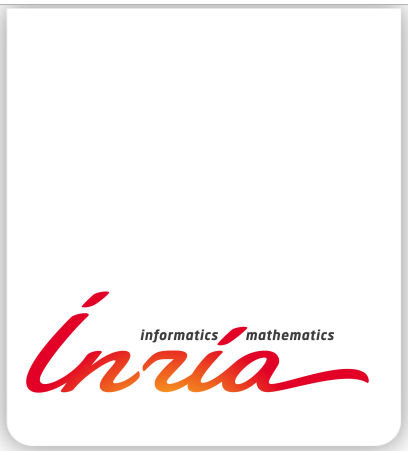
Automatic Curation of SBML Models based on their ODE Semantics

François Fages, Steven Gay, and Sylvain Soliman

RESEARCH REPORT
N° 8014
July 2012
Project-Team Contraintes

ISSN 0249-6399 ISRN INRIA/RR--8014--FR+ENG



Automatic Curation of SBML Models based on their ODE Semantics

François Fages*, Steven Gay, and Sylvain Soliman

Project-Team Contraintes

Research Report n° 8014 — July 2012 — 16 pages

Abstract: Many models in Systems Biology are described as a system of Ordinary Differential Equations. The fact that the Systems Biology Markup Language SBML has become a standard for sharing and publishing models, has helped in making modelers formalize the structure of the reactions and use structure-related methods for reasoning about models. Unfortunately, SBML does not enforce any coherence between the structure and the kinetics of a reaction. Therefore the structural interpretation of models transcribed in SBML may vary according to different choices of representation of the original model and may be incorrect for some analyses.

The first contribution of this paper is to propose a general compatibility condition between the kinetic expression and the structure of a reaction. We show that these well-formedness conditions are satisfied by standard kinetics and that they entail a property of independence from the kinetic expressions for the influence graph associated to the ODEs. We present a heuristic algorithm of low computational complexity for, given an ODE system, inferring a reaction model that preserves the ODE semantics and infers well-formed reactions whenever possible. This algorithm can be used for not only checking whether the network and ODE structures of an SBML model are consistent but also automatically curating SBML models by exporting them as ODE systems and then importing them as well-formed reaction models. We show how this strategy is capable of automatically curating SBML models on a large scale and provide some statistics figures obtained on the whole biomodels.net repository.

The algorithms described in this paper are implemented in the open-source software modeling platform BIOCHAM [Fages and Soliman, 2008a, Calzone et al., 2006] available at `http://contraintes.inria.fr/biocham` The models used in the experiments are available from `http://www.biomodels.net/`

Key-words: systems biology, ordinary differential equations, reaction model

* to whom correspondence should be addressed

RESEARCH CENTRE
PARIS – ROCQUENCOURT

Domaine de Voluceau, - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex

Curation automatique de modèles SBML basée sur la sémantique d'ODE

Résumé : Le format SBML est devenu un format standard de publication de modèles en biologie systémique. Malheureusement, les modèles SBML sont souvent encodés en ignorant la structure en réactions pourtant requise par les concepteurs du format.

Cet article présente une formalisation d'un modèle de réactions bien formé, et un algorithme heuristique qui infère un modèle de réactions si possible bien formé, permettant la curation automatique de modèles SBML.

Mots-clés : biologie systémique, équations differentielles ordinaires, modèle de réactions

1 Introduction

Many models in Systems Biology are described as a system of Ordinary Differential Equations (ODEs), which allows for transient and steady-state analysis via numerical integration (for instance using MATLAB®) or bifurcation analysis (with tools like XPPAUT [Ermentrout, 2002]), but only when kinetic information is available.

Complementary structure-related qualitative analysis techniques have become increasingly popular in recent years, such as qualitative model checking and pathway analysis [Eker et al., 2002, Fages et al., 2004] or Petri net analysis [Reddy et al., 1993, Zevedei-Oancea and Schuster, 2003, Angeli et al., 2007, Chaouiya et al., 2008, Rohr et al., 2010]. They do not rely on kinetic information, but require a structured model with well-identified products, reactants and modifiers (and in certain cases their stoichiometry) for each reaction.

The fact that the Systems Biology Markup Language (SBML) of [Hucka et al., 2003] has become a standard for sharing and publishing models has helped in making modelers formalize the reaction structure of their models. Unfortunately, SBML does not enforce any coherence between the structure and the kinetics of a reaction. Therefore the structural interpretation of models transcribed in formalisms such as SBML may vary according to different choices of representation of the original model. For instance, if the models were originally written as ODEs, a later discrete interpretation as a qualitative or stochastic model of their transcription in SBML may produce wrong results.

The first contribution of this paper is to propose a general compatibility condition between the kinetic expression and the structure of a reaction. We show that these well-formedness conditions are satisfied by standard kinetics such as mass action law, Michaelis-Menten, Hill and negative Hill kinetics. We also show that they entail a property of independence from the kinetics for the influence graph (Jacobian sign matrix) associated to well-formed reaction models.

In [Kaleta et al., 2009], it is elaborated that structural information hidden in kinetic laws may affect the results obtained from structural analyses, such as elementary mode analysis [Schuster et al., 2002], flux balance analysis [Varma and Palsson, 1994], chemical organization theory [Dittrich and di Fenizio, 2007], deficiency analysis or chemical reaction network theory [Feinberg, 1977, Shinar and Feinberg, 2010]. Likewise, the correct structure is mandatory when a reaction network must be interpreted as a stochastic process *à la* [Gillespie, 1977].

It is worth noticing that these structural analyses may directly support dynamic analyses: for instance, [Koh et al., 2006] apply network decomposition for a modular parameter estimation approach, [Angeli et al., 2007] introduce a structural persistence criterion, Petri net place invariants reveal conservation laws in [Soliman, 2008] and transition invariants are used in [Grafahrend-Belau et al., 2008] to identify fragile nodes and the core of a network, and in [Nabli and Soliman, 2010] to determine steady state solutions.

In [Kaleta et al., 2009], the authors present an algorithm that uncovers hidden structural information for some SBML models of the `biomodels.net` repository [le Novère et al., 2006]. The problem of finding a structured model for a given system of ODEs is not new. Actually for the restricted case of models with only Mass Action kinetics a general solution is provided in [Hárs and Tóth, 1979]. This approach was evolved over the years, see for instance [Szederkényi et al., 2011] for sparse/dense/core solutions when numerical values are provided for the parameters, or [Soliman and Heiner, 2010] for unicity conditions in the symbolic case.

The second contribution of this paper is to propose an algorithm for inferring a reaction model corresponding to an ODE system and satisfying our more general well-formedness conditions whenever possible. This algorithm, of low complexity, is shown to preserve the ODE semantics of the reactions and their well-formedness when it is applied to an ODE system coming from a non-decomposable well-formed reaction model.

Then our third contribution is to show that this algorithm can be used to automatically curate SBML models by exporting them as ODE systems and then importing them by inferring well-formed reactions. We show how this strategy is capable of curating automatically the writing in SBML of the models of the cell cycle in `biomodels.net`, and provide some statistics figures obtained on the whole `biomodels.net` repository.

2 Theory of Well-formed Reactions and Kinetics

2.1 Well-formedness Conditions

Let us consider a finite set $V = \{x_1, \ldots, x_v\}$ of v molecular species. A *reaction model* R is a finite set of n reactions, written

$$R = \{\ e_i \texttt{ for } \ r_i \ / \ m_i \texttt{ => } \ p_i \ \}_{i=1,\ldots,n}$$

where e_i is a formal mathematical expression over molecular species concentrations (possibly involving symbolic parameters), and r_i, m_i and p_i are multisets of molecular species which represent respectively, the reactants, the inhibitors, and the products of the reaction. The species that are both reactants and products in a reaction are called catalysts. For a multiset r of molecular species, i.e. a function $V \to \mathbb{N}$, we denote by $r(x)$ the multiplicity of x in r, i.e. $r(x) = 0$ if x does not belong to r, and $r(x) \geq 1$ if x belongs to r, which is also written $x \in r$. The empty multiset is written $\_$. A multiset r will also be sometimes denoted by the linear expression with integer stoichiometric coefficients $\sum_{i=1}^{m} r(x_i) * x_i$.

It is worth remarking that this syntax for reactions is compatible with SBML (and Biocham) reactions except for the distinction between catalysts and inhibitors which are just considered as "modifiers" in SBML annotations (or as catalysts in Biocham). However we find it useful for the theory here to distinguish between the activation or inhibitory effects of a modifier and mark it syntactically as such in the structure of the reactions.

This distinction does not affect the system of ordinary differential equations (ODE) associated to a reaction model given by the classical Reaction Rate Equation, i.e. for $1 \leq j \leq v$:

$$\dot{x_j} = \sum_{i=1}^{n} (p_i(x_j) - r_i(x_j)) * e_i$$

We are interested in mathematical conditions that express the compatibility between the kinetic expression and the structure of a reaction. Since we may want to express with one reaction the dynamics of a more complex system obtained by reduction, we do not content ourselves with elementary kinetic expressions but seek abstract compatibility properties that can be applied to any mathematical expression given as kinetics. This is in contrast to most work on chemical reaction network theory [Feinberg, 1977, Shinar and Feinberg, 2010] but in accordance with the use in SBML of MathML for writing the kinetic expressions without any limitation.

Let us call a *non-decomposable* term a mathematical expression that is syntactically not an addition nor a subtraction, and that cannot be reduced at top-level by the laws of distributivity of the product and division on addition and subtraction. For an expression, by positive (resp. negative) we mean positive or null (resp. negative or null).

Definition 2.1. *A reaction* e `for` r / m `=>` p *over molecular species* $\{x_1, \ldots, x_v\}$ *is* well-formed *if the following conditions hold:*

1. *e is a well-defined, positive and partially differentiable mathematical expression for any values $x_1 \geq 0, \ldots, x_v \geq 0$;*

2. $x \in r$ *if and only if* $\partial e/\partial x > 0$ *for some* $x_1 \geq 0, \ldots, x_v \geq 0$;

3. $x \in m$ *if and only if* $\partial e/\partial x < 0$ *for some* $x_1 \geq 0, \ldots, x_v \geq 0$;

The reaction is non-decomposable *if* e *is a non-decomposable term.*

The first condition expresses that kinetic expressions must be well-defined, differentiable and positive. It can be related to this sentence from the SBML specification [Hucka et al., 2008]: *"...labeling a reaction as irreversible is interpreted as an assertion that the rate expression will not have negative values during a simulation."* The second (resp. third) condition states that the partial derivative of e w.r.t. a reactant (resp. an inhibitor) must be strictly positive (resp. negative) for some positive values of concentrations. The non-decomposability condition excludes the composition of several reactions in a single one with a sum as kinetic expression.

Example 2.2. *A typical example of an ODE that does not correspond to any non-decomposable well-formed reaction model is the equation* $\dot{x} = -k$ *That ODE can be associated to, either the non-decomposable but non well-formed (since condition 2 is violated) reaction* `k for x => _`*, or to the well-formed but decomposable (since the kinetic expression of the first reaction is a sum) reaction model* `k+l*x for x => _` *and* `l*x for x => 2*x`.

Example 2.3. *Let us consider the following three reactions:*

```
k1*[pMPF]*[Cdc25] for pMPF + Cdc25 => MPF + Cdc25
k2*[MPF]*[Wee1]   for MPF + Wee1 => pMPF + Wee1
k3/(k4+[Clock])   for _ / Clock => Wee1
```

where $k1, k2, k3$ *are parameters.*

The ODE system associated to them is

$\dot{pMPF} = k2 * [MPF] * [Wee1] - k1 * [pMPF] * [Cdc25]$

$\dot{MPF} = k1 * [pMPF] * [Cdc25] - k2 * [MPF] * [Wee1]$

$\dot{Wee1} = k3/(k4 + [Clock])$ $\quad \dot{Cdc25} = 0$ $\quad \dot{Clock} = 0$

This reaction model is well-formed and non-decomposable. In particular, we have $\partial\big(k3/(k4 + [Clock])\big)/\partial[Clock] < 0$ *for the inhibitor Clock in the synthesis reaction for Wee1.*

Although currently not met in practice, as we will see in Section 4, the well-formedness conditions should not be restrictive. In particular, they are met by all standard kinetic expressions:

Proposition 2.4. *Reactions with mass action law kinetics:* $k * \prod_j x_j^{n_j}$ `for` $\sum_j n_j * x_j$ `=>` p*, Michaelis-Menten kinetics:* $V * \frac{x}{K+x}$ `for` x `=>` y*, Hill kinetics:* $V * \frac{x^n}{K^n+x^n}$ `for` x `=>` y*, or negative Hill kinetics:* $\frac{V}{K^n+x^n}$ `for` $\_/x$ `=>` y*, with rate constants* $k, V, K > 0$ *and exponent* $n \geq 1$*, are well-formed and non-decomposable.*

Interestingly, some other conditions which would be quite natural to impose are not necessary for the results presented in this paper. This is the case for instance of:

- (monotonicity) for any variable $x \in V$ $\partial e/\partial x$ has a constant sign for all $x_1 \geq 0, \ldots, x_v \geq 0$;
- (strict reactants) $e = 0$ whenever a reactant $x = 0$.

2.2 Properties of the ODE Influence Graph associated to a Well-formed Reaction Model

A influence graph between molecular species induced by the ODE semantics of a well-formed reaction model enjoys some strong properties. Although not necessary to the algorithm presented in the next section, we present these properties here as part of the theory of biochemical networks

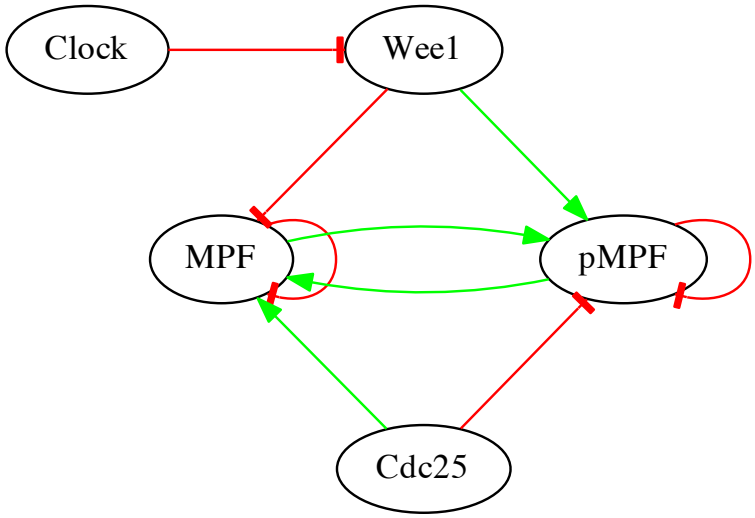
on which the algorithm is based. Formal influence graphs have been introduced in the setting of gene regulatory networks [Thomas et al., 1976] as a simple abstraction of complex regulation mechanisms. These graphs completely abstract from the precise interactions, especially at post-transcriptional level, and retain only the activation and inhibition effects on gene transcription. As conjectured in [Thomas, 1981], the existence of a positive circuit (resp. a negative circuit) in an influence graph has been proved to be a necessary condition for multistationarity, e.g. for cell differentiation, (resp. for oscillations, e.g. for homeostasis) in different formalisms, and in particular for ODE systems in [Kaufman et al., 2007, Soulé, 2006, Soulé, 2003, Snoussi, 1998, Gouzé, 1998].

In an ODE system, the influence graph is mathematically defined by the signs of the coefficients in the Jacobian matrix of the system:

Definition 2.5. *The* differential influence graph (DIG) *associated to a reaction model is the graph that has for vertices the molecular species, and for edge-set the following two kinds of edges:*

$$\{x \to^+ y \mid \partial\dot{y}/\partial x > 0 \text{ for some } x_1 \geq 0, \ldots, x_v \geq 0\}$$
$$\cup\{x \to^- y \mid \partial\dot{y}/\partial x < 0 \text{ for some } x_1 \geq 0, \ldots, x_v \geq 0\}$$

Example 2.6. *The DIG of example 2.3 is*



Definition 2.7. *The* stoichiometric influence graph (SIG) *associated to a finite set R of reactions is the graph that has for vertices the molecular species, and for edges the following set of positive and negative influences:*

$x \to^+ y$ if there exists a reaction i with $p_i(y) - r_i(y) > 0$ and $r_i(x) > 0$, or $p_i(y) - r_i(y) < 0$ and $m_i(x) > 0$,

$x \to^- y$ if there exists a reaction i with $p_i(y) - r_i(y) < 0$ and $r_i(x) > 0$, or $p_i(y) - r_i(y) > 0$ and $m_i(x) > 0$

Obviously, the SIG is trivial to compute, just by parsing the reactions, with a linear time complexity. As shown in [Fages and Soliman, 2008b], the SIG is an over-approximation of the DIG:

Theorem 2.8. *For any finite set R of well-formed reactions, the DIG is a subgraph of the SIG of R.*

This result can be generalized to an equivalence result with an extra assumption. Let us say that a tuple of molecular species (x, y) is in *conflict in an influence graph* if we have both $x \to^+ y$ and $x \to^- y$.

Theorem 2.9. *For any finite set R of well-formed reactions such that the SIG of R contains no conflict, the DIG and the SIG are identical.*

Proof. We just have to prove that the SIG is a subgraph of the DIG. Let us consider an arc $x \to^+ y$ in the SIG. There exists a reaction i with $p_i(y) - r_i(y) > 0$ and $r_i(x) > 0$, or $p_i(y) - r_i(y) < 0$ and $m_i(x) > 0$. Since the reaction is well-formed, we have either $p_i(y) - r_i(y) > 0$ and $\frac{\partial e_i}{\partial x} > 0$, or $p_i(y) - r_i(y) < 0$ and $\frac{\partial e_i}{\partial x} < 0$, for some $x_1 \geq 0, \ldots, x_v \geq 0$. Now, if $p_i(y) - r_i(y) > 0$, the term e_i has to occur in $\dot{y}$ with a positive sign, so $\frac{\partial e_i}{\partial x} > 0$. Furthermore, since there is no conflict in the SIG, we get $\frac{\partial \dot{y}}{\partial x} > 0$, i.e. $x \to^+ y$ is in the DIG. Similarly, if $p_i(y) - r_i(y) < 0$, the term e_i has to occur in $\dot{y}$ with a negative sign, thus $\frac{\partial e_i}{\partial x} < 0$ and $\frac{\partial \dot{y}}{\partial x} > 0$, i.e. $x \to^+ y$ is in the DIG. The proof for an arc $x \to^- y$ in the SIG is symmetrical. □

Corollary 2.10. *The DIG of a finite set of well-formed reactions without conflict in its SIG, is independent of the kinetic expressions.*

Corollary 2.11. *The DIG of a finite set of well-formed reactions without conflict in its SIG, is computable in linear time in the number of reactions.*

Example 2.12. *The SIG of Example 2.3 is trivial to compute and since it contains no conflict, we can predict that it is identical to its DIG depicted in Figure 2.6. In the simplified model of the yeast cell cycle of [Tyson, 1991], the double activation reactions of MPF through Cdc25 and Wee1 are simplified in a single autocatalytic reaction in parallel with a deactivation reaction:*

```
pMPF+MPF => 2*MPF      MPF => pMPF
```

Such reactions create a conflict in the SIG, namely $MPF \to^- pPMF$ and $MPF \to^+ pMPF$. In general, there is a possibility that such conflicting direct influences in the SIG may be balanced in the ODEs and may not appear in the DIG. This situation is however quite pathological and rare in practice and occurs when oversimplifications are made. For instance, the map of the cell cycle control of [Kohn, 1999] which contains 800 reactions does not contain any conflict in its SIG [Fages and Soliman, 2008b].

3 Automatic Curation Method of SBML Models

The principle of our automatic curation method for SBML models implemented in Biocham is to import the SBML reactions, export their ODE semantics in ODE format, import the ODE format as Biocham reactions, and export them in SBML, as follows: `load_sbml(file); export_ode(file); load_ode(file); export_sbml(file).` The cleverness of the method lies in the load ODE command which infers reactions for any ODE system by inferring well-formed reactions whenever possible. As we will see in the following sections on evaluation, this method has the effect of automatically curating the writing of models in SBML format.

3.1 Inference Algorithm for Reactions

Our algorithm for inferring reactions from ordinary differential equations is based on a syntactical normal form for ODE systems which facilitates the recognition of common subterms in the equations.

Definition 3.1. *A formal mathematical expression is in* additive normal form *if it is of the form* $\sum_{s=1}^{t} c_s * f_s$ *where* c_s *are numerical coefficients and* f_s *are non-decomposable terms without numerical coefficients.*

An ODE system is in additive normal normal form if each equation is in additive normal normal form as follows $\dot{x_i} = \sum_{s=1}^{t} c_{i,s} * f_s,\ 1 \leq i \leq v$ *where* t *is the number of non-decomposable terms in the system.*

Obviously, additive normal forms are not unique but any ODE system can be written in additive normal form through standard algebraic transformations. Now, given an ODE system, we can normalize it and infer a corresponding reaction model by sorting the terms of the equations and giving their coefficients as stoichiometric coefficients, as follows:

Algorithm 3.2. input: *ODE system* O

1. $O \leftarrow$ *additive-normal-form*(O)
2. $R \leftarrow \emptyset$
3. *for each non-decomposable term* f *of an equation in* O
   (a) *let* $r \leftarrow \_$, $p \leftarrow \_$, $m \leftarrow \_$
   (b) *for each variable* x *where* f *occurs with coefficient* c *in* $\dot{x}$ *in* O
      i. *if* $c < 0$ *then* $r(x) \leftarrow -c$
      ii. *if* $c > 0$ *then* $p(x) \leftarrow c$
   (c) *for each variable* x *such that* $r(x) = 0$ *and* $\frac{\partial f}{\partial x} > 0$ *for some values*
      i. $r(x) \leftarrow 1$
      ii. $p(x) \leftarrow p(x) + 1$
   (d) *for each variable* x *such that* $\frac{\partial f}{\partial x} < 0$ *for some values*
      i. $m(x) \leftarrow 1$
   (e) $R \leftarrow R \cup \{f$ `for` r / m `=>` $p\}$

output*: reaction model* R

Testing the sign of a partial derivative $\frac{\partial f}{\partial x}$ may involve arbitrary complex operations. In an implementation we shall content ourselves with an approximate test such as comparing the exponents of x in the numerator and denominator of f. With these restrictions, we have

Proposition 3.3 (Complexity)**.** *On an ODE system in additive normal form, Algorithm 3.2 computes a reaction model in time* $O(v * t)$*, where* v *is the number of variables and* t *is the number of non-decomposable terms in the system.*

In mathematical terms, the result of the algorithm is given by

Proposition 3.4 (Inferred reactions)**.** *The reaction model inferred by Algorithm 3.2 from an ODE system in additive normal form* $\dot{x_i} = \sum_{s=1}^{t} c_{i,s} * f_s$ *for* $1 \leq i \leq v$ *is the set of non-decomposable reactions*

$\{f_s$ *for* $r_s / m_s \rightarrow p_s\}_{1 \leq s \leq t}$

$$\text{where}\quad r_s = \sum_{\{i \mid c_{i,s} < 0\}} (-c_{i,s}) * x_i + \sum_{\{i \mid c_{i,s} \geq 0,\ \frac{\partial f_s}{\partial x_i} > 0\}} x_i,$$

$$p_s = \sum_{\{i \mid c_{i,s} > 0\}} c_{i,s} * x_i + \sum_{\{i \mid c_{i,s} \geq 0,\ \frac{\partial f_s}{\partial x_i} > 0\}} x_i$$

and m_s *is the set of variables* x *such that* $\frac{\partial f_s}{\partial x} < 0$*.*

Proposition 3.5 (Soundness)**.** *The ODE system associated to the reaction model inferred from an ODE system* O *is equivalent to* O*.*

Proof. Let us suppose without loss of generality that $O = \{\dot{x_i} = \sum_{s=1}^{t} c_{i,s} * f_s \mid 1 \leq i \leq m\}$ is in additive normal form. The inferred reaction model is the set $\{f_s \text{ for } r_s/m_s \to p_s\}_{1 \leq s \leq t}$
where $r_s = \sum_{\{i \mid c_{i,s}<0\}}(-c_{i,s}) * x_i + \sum_{\{i \mid c_{i,s} \geq 0,\ \frac{\partial f_s}{\partial x_i} > 0\}} x_i$,
$p_s = \sum_{\{i \mid c_{i,s}>0\}} c_{i,s} * x_i + \sum_{\{i \mid c_{i,s} \geq 0,\ \frac{\partial f_s}{\partial x_i} > 0\}} x_i$,
and m_s is the set of variables y such that $\frac{\partial f_s}{\partial y} < 0$.

The ODE system associated to these reactions is thus
$\{\dot{x_i} = \sum_{s=1}^{t}(p_s(x_i) - r_s(x_i)) * f_s\}_{1 \leq i \leq m}$
$= \{\dot{x_i} = \sum_{s=1}^{t} c_{i,s} * f_s\}_{1 \leq i \leq m} \ = \ O.$ □

Algorithm 3.2 always computes a reaction model with an equivalent associated ODE system but this reaction model may not be well-formed. In particular, step 3b adds a variable x to the reactants of the reactions even if x does not appear in the kinetic expression f of the reaction. Therefore the algorithm may infer reactions with reactants that do not occur in the kinetic expression. On the other hand, all variables appearing in the kinetics will now appear in the reaction as either catalysts (step 3c), inhibitors or both (step 3d):

Proposition 3.6. *The reaction models inferred by Algorithm 3.2 contain no reaction with a molecular species x appearing in the kinetic expression f with $\frac{\partial f}{\partial x \neq 0}$ and not appearing as a reactant or modifier.*

As for the completeness of the inference algorithm, Example 2.2 shows that we need to consider ODEs associated to both well-formed and non-decomposable reactions to ensure that the inferred reaction model is well-formed.

Proposition 3.7 (completeness)**.** *The reaction model inferred from the ODEs associated to a non-decomposable well-formed reaction model is well-formed and non-decomposable.*

Proof. Let $R = \{\ e_i \texttt{ for } \ r_i \ / \ m_i \texttt{ => } \ p_i \ \}_{i=1,\ldots,n}$ be a well-formed reaction model with non-decomposable kinetics. The associated ODE is the system $O = \{\dot{x_j} = \sum_{i=1}^{n}(p_i(x_j) - r_i(x_j)) * e_i \mid 1 \leq i \leq m\}$ which is in additive normal form by hypothesis (after evaluation of the integer $p_i(x_j) - r_i(x_j)$). By Prop. 3.4, the inferred reaction model is $\{e_i \text{ for } r'_i/m'_i \to p'_i\}_{1 \leq i \leq n}$ where e_i is non-decomposable by hypothesis,

$$\begin{aligned} r'_i = \ & \sum_{\{j \mid p_i(x_j)<r_i(x_j)\}}(r_i(x_j) - p_i(x_j)) * x_j \\ & + \sum_{\{j \mid p_i(x_j) \geq r_i(x_j),\ \frac{\partial e_i}{\partial x_j} > 0\}} x_j \\ p'_i = \ & \sum_{\{j \mid p_i(x_j)>r_i(x_j)\}}(p_i(x_j) - r_i(x_j)) * x_j \\ & + \sum_{\{j \mid p_i(x_j) \geq r_i(x_j),\ \frac{\partial e_i}{\partial x_j} > 0\}} x_j, \end{aligned}$$

and $m'_i = m_i$. Now for any variable x_j, we have $x_j \in r'_i$ if and only if $x_k \in r_i$ since either $p_i(x_j) < r_i(x_j)$ or $\frac{\partial e_i}{\partial x_j} > 0$. Similarly $x_j \in p'_i$ if and only if $x_j \in p_i$ since either $p_i(x_j) > r_i(x_j)$ or $p_i(x_j) = r_i(x_j)$ and $\frac{\partial e_i}{\partial x_j} > 0$, These equalities between the sets (not multisets) of reactants, products and inhibitors suffice to show the well-formedness for the inferred reactions. □

3.2 Inference Algorithm for Hidden Molecules

ODE models often contain algebraic invariants, among which linear invariants, e.g. mass conservation invariants or Petri-net place invariants, are an important particular case. A linear invariant can be used to simplify a model by eliminating one variable and replacing it with a linear expression. This may have several advantages, but hard coding this elimination in the kinetic expressions of a reaction model may affect the structure of the reactions and may invalidate some structural analyses.

Example 3.8. *The model of Example 2.3 has one invariant: $pMPF + MPF$ is a constant c (the initial value of $pMPF$ and MPF) since $\dot{pMPF} + \dot{MPF} = 0$. One variable, e.g. $pMPF$, can thus be eliminated and replaced by $c - MPF$. This yields the to the ODE systen*

$\dot{MPF} = k1 * (c - [MPF]) * [Cdc25] - k2 * [MPF] * [Wee1]$
$\dot{Wee1} = k3/(k4 + [Clock])$ $\quad \dot{Cdc25} = 0$ $\quad \dot{Clock} = 0$

On this form, Algorithm 3.2 infers the unintended reactions:

```
c*k1*[Cdc25]      for Cdc25 => Cdc25 + MPF
k1*[Cdc25]*[MPF]  for MPF + Cdc25 => Cdc25
k2*[MPF]*[Wee1]   for MPF + Wee1 => Wee1
k3/(k4+[Clock])   for _ / Clock => Wee1
```

In this section we describe an algorithm for reversing the simplifications by linear invariants and restoring *hidden* molecular species. This reversal transformation is applied as a preprocessor before inferring the reactions from the ODEs with Algorithm 3.2.

Let us first note that one can easily add to an ODE system a new variable y equal to a linear combination e of its variables $e = \sum \lambda_i x_i$ without changing its solutions when projected on the x_i. The new equation for y is $\dot{y} = \sum \lambda_i \dot{x}_i$, the $\dot{x}_i$ being given by the rest of the system. If one also imposes as initial condition for y the value of e at the initial state, it becomes even possible to replace in the original system some occurrences of e by y while keeping an equivalent system.

Example 3.9. *Starting with the system from Example 3.8, let us introduce $e = c - [MPF]$ as linear combination. We can now add a new variable y such that $y_0 = c - [MPF]_0$ and since $c = [MPF]_0 + [pMPF]_0$ we get $y_0 = [pMPF]_0$.*

The ODE system is:
$\dot{MPF} = k1 * [y] * [Cdc25] - k2 * [MPF] * [Wee1]$
$\dot{Wee1} = k3/(k4 + [Clock])$ $\quad \dot{Cdc25} = 0$ $\quad \dot{Clock} = 0$
$\dot{y} = -\dot{MPF} = k2 * [MPF] * [Wee1] - k1 * [y] * [Cdc25]$

This procedure can thus be used to reverse the simplification process of a linear invariant elimination. However, the expression e needs be chosen with care, otherwise useless variables may be introduced, for instance if $e = x_i$.

In our implementation, the expressions are first normalized so that expressions like $-1.0 * A + B$ are rewritten as $B - A$. Second, expressions of the form $(K - X) - Y$ or $K - X$ are searched in this order, with K a parameter or constant, and X, Y molecule concentrations. The order does not let $K - X$ be selected when it appears in $(K - X) - Y$. Notice that the first normalization phase allows us to catch expressions such as $-X + K$. Third, for each selected expression, a hidden molecule is inferred. Finally, the hidden molecule is substituted to the expression by replacing in the ODEs: $K - X$ by G, $(J + K) - X$ by $J + G$, $(K + J) - X$ by $J + G$ and $K + (J - X)$ by $J + G$, where J can be any expression.

4 Evaluation Results on biomodels.net

The 409 models from the curated branch of the latest version (release 21) of the `biomodels.net` repository [le Novère et al., 2006] were used as benchmark. Out of those 409 models only 345 define *reactions*, the other ones only describing systems through events and rules. Though the fact that a reaction provides its *kineticLaw* is not compulsory in the SBML specification, 340 of the 345 structured models do provide proper kinetic laws and are thus amenable to automatic curation via export and import of the corresponding system of ODEs.

4.1 Global analysis

The following table sums up the result of the procedure, as detected by BIOCHAM warnings. These warnings correspond to syntactical conditions that indicate that a reaction is not *well-formed*:

- "K not R" denotes the number of models in which the concentration of some compound appears in a kinetic law but this compound is neither a reactant nor a modifier of the reaction;
- "R not K" denotes the number of models in which some compound is marked as reactant or modifier of a reaction but does not appear in the its kinetic law.
- "Negative" denotes the number of models where a minus sign appears in the kinetic expression at some place that is not inside an exponent expression.

Indeed, in a well-formed reaction, if a species is a reactant or an inhibitor, then $\partial e/\partial x \neq 0$, therefore x should appear in e. Similarly, if a species x is neither a reactant nor an inhibitor, then $\partial e/\partial x = 0$ so x should not appear in the kinetic expression e. Moreover, for having $e \geq 0$ and well defined in the whole positive quadrant, one can argue that e should not contain any subtraction.

Note that once again, there is some consistency with the SBML specification of [Hucka et al., 2008], which states: *"Any species appearing in the mathematical formula of the kineticLaw of a Reaction instance must be declared in at least one of that Reaction's lists of reactants, products, and/or modifiers. Put another way, it is an error for a reaction's kinetic law formula to refer to species that have not been declared for that reaction."* In other words there should never be any "K not R" warning in any SBML model. However, as the table below shows, this is the case of many models from the curated branch of the biomodels repository.

Over the 340 reaction models of the original curated part of biomodels.net, our algorithm detects 57 models with *hidden molecules*, 165 models with K not R warning, 120 models with R not K warning and 148 models with negative kinetics warning. Our algorithm is able to automatically curate this database of models by reducing the number of non well-formed models with a warning by more than the half, from 66% to 28%:

Biomodels.net	**K not R**	**R not K**	**Neg.**	**Any warning**
Orig. Curated	165	120	148	225 (66.17%)
Auto. Curated	0	82	39	96 (28.23%)

As predicted by Proposition 3.6, the algorithm 3.2 completely removes the "K not R" warnings. For the two other warnings, since the algorithm focuses on *non-decomposable* kinetics, it results in curated models quite close to the original ones, but does not tackle thoroughly the case of reactions with rates independent of some reactant, as in Example 2.2. For these reasons, 96 over 340 models remain with a non well-formedness warning after automatic curation.

4.2 Models studied in [Kaleta et al., 2009]

[Kaleta et al., 2009] also scan the whole biomodels repository and report finding 5 inconsistencies: models 44, 93, 94, 143 and 151.

Their diagnostics is as follows, some reaction fluxes become negative during the simulations of those models because of missing reversibility indications in models 93, 94 and 143. In the two first cases they report that adding the reverse reactions makes the models consistent, whereas

for 143 it is also necessary to change some kinetic law. For model 151 they report a missing step, but since the opposite reaction is part of the model, once again this amounts to adding a reverse reaction to an existing one. Finally, for model 44 they describe that the issue is that some kinetic expression does not depend on one of the reactants of the reaction, making it possible for that reactant's concentration to become negative.

For models 93, 94 and 151, that indeed are flagged by the "Negative" warning, our algorithm correctly adds the missing reverse reactions, directly from the kinetic expressions. The models automatically curated this way do not raise any warning at the end.

For model 44, the automatic curation allows us to get rid of a "K not R" warning by transforming the reaction `v3`
`A::cyt + Y::ves => A::cyt + Z::cyt` with the kinetic law
$cytosol * Vm3 * [A]^4 * [Y]^2 * [Z]^4/((Ka^4 + [A]^4) * ((Ky^2 + [Y]^2) * (Kz^4 + [Z]^4)))$ into
`Z::cyt + A::cyt + Y::ves => 2*Z::cyt + A::cyt`

However, as expected, the "R not K" warning identified by Kaleta *et al.* remains, the obtained model is still not well-formed. The same happens with model 143 where indeed a "R not K" warning remains after automatic curation, in accordance with the earlier results.

Note that this also shows that three years later, the same flaws are still present in biomodels, which illustrates the need for automatic curation methods.

4.3 Models of the Cell Cycle

As shown in [Gay et al., 2010], the list of models in a repository like `biomodels.net` can be organized in a hierarchy of models related by reduction/refinement relationships between them. These relationships can be computed by an algorithm for detecting subgraph epimorphisms between the reaction graphs of the models. On the models of the cell cycle however, this method did not produce good results because these originally ODE models have been transcribed with different conventions in SBML and the structure of the reaction graphs may not correctly reflect the molecular interactions that are given in the kinetics.

For instance, in [Gay et al., 2010] it was noted that the model referenced as `BIOMD0000000008.xml` of [Gardner et al., 1998], adding a control mechanism to the cell-cycle model of [Goldbeter, 1991], was not easily amenable to structural analysis. In particular, its reaction graph was not connected.

Here are some of the reactions of this original model in the curated branch of biomodels:

```
r4: (1+ -1*[M])*V1*(r4K1+(-1*[M]+1))^-1 for _ => M.
r5: [M]*r5V2*(r5K2+[M])^-1       for M => _.
r6: V3*(1+ -1*[X])*(r6K3+(-1*[X]+1))^-1 for _ => X.
r7: r7V4*[X]*(r7K4+[X])^-1       for X => _.
```

One of the issues is *hidden* in the definition of `V1` and `V3`:

```
macro(V1,[C]*V1p*([C]+K6^-1)).
macro(V3,[M]*V3p).
```

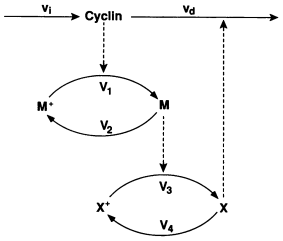
This shows that, as pointed out by a "K not R" warning, `C` is indeed involved in the kinetics of `r4` but it is not marked as modifier.

One can also note that, though encoded in complicated MathML expressions, `1 - [M]` (resp. `1 - [X]`) appears in the synthesis of `M` (resp. `X`) as a *ghost* form of the inactive form of `M` (resp. `X`). Indeed, [Goldbeter, 1991] states that *"(1 - M) thus represents the fraction of inactive (i.e., phosphorylated) cdc2 kinase, while (1 - X) represents the fraction of inactive (i.e., dephosphorylated) cyclin protease"*. After automatic curation the model becomes:

```
r4: V1*[M_i]*(r4K1+[M_i])^-1 for M_i =[C]=> M.
r5: r5V2*[M]*(r5K2+[M])^-1   for M => M_i.
r6: V3*[X_i]*(r6K3+[X_i])^-1 for X_i =[M]=> X.
```

```
r7: r7V4*[X]*(r7K4+[X])^-1   for X => X_i.
```

The fact that the two inactive forms are now explicit and that the action of `C` on `M` and of `M` on `X` are properly indicated provides a well-formed reaction model consistent with the usual graphical representation and suitable for further structural analysis:



5 Conclusion

We have described an algorithm for automatically curating SBML models through their semantics with ordinary differential equations. The method is based on an algorithm which, given an ODE system in input, allows us to infer a reaction model with the same ODE semantics.

We have analyzed the capability of this method to automatically curate the transcription in SBML of the cell cycle models in `biomodels.net`. In particular, we have shown that the inference of well-formed reactions from the ODEs, combined with the inference of hidden molecules through the recognition of their elimination using linear invariants, provide a consistent representation of these ODE models in SBML with which systematic structural analysis methods, such as model comparison by subgraph epimorphism [Gay et al., 2010], can be applied. On the whole curated part of the biomodels repository, we have shown that our automatic curation method significantly improves the writing of the models in SBML by reducing the number of non well-formed models from 66% to 31%.

Interestingly, our algorithm is based on a general mathematical condition for expressing the compatibility between a kinetic expression and the structure of a reaction in terms of its reactants, products and inhibitors. We have shown some general properties enjoyed by the ODE systems associated to such well-formed reaction models. These results militate for distinguishing between catalysts and inhibitors in the modifiers of a reaction.

We believe that our well-formedness conditions which generalize previous restrictions to mass action law, or rational expression kinetics, provide a solid ground for developing a theory of chemical reactions, as needed for the development of SBML and for the building, and efficient use, of model repositories in systems biology.

Acknowledgement: This work has been supported by the French OSEO Biointelligence and ANR Biotempo projects, and by the European project EraNet Sysbio C5Sys.

References

[Angeli et al., 2007] Angeli, D., Leenheer, P. D., and Sontag, E. D. (2007). A petri net approach to persistence analysis in chemical reaction networks. In *Biology and Control Theory: Current Challenges*, volume 357 of *LNCI*, pages 181–216. Springer-Verlag.

[Calzone et al., 2006] Calzone, L., Fages, F., and Soliman, S. (2006). BIOCHAM: An environment for modeling biological systems and formalizing experimental knowledge. *Bioinformatics*, 22(14):1805–1807.

[Chaouiya et al., 2008] Chaouiya, C., Remy, E., and Thieffry, D. (2008). Petri net modelling of biological regulatory networks. *Journal of Discrete Algorithms*, 6(2):165–177.

[Dittrich and di Fenizio, 2007] Dittrich, P. and di Fenizio, P. (2007). Chemical organisation theory. *Bulletin of Mathematical Biology*, 69(4):1199–1231.

[Eker et al., 2002] Eker, S., Knapp, M., Laderoute, K., Lincoln, P., Meseguer, J., and Sönmez, M. K. (2002). Pathway logic: Symbolic analysis of biological signaling. In *Proceedings of the seventh Pacific Symposium on Biocomputing*, pages 400–412.

[Ermentrout, 2002] Ermentrout, B. (2002). *Simulating, Analyzing, and Animating Dynamical Systems: A Guide to XPPAUT for Researchers and Students*. SIAM, Philadelphia.

[Fages and Soliman, 2008a] Fages, F. and Soliman, S. (2008a). Formal cell biology in BIOCHAM. In Bernardo, M., Degano, P., and Zavattaro, G., editors, *8th Int. School on Formal Methods for the Design of Computer, Communication and Software Systems: Computational Systems Biology SFM'08*, volume 5016 of *Lecture Notes in Computer Science*, pages 54–80, Bertinoro, Italy. Springer-Verlag.

[Fages and Soliman, 2008b] Fages, F. and Soliman, S. (2008b). From reaction models to influence graphs and back: a theorem. In *Proceedings of Formal Methods in Systems Biology FMSB'08*, number 5054 in Lecture Notes in Computer Science. Springer-Verlag.

[Fages et al., 2004] Fages, F., Soliman, S., and Chabrier-Rivier, N. (2004). Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM. *Journal of Biological Physics and Chemistry*, 4(2):64–73.

[Feinberg, 1977] Feinberg, M. (1977). Mathematical aspects of mass action kinetics. In Lapidus, L. and Amundson, N. R., editors, *Chemical Reactor Theory: A Review*, chapter 1, pages 1–78. Prentice-Hall.

[Gardner et al., 1998] Gardner, T. S., Dolnik, M., and Collins, J. J. (1998). A theory for controlling cell cycle dynamics using a reversibly binding inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*, 95(24):14190–14195.

[Gay et al., 2010] Gay, S., Soliman, S., and Fages, F. (2010). A graphical method for reducing and relating models in systems biology. *Bioinformatics*, 26(18):i575–i581. special issue ECCB'10.

[Gillespie, 1977] Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361.

[Goldbeter, 1991] Goldbeter, A. (1991). A minimal cascade model for the mitotic oscillator involving cyclin and cdc2 kinase. *PNAS*, 88(20):9107–9111.

[Gouzé, 1998] Gouzé, J.-L. (1998). Positive and negative circuits in dynamical systems. *Journal of Biological Systems*, 6:11–15.

[Grafahrend-Belau et al., 2008] Grafahrend-Belau, E., Schreiber, F., Heiner, M., Sackmann, A., Junker, B. H., Grunwald, S., Speer, A., Winder, K., and Koch, I. (2008). Modularization of biochemical networks based on a classification of petri net by T-invariants. *BMC Bioinformatics*, 9(90).

[Hárs and Tóth, 1979] Hárs, V. and Tóth, J. (1979). On the inverse problem of reaction kinetics. In Farkas, M., editor, *Colloquia Mathematica Societatis János Bolyai*, volume 30 of *Qualitative Theory of Differential Equations*, pages 363–379.

[Hucka et al., 2003] Hucka, M. et al. (2003). The systems biology markup language (SBML): A medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

[Hucka et al., 2008] Hucka, M., Hoops, S., Keating, S. M., Nicolas, L. N., Sahle, S., and Wilkinson, D. (2008). Systems biology markup language (SBML) level 2: Structures and facilities for model definitions. *Nature Precedings*.

[Kaleta et al., 2009] Kaleta, C., Richter, S., and Dittrich, P. (2009). Using chemical organization theory for model checking. *Bioinformatics*, 25(15):1915–1922.

[Kaufman et al., 2007] Kaufman, M., Soulé, C., and Thomas, R. (2007). A new necessary condition on interaction graphs for multistationarity. *Journal of Theoretical Biology*, 248:675–685.

[Koh et al., 2006] Koh, G., Teong, H., Clement, M.-V., Hsu, D., and Thiagarajan, P. (2006). A decompositional approach to parameter estimation in pathway modeling: a case study of the akt and mapk pathways and their crosstalk. *Bioinformatics*, 22(14):e271–e280.

[Kohn, 1999] Kohn, K. W. (1999). Molecular interaction map of the mammalian cell cycle control and DNA repair systems. *Molecular Biology of the Cell*, 10(8):2703–2734.

[le Novère et al., 2006] le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J. L., and Hucka, M. (2006). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acid Research*, 1(34):D689–D691.

[Nabli and Soliman, 2010] Nabli, F. and Soliman, S. (2010). Steady-state solution of biochemical systems, beyond S-Systems via T-invariants. In Quaglia, P., editor, *CMSB'10: Proceedings of the 8th International Conference on Computational Methods in Systems Biology*, pages 14–22. CoSBi, ACM.

[Reddy et al., 1993] Reddy, V. N., Mavrovouniotis, M. L., and Liebman, M. N. (1993). Petri net representations in metabolic pathways. In Hunter, L., Searls, D. B., and Shavlik, J. W., editors, *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pages 328–336. AAAI Press.

[Rohr et al., 2010] Rohr, C., Marwan, W., and Heiner, M. (2010). Snoopy - a unifying petri net framework to investigate biomolecular networks. *Bioinformatics*, 26(7):974–975.

[Schuster et al., 2002] Schuster, S., Fell, D. A., and Dandekar, T. (2002). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*, 18:326–332.

[Shinar and Feinberg, 2010] Shinar, G. and Feinberg, M. (2010). Structural sources of robustness in biochemical reaction networks. *Science*, 327(5971):1389–1391.

[Snoussi, 1998] Snoussi, E. H. (1998). Necessary conditions for multistationarity and stable periodicity. *Journal of Biological Systems*, 6:3–9.

[Soliman, 2008] Soliman, S. (2008). Finding minimal P/T-invariants as a CSP. In *Proceedings of the fourth Workshop on Constraint Based Methods for Bioinformatics WCB'08, associated to CPAIOR'08*.

[Soliman and Heiner, 2010] Soliman, S. and Heiner, M. (2010). A unique transformation from ordinary differential equations to reaction networks. *PLoS One*, 5(12):e14284.

[Soulé, 2003] Soulé, C. (2003). Graphic requirements for multistationarity. *ComplexUs*, 1:123–133.

[Soulé, 2006] Soulé, C. (2006). Mathematical approaches to differentiation and gene regulation. *C. R. Biologies*, 329:13–20.

[Szederkényi et al., 2011] Szederkényi, G., Banga, J. R., and Alonso, A. A. (2011). Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC systems biology*, 5(1):177+.

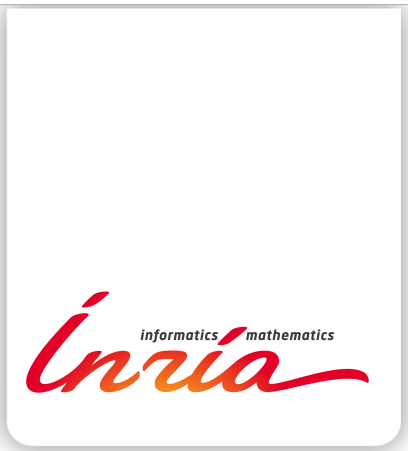
[Thomas, 1981] Thomas, R. (1981). On the relation between the logical structure of systems and their ability to generate multiple steady states or sustained oscillations. *Springer Ser. Synergetics*, 9:180–193.

[Thomas et al., 1976] Thomas, R., Gathoye, A.-M., and Lambert, L. (1976). A complex control circuit : regulation of immunity in temperate bacteriophages. *European Journal of Biochemistry*, 71(1):211–227.

[Tyson, 1991] Tyson, J. J. (1991). Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences*, 88(16):7328–7332.

[Varma and Palsson, 1994] Varma, A. and Palsson, B. (1994). Metabolic flux balancing: basic concepts, scientific and practical use. *Nature Biotechnology*, 12(10):994–998.

[Zevedei-Oancea and Schuster, 2003] Zevedei-Oancea, I. and Schuster, S. (2003). Topological analysis of metabolic networks based on petri net theory. *In Silico Biology*, 3(29).



RESEARCH CENTRE
PARIS – ROCQUENCOURT

Domaine de Voluceau, - Rocquencourt
B.P. 105 - 78153 Le Chesnay Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399