In this context, two χ-paths leading to C represent *vicarious* explanations if the χ-paths resort to different sets of rules.

Theorem 4.6 (Vicariate) *Given two χ-paths p_1 and p_2 in $Gr((R,S_0))$, leading to C*

$$p_1 = S_0 \xrightarrow{\Theta^1_0,\hat{S}^1_0,X^1_0} S^1_1 \xrightarrow{\Theta^1_1,\hat{S}^1_1,X^1_1} S^1_2 \ldots \xrightarrow{\Theta^1_{h-1},\hat{S}^1_{h-1},X_{h-1}} S^1_h$$

$$p_2 = S_0 \xrightarrow{\Theta^2_0,\hat{S}^2_0,X^2_0} S^2_1 \xrightarrow{\Theta^2_1,\hat{S}^2_1,X^2_1} S^2_2 \ldots \xrightarrow{\Theta^2_{m^2-1},\hat{S}^2_{k-1},X_{k-1}} S^2_k$$

p_1 is vicarious *of p_2 iff either $h \neq k$ or there exists at least a j s.t. $\Theta^1_j \notin \bigcup_{0\leq i<h}\{\Theta^2_i\}$.*

To prove checkpoint properties, we exploit the information recorded in $\hat{S}$, in order to check whether a certain metabolite B is *necessary* in the production of a reagent C.

Theorem 4.7 (Checkpoint) *Given R and an initial solution S_0, reagent B is* necessary *for the production of C if for all χ-path $S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} \ldots S_{m-1} \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m$ in $Gr((R,S_0))$, leading to C, then (i) $B \in S_{m-1}$ and (ii) if $B \in S_0$, then $B \in (\hat{S}_0 \cup \ldots \cup \hat{S}_{m-1})$.*

Conditions (i) and (ii) amount to saying that there is a rule Θ_i that has B in its premises.

Example 4.8 *Consider the initial solution $S_0 = \{A,B,D,O\}$ and an m-network R, consisting of the rules reported below on the left, while the corresponding CGF specification is on the right.*

$$\begin{array}{lcl}
(\lambda,\mu) & A \circ B & \rightarrow C \\
(\delta,\eta) & C \circ F & \rightarrow P \\
(\beta,\gamma) & D \circ A & \rightarrow F \\
(\xi,\theta) & B \circ D & \rightarrow H \\
(\psi,\nu) & D \circ H & \rightarrow E \\
(\sigma,\rho) & L \circ O & \rightarrow H \\
(\phi,\pi) & E \circ H & \rightarrow L \\
(o,\iota) & C \circ D & \rightarrow O \\
(\alpha,\zeta) & P \circ O & \rightarrow E
\end{array}$$

$$\begin{array}{l}
A = a^{\lambda}.C + \overline{c}^{\gamma}.0 \\
B = \overline{a}^{\mu}.0 + d^{\xi}.H \\
C = b^{\delta}.P + g^{o}.O \\
D = c^{\beta}.F + \overline{d}^{\theta}.0 + e^{\psi}.E + \overline{g}^{\iota}.0 \\
E = h^{\pi}.L \\
F = \overline{b}^{\eta}.0 \\
H = \overline{e}^{\nu}.0 + \overline{h}^{\pi}.0 \\
L = f^{\sigma}.H \\
O = \overline{f}^{\rho}.0 + \overline{l}^{\zeta}.0 \\
P = l^{\alpha}.E
\end{array}$$

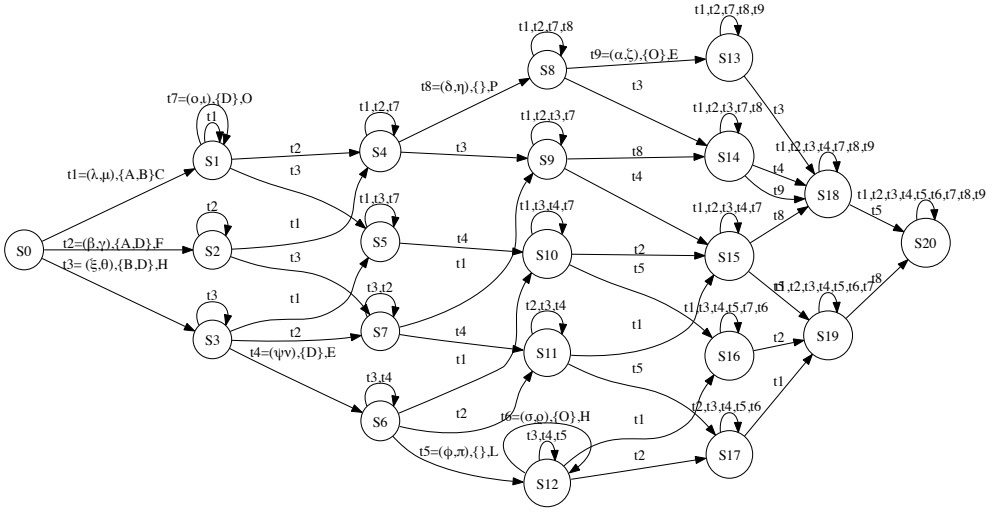

Figure 2: LTS Graph of Ex. 4.8

Figure 2 depicts the LTS semantics where $S_0 = \{A,B,D,O\}$, $S_{20} = \top$, *and*

$S_1 = S_0 \cup \{C\}$ $S_2 = S_0 \cup \{F\}$ $S_3 = S_0 \cup \{H\}$ $S_4 = S_1 \cup \{F\}$ $S_5 = S_3 \cup \{C\}$ $S_6 = S_3 \cup \{E\}$
$S_7 = S_3 \cup \{F\}$ $S_8 = S_4 \cup \{P\}$ $S_9 = S_4 \cup \{H\}$ $S_{10} = S_5 \cup \{E$ $S_{11} = S_6 \cup \{F\}$ $S_{12} = S_6 \cup \{L\}$
$S_{13} = S_8 \cup \{E\}$ $S_{14} = S_8 \cup \{H\}$ $S_{15} = S_9 \cup \{E\}$ $S_{16} = S_{10} \cup \{L\}$ $S_{17} = S_{11} \cup \{L\}$ $S_{18} = S_{13} \cup \{H\}$
$S_{19} = S_{16} \cup \{F\}$

We can observe the following properties.

• *The production of* H *is* necessary *for that of* L, *indeed* $H \cap S_0 = \emptyset$ *and all the states containing* L *come after states that include* H.

• *Rule* (ξ,θ) *is* essential *for the production of* H. *Actually, also rule* (σ,ρ) *is able to produce* H, *but it requires the presence of* L *that in turn requires* H *to be produced, as discussed before.*

• *Neither rule* (ψ,ν), *nor rule* (α,ζ) *is* essential *for the production of* E, *because there exists a* χ*-path* p (p') *leading to* E, *which does not use rule* (ψ,ν) ((α,ζ), *resp.):* $p = S_0 \xrightarrow{t_2} S_2 \xrightarrow{t_1} S_4 \xrightarrow{t_8} S_8 \xrightarrow{t_9} S_{13}$, $p' = S_0 \xrightarrow{t_3} S_3 \xrightarrow{t_4} S_7$. *Nevertheless, rules* (ψ,ν) *and* (α,ζ) *are* mutually essential. *As expected,* p *and* p' *represent therefore two alternative ways of producing* E. *The corresponding explanations:* $E_{(\alpha,\zeta)}[P_{(\delta,\eta)}[C_{(\lambda,\mu)}[A[],B[]],F_{(\beta,\gamma)}[D[],A[]]],O[C_{(\lambda,\mu)}[A[],B[]],D[]]]$ *(for* p*) and* $E_{(\psi,\nu)}[D[],H_{(\xi,\theta)}[B[],D[]]]$ *(for* p'*) represent two* vicarious *explanations for* E.

• *Note that if we* exclude A *from the initial solution, we cannot produce* F, *because rule* (β,γ) *could not be applied, but we still have a way to produce* E.

We now characterize the causality and the robustness properties in our computational framework. We say that a rule *causes* another rule, whenever the second one is always preceded by the first one.

Definition 4.9 (Causality) *Let* Θ,Θ' *two rules used in* $Gr((R,S_0))$. *The rule* Θ causes Θ' ($\Theta \sqsubset \Theta'$) *in* $Gr((R,S_0))$ *iff for all* χ*-path in* $Gr((R,S_0))$ $p = S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m$, $(\Theta' = \Theta_j) \Rightarrow \exists \Theta_i = \Theta$ *with* $i < j$.

Robustness has to do with the capacity of a network to produce a certain metabolite.

Theorem 4.10 (Strong and Weak Robustness) *Given two environments R_1 and R_2,*

- *$R_1 \ll_{S,P} R_2$ for C, iff for all χ-path $p \in Gr((R_1,S_0))$ leading to C, $p \in Gr((R_2,S_0))$ and leads to C.*
- *$R_1 \prec_{S,P} R_2$ for C iff for all χ-path $p \in Gr((R_1,S_0))$ leading to C, then there exists a χ-path $p' \in Gr((R_2,S_0))$ leading to C.*

Finally, we characterize the redundancy and the exclusion properties. Both are related with the role of initial metabolites and the possibilities of the network to produce metabolites, in case of modifications of the initial set. To this aim we resort to ρ-paths and to the following notion, that given a ρ-path p, computes the subset of metabolites $\mathscr{U}(p)$ of the initial solution strictly required to perform each transition of the given path. Such information is obtained by collecting all the subsets $\hat{S}_i$ (i.e. the subsets of the initial solution S_0 used by the transitions in p), and by removing those metabolites (in S_0) that the system itself is able to produce along the path. Recall that since p is a ρ-path, we are guaranteed that the transitions that produce these metabolites always come before the transitions that use them.

Definition 4.11 *Given a ρ-path in $Gr((R,S_0))$, $p = S_0 \xrightarrow{\Theta_0,\hat{S}_0,X_0} S_1 \xrightarrow{\Theta_1,\hat{S}_1,X_1} S_2 \ldots S_{m-1} \xrightarrow{\Theta_{m-1},\hat{S}_{m-1},X_{m-1}} S_m$*
$\mathscr{U}(p) = (\bigcup_{0\leq i<m} \hat{S}_i \setminus \bigcup_{0\leq i<m} \{X_i\})$

An initial solution is *redundant* for the production of a metabolite C, whenever there exists at least a component that is not required from the very beginning, in *all the paths* that lead to C. Moreover, to produce a metabolite C, we can *exclude* a metabolite B from the initial solution S_0, if B is not required from the very beginning, in *at least one path* that leads to C.

Theorem 4.12 (Redundancy and Exclusion) *Given an environment R, an initial solution S_0, and a metabolite $C \notin S_0$, let $\mathscr{P}_C = \{p \mid p \text{ is } \rho\text{-path in } Gr((R,S_0)) \text{ that leads to } C\}$. Then*

- *S_0 is* redundant *for C iff $\bigcup_{p\in\mathscr{P}_C} \mathscr{U}(p) \subset S_0$.*
- *a metabolite B* can be excluded *for the production of C iff $\exists$ a ρ-path p in $\mathscr{P}_C$ s.t. $B \notin \mathscr{U}(p)$.*

Example 4.13 *Consider again the network described in Ex. 4.8.*
• The initial solution $S_0 = \{A,B,D,O\}$ is redundant *for the production of E. Consider indeed the ρ-path leading to E: $p_1 = S_0 \xrightarrow{t_1} S_1 \xrightarrow{t_7} S_1 \xrightarrow{t_1} S_4 \xrightarrow{t_8} S_8 \xrightarrow{t_9} S_{13}$. Note that p_1 is similar to the χ-path p, seen in Ex. 4.8, except that it also includes the self-loop transition t_7 (rule (o,ι)) on the state S_1. This transition corresponds to a reaction that produces O, which is already in S_0, but that it is not required until this point. Therefore O could safely be* excluded *from S_0, since $O \notin \mathscr{U}(p_1)$. Similarly, we can prove that $O \notin \mathscr{U}(p)$ for all the other paths that reach S_{13} and all its successors.*
• Consider again the χ-path $p' = S_0 \xrightarrow{t_3} S_3 \xrightarrow{t_4} S_7$, leading to E. Note that p' is also a ρ-path and that $O \notin \mathscr{U}(p')$. The same result holds for all the paths reaching S_{13} and therefore the successor states. Hence, by Theorem 4.12 we can conclude that the metabolite O could safely be excluded *from the initial solution without compromising the production of the metabolite E. If we are not interested in maintaining all the ways to produce E, but just the general ability of the system to produce it, we can* exclude *A, since p' is a ρ-path leading to E and $A \notin \mathscr{U}(p')$.*
• Note that in this case, checkpoint and exclusion properties rely on the same information: we could have detected indeed that A could be excluded *from the fact that A was not* necessary *for the production of E. However, this is not true in general. Assume, e.g., to modify rule (ψ,ν) in order to require the presence of O, as $(\psi,\nu)' : O+H \to E$. As a consequence, also the paths leading to state S_7 require the presence of O, making also O* necessary *for the production of E. However, we can conclude that while S_0 is not*

(1)	β-D-Glucose $\circ$ ATP	$\rightarrow$	β-D-Glucose-6P $\circ$ ADP
(2)	β-D-Glucose-6P	$\rightarrow$	β-D-Fructose-6P
(3)	β-D-Fructose-6P $\circ$ ATP	$\rightarrow$	β-D-Fructose-1,6bP $\circ$ ADP
(4)	β-D-Fructose-1,6bP	$\rightarrow$	Glyceraldehyde-3-P $\circ$ Dihydroxyacetonephosphate
(5)	Glyceraldehyde-3-P	$\rightarrow$	Dihydroxyacetonephosphate
(6)	Dihydroxyacetonephosphate	$\rightarrow$	Glyceraldehyde-3-P
(7)	Glyceraldehyde-3-P $\circ$ NAD	$\rightarrow$	1,3 Bisphosphoglycerate $\circ$ NADH
(8)	1,3 Bisphosphoglycerate $\circ$ ADP	$\rightarrow$	3-Phosphoglycerate $\circ$ ATP
(9)	3-Phosphoglycerate	$\rightarrow$	2-Phosphoglycerate
(10)	2-Phosphoglycerate	$\rightarrow$	Phosphoenol pyruvate
(11)	Phosphoenol pyruvate $\circ$ ADP	$\rightarrow$	Pyruvate $\circ$ ATP
(12)	β-D-Glucose $\circ$ $NADP^+$	$\rightarrow$	D-Glucono-1,5-Lactone-6P $\circ$ NADPH
(13)	D-Glucono-1,5-Lactone-6P	$\rightarrow$	6-Phospo-D-Gluconate
(14)	6-Phospo-D-Gluconate $\circ$ $NADP^+$	$\rightarrow$	Ribulose-5-P $\circ$ NADPH
(15)	Ribulose-5-P	$\rightarrow$	D-Xylulose-5-P
(16)	Ribulose-5-P	$\rightarrow$	D-Ribose-5P
(17)	D-Ribose-5P $\circ$ D-Xylulose-5P	$\rightarrow$	Glyceraldehyde-3-P $\circ$ D-sedoeptulose-7-P
(18)	D-sedoeptulose-7-P $\circ$ Glyceraldehyde-3-P	$\rightarrow$	D-Erythrose-4P $\circ$ D-Fructose-6-P
(19)	D-Erythrose-4P $\circ$ D-Xylulose-5-P	$\rightarrow$	Glyceraldehyde-3-P $\circ$ β-D-Fructose-6P

Table 2: Rules of the Glycolytic Pathway and of the Pentose Phosphate Pathway

redundant *for the production of E in the modified system, considering p' above, O could be* excluded*, since throughout p' the modified system is still able to produce E.*

• *Finally note that $(\lambda,\mu) \sqsubseteq (\phi,\pi)$, $(\beta,\gamma) \sqsubseteq (\phi,\pi)$ while neither $(\lambda,\mu) \sqsubseteq (\beta,\gamma)$ nor $(\beta,\gamma) \sqsubseteq (\lambda,\mu)$. Indeed, the transitions related to the application of rules (λ,μ) and (β,γ) (t_1 and t_2 resp.) are not causally related, hence, they can be fired in any order.*

5 Properties at work in a metabolic pathway

A precise characterization of the structural role played by the single elements in the overall metabolic networks is relevant both for better understanding living systems and for developing treatments for pathological aspects. As an example consider the clinical studies of primary and metastatic cancers that have clearly demonstrated that human malignancies are characterized by an increased activity of glycolysis when compared to normal tissue [17]. This metabolic peculiarity suggests an inviting target for cancer treatment and various therapeutic strategies aiming at selectively disrupting glycolytic network of malignant cells are under investigation [18].

In this light, we present a simplified glycolytic pathway embedded in a wider context comprising also the Penthose Phosphate Pathway. Through these interconnected pathways the β-D-Glucose-6P is oxidized yielding Pyruvate and energy (ATP) or Ribose and reducing equivalents (NADPH).

The pathway can be formalized as in Tab. 2. For lack of space we do not show here the corresponding LTS graph, however it should be clear how our properties, related with very important biological features, can be verified using the method illustrated in § 4 (see in particular Ex. 4.8 and 4.13).

Reaction properties Suppose that our initial solution is S_α: $\{\beta$-D-Glucose, ATP, $NADP^+$, NAD$\}$; we can verify the following properties.

• The existence of the following causal chains of rules: $(1) \sqsubseteq (2) \sqsubseteq ... \sqsubseteq (11)$ and $(12) \sqsubseteq ... \sqsubseteq (15)$.

• Rule (7) is *essential* for the production of the metabolite *Pyruvate*. Its exclusion interrupts all the possible paths reaching *Pyruvate*.
• Rule (12) is *essential* for the production of various metabolites, e.g. *NADPH* and *D-Erythrose-4P*.
• The metabolite *D-Glucono-1,5-Lactone-6P* is a *checkpoint* for the production of NADPH, produced both by the rules (12) and (14), and for that of *D-Xylulose-5-P*, produced both by the rules (15) and (17). The rule (12) which produces *D-Glucono-1,5-Lactone-6P* corresponds to the reaction catalyzed by the enzyme Glucose-6-Phosphate Dehydrogenase(G6PD), an enzymopathy commonly known as fauvism.
• The metabolite *Glyceraldehyde-3-P* can be produced either by the χ-path composed by the transitions corresponding to the rules: (12), (13), (14), (15), (16), (17), or by the χ-path composed by the transitions corresponding to the rules: (1), (2), (3), (4). The two paths correspond to two *vicarious* explanations.
• The metabolite $NADP^+$ should be *included* in the initial solution in order to produce *D-Xylulose-5-P*, but it can be *excluded* as far as the production of β*-D-Fructose-1,6bP* is concerned.
• Finally, having the initial solution $S_\beta = S_\alpha \cup \{$*Glyceraldehyde-3-P*$\}$, the rules (4) and (5) are *mutually essential* for *Pyruvate* as they must be both removed in order to suppress the Dihydroxyacetonephosphate production. Another example of mutually essential rules is given in [5].

Network properties In order to illustrate our definition of robustness, we consider two pathways: the pathway described above and another one, obtained from the first one by suppressing the two reactions, one inverse of the other, represented by rules (5) and (6). This suppression corresponds to the inhibition of an enzyme, that is related to a severe disease, known as triosephosphate isomerase (TPI) deficiency, see [31] for details. Considering the standard solution S_α, it is easy to verify that the glycolytic pathway results to be more robust than its variant related to the disease with respect to the production of *Pyruvate*: only one of the explanations for *Pyruvate* existing in the original network is viable in the second one. This simple example well highlights the relevance that a study of robustness may have. Quite naturally, this notion can be extended in order to consider robustness with respect to different solutions or with respect to different metabolites of the same network. About the latter, intuitively, it turns out that there are at least two explanations for metabolites like *Glyceraldehyde-3-P*, *2-Phosphoglycerate* and *Pyruvate*, given the solution S_α in the glycolytic pathway. Instead, only one explanation exists for metabolites like β*-D-Glucose-6P* or. Therefore, the network results more robust for the production of *Glyceraldehyde-3-P, 2-Phosphoglycerate* and *Pyruvate* rather than for that of β*-D-Glucose-6P*. From a drug research point of view, targeting the parts of the network involved in the production of the last two metabolites may result more effective than targeting the others. Indeed, a drug targeting the reaction of hexokinase, leading to the production of β*-D-Glucose-6P* is under development [18].

6 Conclusions

We have presented a taxonomy of biological properties of interest regarding metabolic network. Based on a (formal) notion of *causality*, this taxonomy translates a bunch of properties in use within biologists into a formal framework. We have also proposed a computational counterpart of the framework, which, allowing the automated verification of the properties, paves the way to the development of software tools supporting the analysis of metabolic networks. We have chosen a reading of causality and the computational mechanisms that rely on theories developed in concurrency and particularly suitable to describe causality in interactive behaviours and providing a wealth of analysis techniques. Definitions do not depend on the computational framework, and this can be changed whenever another computational support may result more convenient for a specific domain or set of (causally-based) properties.

Future work regards the extension of the set of proposed properties, experimentation with case-

studies of interest for wet-lab research, possibly contrasting the framework with other analogous proposals especially as far as the trade-off between expressiveness and efficiency is concerned. Moreover, we would like to attempt a characterization of properties of sets of interconnected signaling pathways, like the ones involved in cancerogenesis, since the understanding of the structural features underlying their interactions may provide useful hints for drug research. In this sense, it could be worth studing possible integrations of our framework with the qualitative logical view adopted in [3] for signaling networks.

References

[1] R. Albert, H. Jeong & A. Barabasi (2000): *Error and attack tolerance of complex networks*. Nature 406.

[2] U. Alon (2007): *Network motifs: theory and experimental approaches*. Nature reviews Genetics 8.

[3] C. Baral, W. Kolch, C. Shankland & N. Tran (2005): *Reasoning about the ERK signal transduction pathway using BioSigNet-RR*. In: Proc. of CMSB'05.

[4] C. Bodei (2009): *A Control Flow Analysis for Beta-binders with and without static compartments*. Theoretical Computer Science 410(33-34).

[5] C. Bodei, A. Bracciali & D. Chiarugi (2008): *On deducing causality in metabolic networks*. BMC Bioinformatics 9 (4).

[6] L. Cardelli (2005): *Brane calculi*. In: Proc. of Computational Methods in Systems Biology (CMSB'04), LNCS 3082. Springer, pp. 257–280.

[7] L. Cardelli (2008): *On process rate semantics*. Theoretical Computer Science 391(3), pp. 190–215.

[8] N. Chabrier, F. Fages & S. Soliman (2005): *The Biochemical abstract machine BIOCHAM*. In: Proc. of CMSB'04, LNCS 3082. Springer, pp. 172–191.

[9] R.G. Clyde, J.L. Bown, T.R. Hupp, N. Zhelev & J.W. Crawford (2006): *The role of modelling in identifying drug targets for diseases of the cell cycle*. J R Soc Interface 3, pp. 617–627.

[10] A. Coletta, R. Gori & F. Levi (2009): *Approximating Probabilistic Behaviors of Biological Systems Using Abstract Interpretation*. In: Proc. of FBTC'08, ENTCS 229(1). pp. 165–182.

[11] M. Curti, P. Degano, C. Priami & C.T. Baldari (2004): *Modeling biochemical pathways through enhanced pi-calculus*. Theoretical Computer Science 325(1), pp. 111–140.

[12] V. Danos & J. Krivine (2005): *Transactions in RCCS*. In: Proc. of CONCUR'05, LNCS 3653. Springer.

[13] V. Danos & C. Laneve (2003): *Graphs for Core Molecular Biology*. In: Proc. of CMSB'03, LNCS 2602. Springer, pp. 34 – 46.

[14] S. Eker, M. Knapp, P. Lincoln, K. Laderoute & C. Talcott (2002): *Pathway Logic: Executable Models of Biological Network*. In: Proc. of the Fourth International Workshop on Rewriting Logic and Its Applications (WRLA'02), ENTCS 71. Elsevier, pp. 144–161.

[15] S. Fatumo, K. Plaimas, J.P. Mallm, G. Schramm, E. Adebiyi, M. Oswald, R. Eils & R. Knig (2009): *Estimating novel potential drug targets of Plasmodium falciparum by analysing the metabolic network of knock-out strains in silico*. Infect. Genet. Evol. 9(3), pp. 351–8.

[16] D.A. Fell (1997): *Understanding the control of metabolism*. Portland Press, London, United Kingdom.

[17] S.S Gambhir (2002): *Molecular imaging of cancer with positron emission tomorgraphy*. Nat. Rev. Cancer 2, pp. 891–899.

[18] R.A. Gatenby & R.J. Gillies (2007): *Glycolysis in cancer: A potential target for therapy*. The International Journal of Biochemistry and Cell Biology 39, pp. 1358–1366.

[19] S. Gerdes, R. Edwards, M. Kubal, M. Fonstein, R. Stevens & A. Osterman (2006): *Essential genes on metabolic maps*. Current Opinion in Biotechnology 17, pp. 448–456.

[20] G.T. Gillespie (2000): *The chemical Langevin equation*. Journal of Chemical Physics 113(1), pp. 297–306.

[21] R. Gori & F. Levi (2009): *Abstract Interpretation based Verification of Temporal Properties for BioAmbients.* To appear in Info &Co.

[22] R. Gori & F. Levi (2009): *Abstract Interpretation for Probabilistic Termination of Biological Systems.* In: *Proc. of MeCBIC'09*, EPTCS 11.

[23] H Kitano (2002): *Systems Biology: a brief overview*. *Science* 295(5560), pp. 1662–1664.

[24] H. Kitano (2004): *Biological robustness*. *Nat Rev Genet* 5, pp. 826–837.

[25] H. Kitano (2007): *Towards a theory of biological robustness*. *Molecular Systems Biology* 3, p. 137.

[26] S. Klamt & E.D. Gilles (2002): *Minimal cut sets in biochemical reaction networks*. *Nature* 420.

[27] R. Milner (1999): *Communicating and mobile systems: the π-calculus*. Cambridge University Press.

[28] F. Nielson, H. Riis-Nielson, D. Schuch-Da-Rosa & C. Priami (2004): *Static analysis for systems biology*. In: *Proc. of workshop on Systeomatics-dynamic biological systems informatics.* Computer Science Press, Trinity College Dublin,, pp. 1–6.

[29] F. Nielson, H. Riis-Nielson, D. Schuch-Da-Rosa & C. Priami (2007): *Control Flow Analysis for BioAmbients*. *ENTCS* 180(3), pp. 65–79.

[30] Y-K. Oh, B.. Palsson, S.M. Park, C.H. Schilling & R. Mahadevan (2007): *Genome-scale Reconstruction of Metabolic Network in Bacillus subtilis Based on High-throughput Phenotyping and Gene Essentiality Data*. *TJ Biol Chem* 10, pp. 693–706.

[31] F. Orosz, J. Olh & J. Ovdi (2006): *Triosephosphate isomerase deficiency: facts and doubts. IUBMB life* 12, pp. 703–715.

[32] A. Phillips & L. Cardelli (2004): *A correct abstract machine for the stochastic pi-calculus*. In: *Proc. of Bioconcur'04.* ENTCS, Elsevier.

[33] H. Pilegaard, F. Nielson & H. Riis Nielson (2008): *Pathway analysis for BioAmbients*. *J. Log. Algebr. Program.* 77(1-2), pp. 92–130.

[34] H. Pilegaard, H. Riis Nielson & F. Nielson (2006): *Static Analysis of a Model of the LDL Degradation Pathway*. In: *Simulation and Verification of Dynamic Systems.*

[35] C. Priami (1995): *Stochastic π-calculus*. *The Computer Journal* 38, pp. 578–589.

[36] C. Priami & P. Quaglia (2005): *Beta Binders for Biological Interactions*. In: *Proc. of CMSB'04*, LNCS 3082. Springer, pp. 20–33.

[37] C. Priami, A. Regev, E. Shapiro & W. Silvermann (2001): *Application of a stochastic name-passing calculus to representation and simulation of molecular processes*. *Inf. Process. Lett.* 80(1), pp. 25–31.

[38] A. Regev, E. Panina, W. Silverman, L. Cardelli & E. Shapiro (2004): *Bioambients: An abstraction for biological compartements*. *Theoretical Computer Science* 325(1), pp. 141–167.

[39] A. Regev, W. Silvermann & E. Shapiro (2001): *Representation and Simulation of Biochemical Processes Using the pi-Calculus Process Algebra*. In: *Pacific Symposium on Biocomputing.* pp. 459–470.

[40] C.H. Schilling, D. Letscher & B.O. Palsson (2000): *Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective*. *J. Theor. Biol.* 203.

[41] S. Schuster, D. Fell & T. Dandekar (2000): *A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks*. *Nat. Biotechnol* 18, pp. 226–232.

[42] S. Schuster, D. Fell & T. Dandekar (2002): *Metabolic network structure determines key aspects of functionality and regulation*. *Nature* 420, pp. 190–193.

[43] B. Shargel, H. Sayama, I.R. Epstein & Y. Bar-Yam (2003): *Optimization of Robustness and Connectivity in Complex Networks*. *Physical Review Letters* 90.

[44] J. Stelling, U. Sauer, Z. Szallasi, F.J. Doyle III & J. Doyle (2004): *Robustness of cellular functions*. *Cell* 118, pp. 675–685.

[45] J. Thykaer & J. Nielsen (2003): *Metabolic engineering of beta-lactam production. Metabolic Eng.* 5 (1).

[46] B.J. Yu, B.H. Sung, J.Y. Lee, S.H. Son, M.S. Kim & S.C. Kim (2006): *sucAB and sucCD are mutually essential genes in Escherichia coli*. FEMS Microbiol Lett 254, pp. 245–250.

[47] H. Yu, P.M. Kim, E. Sprecher, V. Trifonov & M. Gerstein (2009): *The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics.* PLOS Computational Biology 3 (4).