Simulation techniques for the calculus of wrapped compartments

Simulation Techniques for the Calculus of Wrapped Compartments[☆]

Mario Coppo[a], Ferruccio Damiani[a], Maurizio Drocco[a], Elena Grassi[b], Eva Sciacca[a], Salvatore Spinella[a], Angelo Troina[a]

[a] *Dipartimento di Informatica, Università di Torino*
[b] *Molecular Biotechnology Center, Dipartimento di Genetica, Biologia e Biochimica, Università di Torino*

Abstract

The modelling and analysis of biological systems has deep roots in Mathematics, specifically in the field of Ordinary Differential Equations (ODEs). Alternative approaches based on formal calculi, often derived from process algebras or term rewriting systems, provide a quite complementary way to analyse the behaviour of biological systems. These calculi allow to cope in a natural way with notions like compartments and membranes, which are not easy (sometimes impossible) to handle with purely numerical approaches, and are often based on stochastic simulation methods. Recently, it has also become evident that stochastic effects in regulatory networks play a crucial role in the analysis of such systems. Actually, in many situations it is necessary to use stochastic models. For example when the system to be described is based on the interaction of few molecules, when we are at the presence of a chemical instability, or when we want to simulate the functioning of a pool of entities whose compartmentalised structure evolves dynamically. In contrast, stable metabolic networks, involving a large number of reagents, for which the computational cost of a stochastic simulation becomes an insurmountable obstacle, are efficiently modelled with ODEs. In this paper we define a hybrid simulation method, combining the stochastic approach with ODEs, for systems described in the Calculus of Wrapped Compartments (CWC), a

[☆]This research is founded by the BioBITs Project (*Converging Technologies* 2007, area: Biotechnology-ICT), Regione Piemonte.
During the final revision phase of this paper, Angelo Troina was visiting the Instituto de Ecología at Universidad Tecnica Particular de Loja (Ecuador), in the context of the Prometeo program founded by SENESCYT.

calculus on which we can express the compartmentalisation of a biological system whose evolution is defined by a set of rewrite rules.

Keywords: Term Rewriting Systems, Stochastic/Deterministic Simulation Methods, Hybrid Simulation, Biochemical Systems

1. Introduction

The most common approaches used by biologists to describe biological systems have been mainly based on the use of deterministic mathematical means like, e.g., Ordinary Differential Equations (ODEs for short). ODEs make it possible to abstractly reason on the behaviour of biological systems and to perform a quantitative *in silico* investigation. However, this kind of modelling becomes more and more difficult, both in the specification phase and in the analysis processes, when the complexity of the biological systems taken into consideration increases. More recently, the observation that biological systems (for example in the case of chemical instability) are inherently stochastic [1], has led a growing interest in the stochastic modelling of chemical kinetics.

Besides, the concurrently interacting structure of biological systems has inspired the possibility to describe them by means of formalisms developed in Computer Science for the description of computational entities [2]. Different formalisms have either been applied to (or have been inspired from) biological systems. Automata-based models [3, 4] have the advantage of allowing the direct use of many verification tools such as model checkers. Rewrite systems [5, 6, 7] usually allow describing biological systems with a notation that can be easily understood by biologists. Process calculi, including those commonly used to describe biological systems [2, 8, 9], have the advantage of being compositional, but their way of describing biological entities is often less intuitive. Quantitative simulations of biological models represented with these kind of frameworks (see, e.g. [8, 10, 11, 12, 13, 14]) are usually developed via a stochastic method derived from Gillespie's algorithm [15]. These formalisms, in fact, are mainly used to represent interactions between complex biological structures like cells, bacteria or genes whose results would be hard to validate in the ODE approach.

The ODE description of biological systems determines *continuous* and deterministic models in which variables describe the concentrations of the species involved in the system as functions of the time. These models are

based on average reaction rates, measured from real experiments which relate to the change of concentrations over time, taking into account the known properties of the involved chemicals, but possibly abstracting away some unknown mechanisms.

In contrast to the deterministic model, *discrete* and stochastic simulations involve random variables. Since the basic steps of a molecular reaction are described in terms of their probability of occurrence, the behaviour of a reaction is not determined a priori but characterised statistically. Thus, biological reactions fall in the category of stochastic systems and stochastic models for their kinetics are widely accepted as the best way to represent and simulate genetic and biochemical networks. In particular, when the system to be described is based on the interaction of few molecules, or we want to simulate the functioning of a little pool of cells, the system may expose several different behaviour, observed with different probabilities. As a consequence, the stochastic approach is always valid when the deterministic one is (i.e. when the system is stable and exposes only one possible behaviour), and it may be valid when the ordinary deterministic is not (i.e. in a nonlinear system in the neighbourhood of a chemical instability).

Actually, in the last years it has become evident that stochastic effects in regulatory networks play a crucial role in the analysis of multi-stable systems. In contrast, metabolic networks involving large numbers of molecules are most often modelled deterministically. Summing up, it may happen that a purely deterministic model does not accurately capture the dynamics of the considered system, and a stochastic description is needed.

However, the computational cost of a discrete stochastic simulation often becomes an insurmountable obstacle. Instead, the ODEs method is extremely more computationally efficient. Thus, when the deterministic approach is applicable and provides a good approximation of the system behaviour, it might be profitable to take advantage of its efficiency, and move to the stochastic approach when this is not true any more.

In a hybrid model, some reactions are modelled in a discrete way (i.e. computed, probabilistically, according to an exact stochastic method) and others in a continuous way (i.e. computed, in a deterministic way, by a set of ODEs). These models for the simulation of biological systems have been presented in the last few years for purely mathematical models [16, 17, 18], i.e. models in which all reactions take place in single, "flat" ambient (without compartmentalisation). In this paper we adapt the hybrid simulation technique within the programming language approach to describe and analyse

the dynamics of biological systems.

Approximate methods of stochastic simulations have been developed, like the ones employing stochastic differential equations (SDEs), in particular the Langevin type [19], or the tau-leap technique introduced by Gillespie [20, 21]. On the one hand, in the analysis with stochastic differential equations, the change in molecule concentrations is modelled in a continuous space. Such an approximation is valid only when a sufficiently large number of molecules is involved, since effects due to discreteness can bring the system in a state that is not captured with a continuous analysis. On the other hand, the tau-leaping method is no longer efficient if a system contains even a single reaction with very small numbers of substrate molecules because the length of the correct time step is of the order of waiting times occurring in exact simulation algorithms.

The mentioned approximate methods, however, for systems involving fast reactions and a large number of molecules converge, at the thermodynamic limit, with the solution obtained with the exact discrete methods (see e.g. [19]).

In this paper we will introduce an hybrid simulation algorithm for the *Calculus of Wrapped Compartments* (CWC for short), a variant of the Calculus of Looping Sequences (CLS for short) [7, 12] and develop a hybrid simulation algorithm for it. Starting from an alphabet of atomic elements and from an alphabet of labels (representing compartment types), CWC terms are defined as multisets of elements and labelled compartments. Elements can be localized by compartmentalisation and the structure of a compartment can be specified by detailing the elements of interest on its membrane (as atomic elements) and its type (as a label). This allows, for instance, to represent a cell as a compartment and its nucleus with a separate, nested, compartment. The evolution of the system is driven by a set of rewrite rules modelling the reactions of interest that can be local to a single compartment, or involve different compartments (for instance in the movement of a component from a compartment to another). Compartments can be dynamically created or destroyed.

We will apply our approach to a variant of Lotka-Volterra dynamics and an HIV-1 transactivation mechanism.

The CWC simulator, currently under development at the Computer Science department of Turin University, has been enriched with a prototype implementation of the hybrid simulation algorithm [22]. It has been proven, by means of several case studies, that hybrid simulations produce results con-

sistent with those obtained by the exact stochastic simulation method, but with a considerable gain in computing time. This paper surveys and extends the preliminary results presented in [14, 23].

Summary. Section 2 introduces the CWC formalism. Section 3 recalls the stochastic and the deterministic simulation methods. Section 4 introduces the hybrid simulation technique and Section 5 applies it to the analysis of the HIV-1 transactivation mechanism. Related work is discussed in Section 6. We draw our conclusions in Section 7. Appendix A briefly discusses the criterion for the dynamical partition of deterministic and stochastic rules.

2. The Calculus of Wrapped Compartments

Like most modelling languages based on term rewriting (notably CLS), a CWC model consists of a term, representing the (biological) system and a set of rewrite rules which model the transformations determining the system's evolution. Terms are defined from a set of atomic elements via an operator of compartment construction. Compartments are enriched with a nominal type, represented as a label, which identifies the set of rewrite rules that may be applied to them.

2.1. Terms and Structural Congruence

Terms of the CWC calculus are intended to represent a biological system. A *term* is a multiset of *simple terms*. Simple terms, ranged over by t, u, v, ... are built by means of the *compartment* constructor, $(- \rfloor -)^{-}$, from a set $\mathcal{A}$ of *atomic elements* (*atoms* for short), ranged over by a, b, c, ... and from a set $\mathcal{L}$ of *compartment types* (represented as *labels* attached to compartments), ranged over by $\ell, \ell', \ell_1, \ldots$ and containing a distinguished element $\top$ which characterises the top level compartment. The syntax of simple terms is given in Figure 1. We write $\overline{t}$ to denote a (possibly empty) multiset of simple terms $t_1 \cdots t_n$. Similarly, with $\overline{a}$ we denote a (possibly empty) multiset of atoms. The set of simple terms will be denoted by $\mathcal{T}$.

Then, a simple term is either an atom or a compartment $(\overline{a} \rfloor \overline{t})^{\ell}$ consisting of a *wrap* (represented by the multiset of atoms $\overline{a}$), a *content* (represented by the term $\overline{t}$) and a *type* (represented by the label ℓ). Note that we do not allow nested structures within wraps but only in compartment contents. We write $\bullet$ to represent the empty multiset and denote the union of two multisets $\overline{u}$ and $\overline{v}$ as $\overline{u}\ \overline{v}$. The notion of inclusion between multisets, denoted as usual

Simple terms syntax

$t \quad ::= \quad a \;\big|\; (\overline{a} \rfloor \overline{t})^{\ell}$

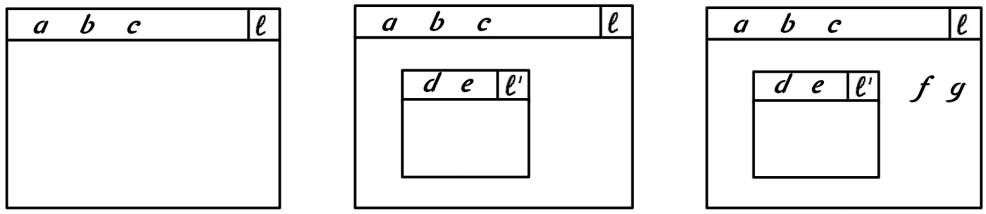
Structural congruence

$\overline{t}\ u\ w\ \overline{v} \equiv \overline{t}\ w\ u\ \overline{v}$

if $u \equiv w$ then $\overline{t}\ u\ \overline{v} \equiv \overline{t}\ w\ \overline{v}$

if $\overline{a} \equiv \overline{b}$ and $\overline{t} \equiv \overline{u}$ then $(\overline{a} \rfloor \overline{t})^{\ell} \equiv (\overline{b} \rfloor \overline{u})^{\ell}$

Figure 1: CWC term syntax and structural congruence rules



(a) (b) (c)

Figure 2: **(a)** represents $(a\ b\ c \rfloor \bullet)^{\ell}$; **(b)** represents $(a\ b\ c \rfloor (d\ e \rfloor \bullet)^{\ell'})^{\ell}$; **(c)** represents $(a\ b\ c \rfloor (d\ e \rfloor \bullet)^{\ell'}\ f\ g)^{\ell}$

by $\subseteq$, is the natural extension of the analogous notion between sets. The set of terms (multisets of simple terms) and the set of multisets of atoms will be denoted by $\overline{\mathcal{T}}$ and $\overline{\mathcal{A}}$, respectively. Note that $\overline{\mathcal{A}} \subseteq \overline{\mathcal{T}}$.

Since a term $\overline{t} = t_1 \cdots t_n$ is intended to represent a multiset we introduce a relation of structural congruence between terms of CWC defined as the least equivalence relation on terms satisfying the rules given in Figure 1. From now on we will always consider terms modulo structural congruence. To denote multisets of atomic elements we will sometime use the compact notation na where a is an atomic element and n its multiplicity so for instance $3a\,2b$ is a notation for the multiset $a\,a\,a\,b\,b$.

An example of term is $\overline{t} = 2a\ \ 3b\ \ (c\ \ d \rfloor e\ \ f)^{\ell}$ representing a multiset consisting of two atoms a and three b (for instance five molecules) and an ℓ-type compartment $(c\ \ d \rfloor e\ \ f)^{\ell}$ which, in turn, consists of a wrap (a membrane) with two atoms c and d (for instance, two proteins) on its surface, and containing the atoms e (for instance, a molecule) and f (for instance a DNA strand whose functionality can be modelled as an atomic element). See Figure 2 for some graphical representations.

Notation 2.1 (Top-level compartment). *For sake of uniformity we assume that the term representing the whole system is always a single compartment labelled $\top$ with an empty wrap, i.e., all systems are represented by a term of the shape $(\bullet \rfloor \overline{t})^{\top}$, which we will also write as $\overline{t}$ for simplicity.*

2.2. Contexts

The notion of reduction in CWC is formalised via the notion of reduction context. To define them, the syntax of terms is enriched with a new element $\Box$ representing a hole which can be filled only by a *single* compartment. *Reduction contexts* (ranged over by C) are defined by:

$$C \quad ::= \quad \Box \quad \Big| \quad (\overline{a} \rfloor C \ \overline{t})^{\ell}$$

where $\overline{a} \in \overline{\mathcal{A}}$, $\overline{t} \in \overline{\mathcal{T}}$ and $\ell \in \mathcal{L}$. Note that, by definition, every context contains a single hole $\Box$. The set of contexts is denoted by $\mathcal{C}$.

Given a compartment $t = (\overline{a} \rfloor \overline{u})^{\ell}$ and a context C, the compartment obtained by filling the hole in C with t is denoted by $C[t]$. For instance, if $t = (\overline{a} \rfloor \overline{u})^{\ell}$ and $C = (\overline{b} \rfloor \Box \ \overline{c})^{\ell'}$, then $C[t] = (\overline{b} \rfloor (\overline{a} \rfloor \overline{u})^{\ell} \ \overline{c})^{\ell'}$.

The composition of two contexts C and C', denoted by $C[C']$, is the context obtained by replacing $\Box$ with C' in C. For example, given $C = (a \rfloor \Box \ b)^{\ell}$, $C' = (c \rfloor \Box \ d \ e)^{\ell'}$, we get $C[C'] = (a \rfloor (c \rfloor \Box \ d \ e)^{\ell'} \ b)^{\ell}$.

2.3. Rewrite Rules and Qualitative Reduction Semantics

A rewrite rule is defined by a pair of compartments (possibly containing variables), which represent the patterns along which the system transformations are defined. The choice of defining rules at the level of compartments simplifies the formal treatment, allowing a uniform presentation of the system semantics.

In order to formally define the rewrite rules, we introduce the notion of open term (a term containing variables) and pattern (an open term that may be used as left part of a rewrite rule). To respect the syntax of terms, we distinguish between "wrap variables" which may occur only in compartment wraps (and can be replaced only by multisets of atoms) and "content variables" which may only occur in compartment contents or at top level (and can be replaced by arbitrary terms)

Let $\mathcal{V}_{\overline{\mathcal{T}}}$ be a set of *content variables*, ranged over by X, Y, Z, and $\mathcal{V}_{\overline{\mathcal{A}}}$ a set of *wrap variables*, ranged over by x, y, z such that $\mathcal{V}_{\overline{\mathcal{T}}} \cap \mathcal{V}_{\overline{\mathcal{A}}} = \emptyset$. We denote by $\mathcal{V}$ the set of all variables $\mathcal{V}_{\overline{\mathcal{T}}} \cup \mathcal{V}_{\overline{\mathcal{A}}}$, and with ρ any variable in $\mathcal{V}$. *Open terms*

are terms which may contain occurrences of wrap variables in compartment wraps and content variables in compartment contents. Similarly to terms, open terms are defined as multisets $\overline{o}$ of simple open terms defined in the following way:

$$\begin{array}{lcll} o & ::= & a \mid X & \mid \; (\overline{q} \rfloor \overline{o})^{\ell} \\ q & ::= & a \mid x & \end{array}$$

(i.e. $\overline{q}$ denotes a multiset formed only of atomic elements and wrap variables). Let $\mathcal{O}$ and $\overline{\mathcal{O}}$ denote the set of simple open terms and the set of open terms (multisets of simple open terms), respectively. An open term is *linear* if each variable occurs in it at most once.

An *instantiation* (or substitution) is defined as a partial function $\sigma : \mathcal{V} \to \overline{\mathcal{T}}$. An instantiation must preserve the type of variables, thus for $X \in \mathcal{V}_{\overline{\mathcal{T}}}$ and $x \in \mathcal{V}_{\overline{\mathcal{A}}}$ we have $\sigma(X) \in \overline{\mathcal{T}}$ and $\sigma(x) \in \overline{\mathcal{A}}$, respectively. Given $\overline{o} \in \overline{\mathcal{O}}$, with $\overline{o}\,\sigma$ we denote the term obtained by replacing each occurrence of each variable $\rho \in \mathcal{V}$ appearing in $\overline{o}$ with the corresponding term $\sigma(\rho)$.

Let Σ denote the set of all the possible instantiations and $\mathit{Var}(\overline{o})$ denote the set of variables appearing in $\overline{o} \in \overline{\mathcal{O}}$.

To define the rewrite rules, we first introduce the notion of patterns, which are particular simple open terms representing the left hand side of a rule. *Patterns*, ranged over by p, are the linear simple open terms defined in the following way:

$$p \quad ::= \quad (\overline{a}\ \ x \rfloor \overline{b}\ \ \overline{p}\ \ X)^{\ell}$$

where $\overline{a}$ and $\overline{b}$ are multisets of atoms, $\overline{p}$ is a multiset of pattern, x is a wrap variable, X is a content variable and the label ℓ is called the *type of the pattern*. The set of patterns is denoted by $\mathcal{P}$. Patterns are intended to match with compartments. Note that we force *exactly* one variable to occur in each compartment content and wrap. This prevents ambiguities in the instantiations needed to match a given compartment.[1] The linearity condition, in biological terms, corresponds to excluding that a transformation can depend on the presence of two (or more) identical (and generic) components in different compartments (see also [24]).

Some examples of patterns are:

[1]The presence of two (or more) variables in the same compartment content or wrap, like in $(x \rfloor a\ \ X\ \ Y)^{\ell}$, would introduce the possibility of matching the same path in different although equivalent ways. For instance we could match a term $(x \rfloor a\ \ a\ \ b\ \ b)^{\ell}$ with $X = a,\ Y = b\ \ b$ or $X = a\ \ b,\ Y = b$, etc.

- $(x \rfloor a\ \ b\ \ X)^\ell$ which matches with all compartments of type ℓ containing at least one occurrence of a and one of b.
- $(x \rfloor (a\ \ y \rfloor Y)^{\ell_1}\ \ X)^{\ell_2}$ which matches with compartments of type ℓ_2 containing a compartment of type ℓ_1 with at least an a on its wrap.

A *rewrite rule* is a pair (p, o), denoted by $p \longmapsto o$, where $p = (\overline{a}\ \ x \rfloor \overline{b}\ \ \overline{p}\ \ X)^\ell \in \mathcal{P}$ and $o = (\overline{q} \rfloor \overline{o})^\ell \in \mathcal{O}$ are such that $Var(o) \subseteq Var(p)$. Note that rewrite rules must respect the type of the involved compartments. A rewrite rule $p \longmapsto o$ then states that a compartment $p\sigma$, obtained by instantiating variables in p by some instantiation function σ, can be transformed into the compartment $o\sigma$ with the same type ℓ of p. Linearity is not required in the r.h.s. of a rule, thus allowing duplication or erasure.

A CWC *system* over a set $\mathcal{A}$ of atoms and a set $\mathcal{L}$ of labels is represented by a set $\mathcal{R}_{\mathcal{A},\mathcal{L}}$ ($\mathcal{R}$ for short when $\mathcal{A}$ and $\mathcal{L}$ are understood) of rewrite rules over $\mathcal{A}$ and $\mathcal{L}$.

A *transition* between two terms t and u of a CWC system $\mathcal{R}$ (denoted $t \longrightarrow u$) is defined by the following rule:

$$\frac{R = p \longmapsto o \in \mathcal{R} \quad \sigma \in \Sigma \quad C \in \mathcal{C}}{C[p\sigma] \xrightarrow{R} C[o\sigma]}$$

where $C[p\sigma] \equiv t$ and $C[o\sigma] \equiv u$.

In a rule $p \longmapsto o$ the pattern p represents a compartment containing the reactants of the reaction that will be simulated. The crucial point for determining an application of the rule to a term t is to find the compartments matching with p (i.e. the compartments in which the corresponding reaction can take place).

Note that the applicability of a rewrite rule depends on the type of the involved compartments but not on the context in which it occurs. This corresponds to the assumption that only the compartment type can influence the kind of reaction that takes place in them but not their position in the system.

Note that a same pattern can have more than one match in a term. Take for instance the term $(a \rfloor\ \ 2b\ \ 2c)^\ell\ \ (a\ \ b \rfloor 4b)^\ell$ and the pattern $p = (a\ \ x \rfloor b\ \ X)^\ell$. Then p matches the first compartment with the substitution $\sigma_1(x) = \bullet,\ \sigma_1(X) = b\ \ 2c$ and the second compartment with the substitution $\sigma_2(x) = b,\ \sigma_2(X) = 3b$

Remark 2.2. *For some rewrite rules $\ell : p \longmapsto o$ there may be, in general, different substitutions σ such that $p\sigma \equiv t$ (for some term t) but the results $o\sigma$ produced by them are different. Consider, for instance, the rewrite rule $(y \rfloor a\ \ (b\ \ x \rfloor X)^{\ell}\ \ Y)^{\top} \longmapsto (y \rfloor (a\ \ b\ \ x \rfloor X)^{\ell}\ \ Y)^{\top}$ modelling a catalysed membrane joining at top level. In this case, a term $t = (\bullet \rfloor a\ \ (b\ \ b \rfloor c)^{\ell}\ \ (b \rfloor c)^{\ell})^{\top}$ can make a transition in two different terms, depending on which membrane will be joined by the element a. Namely, $(\bullet \rfloor (a\ \ b\ \ b \rfloor c)^{\ell}\ \ (b \rfloor c)^{\ell})^{\top}$, given an instantiation σ such that $\sigma(x) = b$ and $(\bullet \rfloor (b\ \ b \rfloor c)^{\ell}\ \ (a\ \ b \rfloor c)^{\ell})^{\top}$, given an instantiation σ' such that $\sigma'(x) = \bullet$. We remark that this can happen only when compartments are involved in the rewriting. We will need to take it into account in the stochastic approach.*

Notation 2.3 (Rules that involve only content or wrap). *Usually, rules involve only the content or the wrap of a compartment. Moreover, in a rule $(\overline{a}\ \ x \rfloor \overline{b}\ \ \overline{p}\ \ X)^{\ell} \longmapsto (\overline{q} \rfloor \overline{o})^{\ell}$ very often X has single occurrence, at top level, in $\overline{o}$ and x in $\overline{q}$. We therefore introduce the following notations:*

- *$\ell : \overline{a}\ \ \overline{p} \longmapsto_{\mathsf{c}} \overline{o}$ (or simply $\ell : \overline{a}\ \ \overline{p} \longmapsto \overline{o}$) as a short for $(x \rfloor \overline{a}\ \ \overline{p}\ \ X)^{\ell} \longmapsto (x \rfloor \overline{o}\ \ X)^{\ell}$, and*
- *$\ell : \overline{a} \longmapsto_{\mathsf{w}} \overline{b}$ as a short for $(\overline{a}\ \ x \rfloor X)^{\ell} \longmapsto (\overline{b}\ \ x \rfloor X)^{\ell}$*

where x and X are canonically chosen variables not occurring in $\overline{a}$, $\overline{p}$, $\overline{o}$ or $\overline{b}$. Note that, according to Notation 2.1, rules of the shape $\top : \overline{a} \longmapsto_{\mathsf{w}} \overline{b}$ are forbidden (since the top level compartment must always have an empty wrap).

2.4. Modelling Guidelines

In this section we give some explanations and general hints about how CWC could be used to represent the behaviour of various biological systems. Here, entities are represented by terms of the rewrite system, and events by rewrite rules.

First of all, we should select the biomolecular entities of interest. Since we want to describe cells, we consider molecular populations and membranes. Molecular populations are groups of molecules that are in the same compartment of the cells and inside them. As we have said before, molecules can be of many types: we classify them as proteins, chemical moieties and other molecules.

Membranes are considered as elementary objects: we do not describe them at the level of the phospholipids they are made of. The only interesting properties of a membrane are that it may have a content (hence, create a compartment) and that in its phospholipid bilayer various proteins are embedded, which act for example as transporters and receptors. Since membranes are represented as multisets of the embedded structures, we are modelling a fluid mosaic in which the membranes become similar to a two-dimensional liquid where molecules can diffuse more or less freely [25].

Compartment labels are useful to identify the kind of a compartment and the properties which are common for each compartment of that type. For example, we may use compartment labels to denote the nuclei of a set of cells, the different organelles, etc..

Table 1 lists the guidelines for the abstraction into CWC rules of some basic biomolecular events, some of which will be used in our applications. Entities are associated with CWC terms: elementary objects (genes, domains, etc...) are modelled as atoms, molecular populations as CWC terms, and membranes as atom multisets. Biomolecular events are associated with CWC rewrite rules.

The simplest kind of event is the change of state of an elementary object. Then, there are interactions between molecules: in particular complexation, decomplexation and catalysis. Interactions could take place between simple molecules, depicted as single symbols, or between membranes and molecules: for example a molecule may cross or join a membrane. There are also interactions between membranes: in this case there may be many kinds of interactions (fusion, vesicle dynamics, etc. . .). Finally, we can model a state change of a compartment (for example a cell moving onto a new phase during the cell cycle), by updating its label.[2] Changing a label of a compartment implies changing the set of rules applied to it. This can be used, e.g., to model the different activities of a cell during the different phases of its cycle.

3. Quantitative Simulation Models for CWC

In this section we introduce two quantitative simulation methods for CWC based respectively on a stochastic simulation method and on the de-

[2]Note that, like in the other cases, this reaction is intended to take place in a compartment of a type ℓ''. Without the simplification made in Table 1 this rule should be written as $\ell'' : (x \rfloor X)^{\ell} \longmapsto (x \rfloor X)^{\ell'}$.

Biomolecular Event	**CWC Rewrite Rules**
State change (in content)	$a \longmapsto_{\mathsf{c}} b$
State change (on membrane)	$a \longmapsto_{\mathsf{w}} b$
Complexation (in content)	$a\ b \longmapsto_{\mathsf{c}} c$
Complexation (on membrane)	$a\ b \longmapsto_{\mathsf{w}} c$ $a\ (b\ x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (c\ x \rfloor X)^{\ell}$ $(b\ x \rfloor a\ X)^{\ell} \longmapsto_{\mathsf{c}} (c\ x \rfloor X)^{\ell}$
Decomplexation (in content)	$c \longmapsto_{\mathsf{c}} a\ b$
Decomplexation (on membrane)	$c \longmapsto_{\mathsf{w}} a\ b$ $(c\ x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} a\ (b\ x \rfloor X)^{\ell}$ $(c\ x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (b\ x \rfloor a\ X)^{\ell}$
Membrane crossing	$a\ (x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (x \rfloor a\ X)^{\ell}$ $(x \rfloor a\ X)^{\ell} \longmapsto_{\mathsf{c}} a\ (x \rfloor X)^{\ell}$
Catalysed membrane crossing	$a\ (b\ x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (b\ x \rfloor a\ X)^{\ell}$ $(b\ x \rfloor a\ X)^{\ell} \longmapsto_{\mathsf{c}} a\ (b\ x \rfloor X)^{\ell}$
Membrane joining	$a\ (x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (a\ x \rfloor X)^{\ell}$ $(x \rfloor a\ X)^{\ell} \longmapsto_{\mathsf{c}} (a\ x \rfloor X)^{\ell}$
Catalysed membrane joining	$a\ (b\ x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (a\ b\ x \rfloor X)^{\ell}$ $(b\ x \rfloor a\ X)^{\ell} \longmapsto_{\mathsf{c}} (a\ b\ x \rfloor X)^{\ell}$ $(x \rfloor a\ b\ X)^{\ell} \longmapsto_{\mathsf{c}} (a\ x \rfloor b\ X)^{\ell}$
Compartment state change	$(x \rfloor X)^{\ell} \longmapsto_{\mathsf{c}} (x \rfloor X)^{\ell'}$

Table 1: Guidelines for modelling biomolecular events in CWC, written in the compact notation of 2.3. The types (labels) associated to the rules are omitted for simplicity

terministic solution of ordinary differential equations (ODE).

In order to make the formal framework suitable for modelling quantitative aspects of biological systems, each transition is usually associated with a numerical parameter characterising the kinetic rate of the corresponding reaction.

In a stochastic simulation algorithm, this parameter and the quantity of reagents involved contribute *stochastically* to determine the next state of the system and the time needed to reach it. The system is then described as a Continuous Time Markov Chain (CTMC) [26]. This allows to simulate its evolution by means of standard simulation algorithms (see e.g.[15]). Stochastic simulation techniques can be applied to all CWC systems but, in several cases, at a high computational cost. The deterministic method based on ODE is computationally more efficient, but can be applied, in general, only to systems in which compartments are absent or have a fixed, time-independent, structure. These two approaches, presented separately in this

section, will be integrated in the next section defining an hybrid simulation algorithm for CWC that keeps the generality of the stochastic approach but can reduce its computational cost exploiting, when possible, the efficiency of the ODE simulation method.

In our calculus we will associate to a reaction a *rate function* having a parameter depending on the overall content of the compartment in which the reaction takes place. This allows to tailor the reaction rates on the specific characteristics of the system, as for instance when representing nonlinear reactions as it happens for Michaelis–Menten kinetics, or to describe more complex interactions involving compartments that may not follow the standard mass action rate. These latter, more classical, collision based stochastic semantics (see [15]) can be encoded as a particular choice of the rate function (see Section 3.1.1). A similar approach is used in [27] to model reactions with inhibitors and catalysts in a single rule.

Obviously some care must be taken in the choice of the rate function: for instance it must be complete (defined on the domain of the application) and non-negative.

Definition 3.1. *A* quantitative *rewrite rule is a triple* (p, o, f)*, denoted* $p \overset{f}{\longmapsto} o$*, where* (p, o) *is a rewrite rule and* $f : \Sigma \to \mathbb{R}^{\geq 0}$ *is the* rate function *associated to the rule.*[3]

The rate function takes an instantiation σ as parameter. Such an instantiation models the actual compartment content determining the structure of the environment in which the l.h.s. of a rule matches and that may actively influence the rule application. Notice that, different instantiations that allow the l.h.s. p of a rule to match a term $\overline{t}$ can produce different outcomes which could determine different rates in the associated transitions.

In the following we will use the function $\textsc{Occ} : \mathcal{A} \times \overline{\mathcal{T}} \to \mathbb{N}$ to count the occurrences of an atom within the multiset of atoms at the top level of a term. Namely, $\textsc{Occ}(b, \overline{t})$ returns the number of occurrences of the atom b at the top level of $\overline{t}$.

[3]The value 0 in the codomain of f models the situations in which the given rule cannot be applied, for example when the particular environment conditions forbid the application of the rule.

Example 3.2. *Consider again the term given in Remark 2.2. If the rate function of the rewrite rule is defined as* $f(\sigma) = 0.0002 \cdot (\text{Occ}(b, \sigma(x)) + 1)$*, the initial term* $\bar{t}$ *results in* $(a\ \ b\ \ b \rfloor c)^\ell\ \ (b \rfloor c)^\ell$ *with a rate* 0.0004 *and in the term* $(b\ \ b \rfloor c)^\ell\ \ (a\ \ b \rfloor c)^\ell$ *with rate* 0.0002*.*

We already mentioned that equipping rewrite rules with a function leads to the definition of a stochastic semantics that can abstract from the classical one based on collision analysis (practical for very low level simulations, for example chemical interactions), and allows defining more complex rules (for higher simulation levels, for example cellular or tissue interactions) which might follow different probability distributions. An intuitive example could be a simple membrane interaction: in the presence of compartments, a system could not be considered, in general, as well stirred. In such a case, the classical collision based analysis could not always produce faithful simulations and more factors (encapsulated within the context in which a rule is applied) should be taken into account.

A *quantitative* CWC *system* over a set $\mathcal{A}$ of atoms and a set $\mathcal{L}$ of labels is represented by a set $\mathcal{R}_{\mathcal{A},\mathcal{L}}$ ($\mathcal{R}$ for short when $\mathcal{A}$ and $\mathcal{L}$ are understood) of quantitative rewrite rules over $\mathcal{A}$ and $\mathcal{L}$.

3.1. Stochastic Evolution

In the stochastic framework, the rate of a transition is used as the parameter of an exponential distribution modelling the time spent to complete the transition. A quantitative CWC system $\mathcal{R}$ defines a Continuous Time Markov Chain (CTMC) in which the rate of a transition $C[p\sigma] \xrightarrow{R} C[o\sigma]$ is given by $f(\sigma)$, where the rule $R = (p, o, f) \in \mathcal{R}$ is the quantitative rule which determines a transition.[4] The so defined CTMC determines the *stochastic reduction semantics* of CWC.

When applying a simulation algorithm to a CWC system we must take into account, at a given time, all the system transitions (with their associated rates) that are possible at that point. They are identified by:

- the rewrite rule applied;
- context which selects the compartment in which the rule is applied;

[4]When it does not give rise to ambiguities we omit the label R, identifying the rewrite rule, from the stochastic transition.

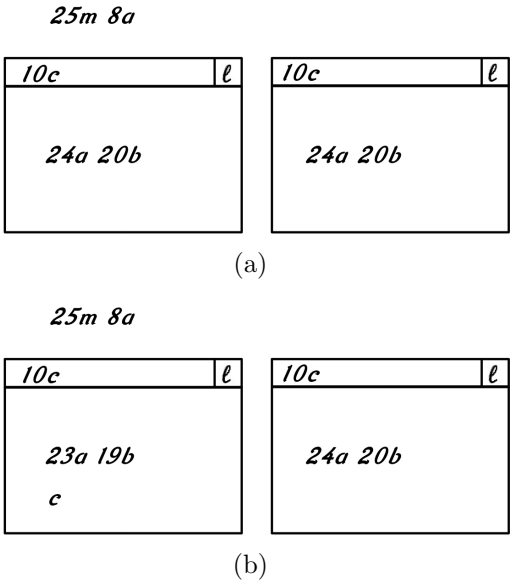


Figure 3: **(a)** $\overline{t}$; **(b)** $\overline{t'}$

- the outcome of the transition.

Remark 3.3. *We must take some care in identifying transitions involving compartments. For instance, if we consider the CWC term*

$$\overline{t} = 25m\ \ 8a\ \ (10c \rfloor 24a\ \ 20b)^{\ell}\ \ (10c \rfloor 24a\ \ 20b)^{\ell}$$

shown in Figure 3 (a) there are two compartments that are exactly the same. If we apply to $\overline{t}$ *the rule* $\ell : a\ \ b \longmapsto c$ *we obtain the term* $\overline{t'}$ *shown in Figure 3 (b). Actually, starting from* $\overline{t}$ *there are two compartments on which the rule can be applied, producing the same term* $\overline{t'}$ *(up to structural congruence). Although the transition is considered as one (up to structural congruence), the quantitative evolution must take this possibility into account by counting two transitions.*

From the transition rates we can define, following a standard simulation procedure [15], the exponential probability distribution of the moment in which the next reaction will take place and the probability distribution of the next transition that will take place.

In particular, given a term $\overline{t}$ and a global time δ, we first identify all the transitions $e_1, \ldots, e_M$ that can be applied to $\overline{t}$. Note that a transition

is identified by both the corresponding rewrite rule and the compartment in which it takes place (see also Remarks 2.2 and 3.3). Let $\pi_1, \ldots, \pi_M$ be the corresponding rates. Defining $\pi = \sum_{i=1}^{M} \pi_i$, the simulation procedure allows to determine, following Gillespie's direct method:

1. The time $\delta + \tau$ at which the next stochastic transition will occur, randomly chosen with τ exponentially distributed with parameter π;
2. The transition e_i that will occur at time $\delta + \tau$, randomly chosen with probability $\frac{\pi_i}{\pi}$.

We will detail this technique in Section 4, where we will present the hybrid simulation algorithm.

3.1.1. Mass Action Law

Gillespie's stochastic simulation algorithm is defined essentially for well stirred systems, confined to a constant volume and in thermal equilibrium at some constant temperature. In these conditions we can describe the system state by specifying only the molecular populations, ignoring the positions and velocities of the individual molecules. Different approaches such as Molecular Dynamics, Partial Differential Equations or Lattice-based methods are required in case of molecular crowding, anisotropy of the medium or canalisation.

We might restrict CWC in order to match Gillespie's framework. Namely, since we just need to deal with simple molecular populations, we might restrict terms to multisets of atoms.

The usual notation for chemical reactions can be expressed by:

$$n_1 a_1 + \ldots + n_\rho a_\rho \xrightarrow{k} n'_1 b_1 + \ldots + n'_\gamma b_\gamma \tag{1}$$

where, $a_i, \ldots, a_\rho$ and $b_i, \ldots, b_\gamma$ are the reagents and product molecules, respectively, $n_i, \ldots, n_\rho$ and $n'_i, \ldots, n'_\gamma$ are the stoichiometric coefficients and k is the kinetic constant.

We only consider now rewrite rules modelling chemical reactions as in reaction 1. A chemical reaction of the form 1 (that takes place within a compartment of type ℓ) can be expressed by the following CWC rewrite rule:

$$\ell : n_1 a_1 \ \ldots \ n_\rho a_\rho \xmapsto{f} n'_1 b_1 \ \ldots \ n'_\gamma b_\gamma \tag{2}$$

which is short for $(x \rfloor n_1 a_1 \ \ldots \ n_\rho a_\rho \ X)^\ell \stackrel{f}{\longmapsto} (x \rfloor n'_1 b_1 \ \ldots \ n'_\gamma b_\gamma \ X)^\ell$, where the rate function f of rule 2 should be suitably defined to model Gillespie's assumption of well stirred systems. In particular, the framework defined by Gillespie, based on molecular collision analysis, leads to binomial distributions of the reagents involved. Namely, we define the rate function f as:

$$f(\sigma) = \binom{\textsc{Occ}(a_1, \sigma(X)) + n_1}{n_1} \cdot \ldots \cdot \binom{\textsc{Occ}(a_\rho, \sigma(X)) + n_\rho}{n_\rho} \cdot k \tag{3}$$

where k is the kinetic constant of the modelled chemical reaction.

When the stoichiometric coefficients are low (e.g. ≤ 2) and the molecular populations are high, this can be approximated as:

$$\frac{(\textsc{Occ}(a_1, \sigma(X)) + n_1)^{n_1} \cdot \ldots \cdot (\textsc{Occ}(a_\rho, \sigma(X) + n_\rho))^{n_\rho}}{n_1 \cdot \ldots \cdot n_\rho} \cdot k \tag{4}$$

By construction, the following holds.

Fact 3.4. *Molecular populations defined as multisets of atoms that evolve according to a fixed set of transformations of the form given by reaction 1, represented by rule 2, interpret into the stochastic semantics of CWC the law of mass action within Gillespie's framework for the evolution of chemically reacting systems.*

Notation 3.5. *We will denote* biochemical rewrite rules *as defined in rule 2 with the simplified notation:*

$$\ell : \overline{a} \stackrel{k}{\longmapsto} \overline{b}$$

where $\overline{a}$ and $\overline{b}$ are multisets of atomic elements, and the rate function is represented by just the kinetic constant of the chemical reaction. We will use this notation also for non strictly biochemical rules in which the rate is calculated following the mass action law.

When the counting is done with the law of mass action, we will extend the simplified notation for biochemical rewrite rules (using a constant rate instead of a function) also for rules involving compartments:

$$\ell : \overline{a} \ \overline{p} \stackrel{k}{\longmapsto} \overline{o}$$

Example 3.6. *Given a term* $\bar{t} = 100a\ 50b\ 9c$*, the biochemical rewrite rule* $\top : 2a\ b \overset{0.002}{\longmapsto} c$ *generates the transition:* $\bar{t} \xrightarrow{k} 98a\ 49b\ 10c$*. The rate* k *is computed in the exact way (according to Equation 3) as* $k = \binom{100}{2} \cdot \binom{50}{1} \cdot 0.002 = 495 \cdot 50 \cdot 0.002 = 495$ *and in the approximate way (according to Equation 4) as* $k = \frac{100^2 \cdot 50}{2 \cdot 1} \cdot 0.002 = 500$*.*

3.1.2. Running Example: Stochastic Simulations

In order to illustrate the quantitative semantics of CWC we consider, as a running example, a toy case study derived from a Lotka-Volterra prey-predator dynamics. Let us consider the prey-predator oscillatory dynamics to be confined into a compartment *IN* interfered with rare events causing dramatic changes in the species evolution. A rare event like these could be schematically represented as a viral epidemic entering and exiting compartment *IN* with a relatively slow rate. Once inside the compartment *IN* the viral epidemic has the capability of killing some preys.

The set of CWC rules modelling this example is given in Figure 4. The preys (atoms a) and predators (atoms b) are located in compartment *IN* and follow a dynamics given by the rules (B_1),(B_2) and (B_3). The viral epidemic (atom *Vir*) enters and leaves the compartment with rules (N_1) and (N_2) respectively, and kills the preys with rule (B_4).

(N_1)	$\top : Vir\ (x \rfloor\ X)^{IN} \overset{0.03}{\longmapsto} (x \rfloor\ Vir\ X)^{IN}$
(N_2)	$\top : (x \rfloor\ Vir\ X)^{IN} \overset{0.1}{\longmapsto} Vir\ (x \rfloor\ X)^{IN}$
(B_1)	$IN : a\ X \overset{1}{\longmapsto} a\ a\ X$
(B_2)	$IN : a\ b\ X \overset{0.001}{\longmapsto} b\ b\ X$
(B_3)	$IN : b\ X \overset{1}{\longmapsto} X$
(B_4)	$IN : Vir\ a\ X \overset{0.1}{\longmapsto} Vir\ X$

Figure 4: CWC rules for the prey-predator dynamics

The simulations are performed for 60 time units, with the starting term:

$$Vir\ (\bullet \rfloor\ 1200a\ 1200b)^{IN}.$$

Several stochastic simulations of the toy case study were performed, showing different possible system evolutions of the dynamics of the species inside the compartment *IN* depending on the viral epidemic factor. Two of these runs are shown in Figure 5. A characteristic of this example is that the evolution

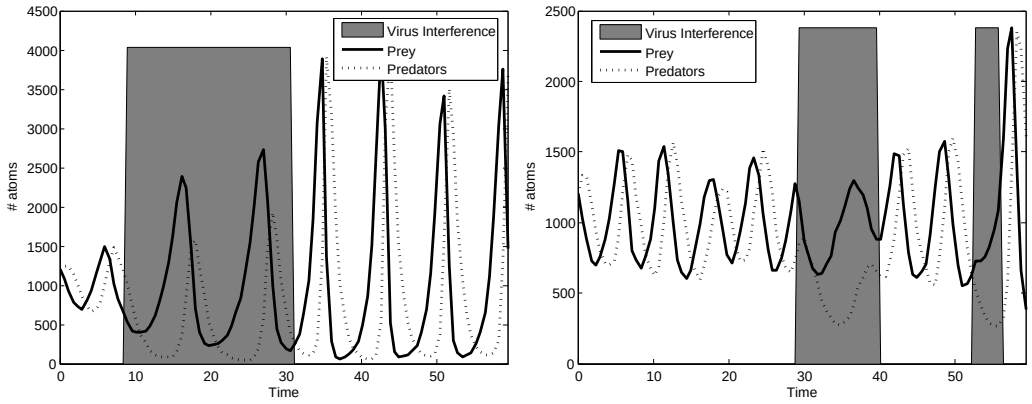


Figure 5: Two different runs obtained with a purely stochastic simulation showing the different behaviour of the dynamics of the species inside the compartment *IN* depending on the viral epidemic factor

of the system is strongly determined by the virus epidemic random event that can change dramatically the dynamics of the species.

3.2. Deterministic Evolution

The standard way to express the evolution of a biochemical system is via ODEs. We define the deterministic reduction semantics for the subset of CWC quantitative rewrite rules, presented in Section 3.1.1, that we called biochemical rewrite rules and express simple biochemical reactions.

Biochemical reactions are local to a single compartment. Reactions that invoke and/or change the structure of compartments cannot be expressed with biochemical rewrite rules. Actually, referring to Table 1, we notice that biochemical rewrite rules can be used to model state change, complexation and decomplexation: these are exactly the kinds of reactions naturally eligible to be simulated with ODEs.

A CWC system $\mathcal{R}$ consisting of r biochemical rewrite rules represents a system of r biochemical reactions. Its deterministic semantics is defined by extracting from $\mathcal{R}$ a system of ODEs to be used for simulating the evolution of the involved multisets of atoms [28]. For every label ℓ, let

- $a_1, \ldots, a_{n_\ell}$ denote the n_ℓ species of atoms that may occur at top level within a compartment of type ℓ, and
- $\mathcal{R}_\ell$ denote the set of rules with label ℓ.

The i-th rule in the set $\mathcal{R}_\ell$ is denoted by

$$\ell : \bar{a}_i \overset{k_i}{\longmapsto}_{\mathsf{c}} \bar{b}_i \quad i = 1, 2, \ldots, |\mathcal{R}_\ell|$$

For all species a_j $(j = 1, 2, \ldots, n_\ell)$ let $\alpha^-_{i,j}$ be the number of atoms of species a_j consumed by the i-th rule and $\alpha^+_{i,j}$ the number of atoms of species a_j produced by the i-th rule. The $n_\ell \times |\mathcal{R}_\ell|$ stoichiometric matrix Λ_ℓ is defined by $\nu_{i,j} = \alpha^+_{i,j} - \alpha^-_{i,j}$.[5]

Let $[a]$ denote the number of the atoms of species a occurring at top level in a given compartment of type ℓ. If $\bar{a}_i = n_{i_1} a_{i_1} \ldots n_{i_{r_i}} a_{i_{r_i}} (r_i \geq 1)$, the evolution of the given compartment of type ℓ is modelled by the following system of ODEs:

$$\ell : \frac{d[a_j]}{dt} = \sum_{i=1}^{|\mathcal{R}_\ell|} \nu_{i,j} \cdot k_i \cdot [a_{i_1}]^{n_{i_1}} \cdot \ldots \cdot [a_{i_{r_i}}]^{n_{i_{r_i}}}$$

Computationally, ODEs are well studied and understood. They can be solved using a variety of numerical methods, from the Euler method to higher-order Runge-Kutta methods or stiff methods, many of which are readily available in software packages that can be easily incorporated into existing simulation code. In all the examples presented in this paper we use a GNU library implementing an explicit embedded Runge-Kutta Prince-Dormand method. We do not need, however, to employ a particular library. Actually, we could change method according to the features of the particular system under consideration.

3.2.1. Running Example: Deterministic Simulations

To perform a deterministic simulation of the toy case study we have to remodel the Lotka-Volterra dynamics presented in Section 3.1.2 by *eliminating* the stochastic influence of the virus epidemics. The equation system governing the population evolution is:

$$\frac{da}{dt} = a - 10^{-3} a \cdot b$$
$$\frac{db}{dt} = -b + 10^{-3} a \cdot b$$

[5]Many of the $\alpha^-_{i,j}$, $\alpha^+_{i,j}$ are usually 0.

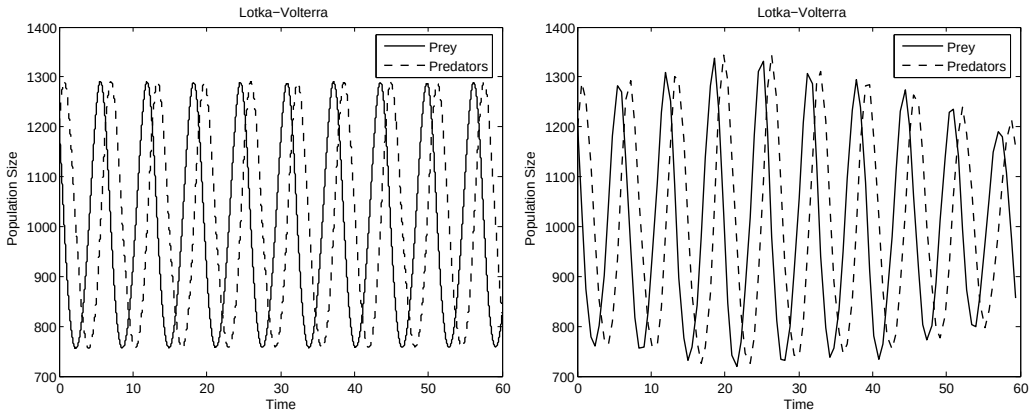


Figure 6: Deterministic simulation of the Lotka-Volterra dynamics (left figure) and the mean of 100 stochastic simulations (right figure)

with an initial condition of $a(0) = b(0) = 1200$.

The evolution of the system is represented in Figure 6. Notice that the mean of 100 stochastic simulations of the Lotka-Volterra dynamics without the viral epidemics interference tends to the deterministic simulation.

4. Hybrid Evolution

The stochastic approach is based on a probabilistic simulation method that manages the evolution of exact integer quantities and often requires a huge computational time to complete a simulation. The ODEs numerical approach computes a unique deterministic and fractional evolution of the species involved in the system and achieves very efficient computations.

A numerical background for the hybrid evolution in which fast reactions are approximated using a deterministic evolution while slow reactions are treated as stochastic events can be found in [29, 30]. This schema allows to accurately solve fast reactions using an ODE solver at the thermodynamic limit. This condition is of course ideal and unattainable in biological systems. However, the analytical knowledge of the system allows the use of this approximation if the variation of the species affected by slow reactions are almost insensitive with respect to species affected by fast reactions.

In this section we adapt this approximation method within CWC, defining a hybrid simulation technique. Given a CWC system $\mathcal{R}$ we partition it into a set of biochemical rewrite rules $\mathcal{B}$ and a set of non-biochemical rewrite rules $\mathcal{N}$. Rules in $\mathcal{N}$ are always applied by using the stochastic method.

Rules in $\mathcal{B}$ might be applied with the ODEs approach. In general $\mathcal{B}$ might contain both rules that model evolution of large numbers of molecules according to very fast reactions (whose execution is suitable to be correctly computed with ODEs) and rules that model very slow reactions or reactions that involve a very small number of reagents. In the latter case it is convenient to compute the execution of the associated rule according to the stochastic approach.

According to the state of the system, a rule might be dynamically interpreted either as stochastic or deterministic. For instance, during a simulation, it might happen that a given biochemical rewrite rule $\ell : \bar{a}_i \stackrel{k_i}{\longmapsto} \bar{b}_i \in \mathcal{B}$ is applied initially according to the stochastic semantics, since the associated compartment contains a very small number of reagents. After the system has evolved for some time, however, the number of the reagents involved in the rule can be substantially increased and it becomes convenient to model the corresponding reaction according to the deterministic approach.

Actually, at the beginning of each simulation step we build, for each compartment in the term, a system of ODEs for the simulation of the biochemical rules in that compartment which (1) are sufficiently fast and (2) involve a sufficient number of reagents. For the remaining rules the evolution is determined by the stochastic simulation algorithm.

In order to describe the hybrid semantics we assume that, given a CWC term $\bar{t}$, each compartment of $\bar{t}$ is univocally identified by an index ι. The index of the (implicit) compartment at the top level will be denoted by ι_0. The *biochemical reagents* of a compartment $(\overline{a} \rfloor \bar{t})^{\ell}$ with index ι, written $\mathrm{BR}(\iota)$, are expressed by the multiset of the atomic elements appearing in the top level of $\bar{t}$. For example, given the term

$$\bar{t} = 2a \ \ (b \rfloor (c \ \ d \rfloor \bullet)^{\ell'} \ \ e)^{\ell} \ \ (b \rfloor d \ \ e)^{\ell}$$

and assuming that the compartment $(b \rfloor (c \ \ d \rfloor \bullet)^{\ell'} \ \ e)^{\ell}$ has index ι_1, the compartment $(c \ \ d \rfloor \bullet)^{\ell'}$ has index ι_2 and the compartment $(b \rfloor d \ \ e)^{\ell}$ has index ι_3, we have that $\mathrm{BR}(\iota_0) = 2a$, $\mathrm{BR}(\iota_1) = e$, $\mathrm{BR}(\iota_2) = \bullet$ and $\mathrm{BR}(\iota_3) = d \ \ e$ where ι_0 is the index of the to level compartment.

A basic point of our hybrid approach is the criterion to determine, at each computation stage, the reductions to compute in the stochastic or in the deterministic way. In this paper we have chosen simply to put a threshold on the number of possible reagents and on the speed of the reaction, but other more sophisticated criteria should be investigated.

Let $\bar{t}$ denote the whole term and let I denote the set of compartment indexes occurring in $\bar{t}$.

1. For each compartment $\iota \in I$:
   - Let ℓ be the label of ι, let $\mathcal{S}_\iota = \mathcal{B}_\ell$ and let $\mathcal{D}_\iota = \emptyset$.
   - For each biochemical rule $B_i = \ell : \bar{a}_i \overset{k_i}{\longmapsto} \bar{b}_i \in \mathcal{B}_\ell$ let $\bar{a}_i = n_{i_1} a_{i_1} \ldots n_{i_{r_i}} a_{i_{r_i}} (r_i \geq 1)$ and let $[a_{i_j}]_\iota$ denote the number of a_{i_j} atoms occurring in $\mathrm{BR}(\iota)$. Let $\pi_i^\iota = k_i \cdot ([a_{i_1}]_\iota^{n_{i_1}} / n_{i_1}) \cdot \ldots \cdot ([a_{i_{r_i}}]_\iota^{n_{i_{r_i}}} / n_{i_{r_i}})$ be the rate of the rule B_i in the compartment ι. If $\pi_i^\iota > \phi$ and $\min_{j=1}^{r_i} [a_{i_j}]_\iota > \psi$ remove B_i from $\mathcal{S}_\iota$ and put it into $\mathcal{D}_\iota$.
2. Considering the rules in $\bigcup_{\iota \in I} \mathcal{S}_\iota \cup \mathcal{N}$ selected according to Gillespie's method and to the semantics given in Section 3.1 a stochastic transition step $C[p\sigma] \xrightarrow{f(\sigma)} C[o\sigma]$, where $R = \ell : p \overset{f}{\longmapsto} o \in \mathcal{S}_{\iota'} \cup \mathcal{N}_\ell$.
   Let τ be the corresponding time interval.†
3. For each compartment ι in I:
   - Let $\mathcal{E}_\iota$ denote the system of ODEs for the rules in $\mathcal{D}_\iota$ in the compartment ι as explained in Section 3.2 without considering, in the compartment ι' where the stochastic transition step takes place, the active reagents appearing in the left part p of the stochastically applied rule. (If $\mathcal{D}_\iota = \emptyset$ then $\mathcal{E}_\iota = \emptyset$.)
   - Apply the system of ODEs $\mathcal{E}_\iota$ to the biochemical reagents $\mathrm{BR}(\iota)$ of the compartment for a time duration τ.
4. Update the term $\bar{t}$ according to the right part o of the chosen stochastic rule and to the applications of the systems of ODEs.

† It may happen that no rule in $(\bigcup_{\iota \in I} \mathcal{S}_\iota) \cup \mathcal{N}$ is applicable. In such cases the evolution of the system must be determined for some time τ according to the deterministic semantics only. In our implementation we choose as τ the maximum time calculated by Gillespie's algorithm for each of the applicable biochemical rules in $\bigcup_{\iota \in I} \mathcal{D}_\iota$.

Figure 7: Steps performed by an hybrid simulation iteration

These two thresholds are named:

ϕ defining a minimum rate to consider a rule deterministically,

ψ defining a minimum quantity to consider the involved rule deterministically.

A method for the choice of ψ and ϕ is presented in Appendix A. Notice that these thresholds must be hold during all deterministic evolution in order to validate the hybrid approach.

Given as input a term $\bar{t}$ to reduce, a rate threshold ϕ and a quantity threshold ψ, each iteration of the hybrid reduction semantics performs the four steps listed in Figure 7. For every label ℓ, the subsets of $\mathcal{B}$ and $\mathcal{N}$ containing the rules with label ℓ are denoted by $\mathcal{B}_\ell$ and $\mathcal{N}_\ell$, respectively. The first step identifies, for each compartment $\iota \in I$ (where I is the set of all compartment indexes occurring in $\bar{t}$), two disjoint sets of biochemical rules, namely $\mathcal{D}_\iota$ (to be applied deterministically) and $\mathcal{S}_\iota$ (to be applied, together with the rules in $\mathcal{N}$, according to the stochastic method). The second step selects, considering only the rules in $\bigcup_{\iota \in I} \mathcal{S}_\iota \cup \mathcal{N}$, the next rule to be applied stochastically. When the stochastic transition is chosen, we "lock" the reagents involved in such a reaction. They will not contribute to the ODEs evolution (since they are already active in the stochastic sense), and their product is added at the end of the stochastic time step. The third step computes a system of ODEs $\mathcal{E}_\iota$ for each compartment $\iota \in I$ and applies the ODEs for the time duration selected by the stochastic step. The calculation are performed in the same units as the stochastic computation, namely in terms of number of molecules and time. The fourth step updates the terms according to the results of the simulation.

Note that during the deterministic step there was the implicit assumption that the evolution of the species involved in fast reactions, calculated with the ODEs, did not significantly alter the propensity of the slow reactions changing their priority with respect to fast reactions. Omission of this hypothesis requires a control on the evolutionary trajectory of species calculated by the ODEs in order to stop deterministic evolution if the priorities of the reactions have changed its structure.

Before computing the next stochastic step after the deterministic one, the fractional numbers of molecules computed by the ODEs need to be converted into integer numbers. As suggested by [31], fractions can be handled

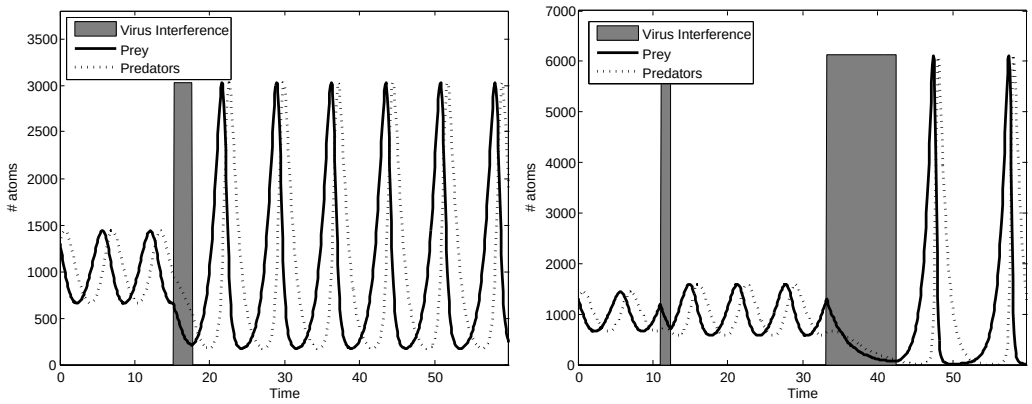


Figure 8: Two different runs of the hybrid simulations showing the different behaviour of the dynamics of the species inside the compartment *IN* depending on the virus epidemic factor

by the two following methods: rounding to the nearest integer and probabilistic rounding. Rounding to the nearest integer may introduce a bias. For example, if the number of molecules is 200.3, it will always be rounded to 200 and this bias may be amplified during the whole simulation. With a probabilistic rounding, instead, a molecular number of 200.3 is rounded up to 201 with probability 30%, and down to 200 with probability 70%. Using such a technique, the average number of the molecular counts becomes identical to the one obtained using deterministic calculations.

In general, if reactions are fast enough, the deterministic ODEs simulation approximate better the exact stochastic simulations. This is the idea behind the use of the threshold ϕ. The use of ψ, instead, allows to prevent the rounding approximation error that may derive when we are dealing with species of few elements. Combined together, the thresholds ϕ and ψ affect the level of approximation we want to use in our simulations. Notice that with $\phi = +\infty$ all reactions will be considered *too slow* and the simulation will be computed with the purely stochastic method.

4.1. Running Example: Hybrid Simulations

Hybrid simulations of the system presented in Section 3.1.2 were performed by using thresholds $\phi = 0.5$ and $\psi = 10$. In Figure 8 we report two runs of the hybrid simulations showing two different evolutions of the species.

Notice that until the viral epidemic factor does not reach the compartment, the prey-predator dynamics (rules (B_1),(B_2) and (B_3) of the model in

Figure 4) is treated deterministically since high propensities drive the subsystem. Conversely, a rare viral outbreak introduces a stochastic change inside the compartment influencing the amount of preys (and, consequently, the predators dynamics) through the rule (B_4).

The long period of evolution, waiting for the epidemic rare event given by the rules (N_1) and (N_2), allows an efficient deterministic computation. In this case, other approximation techniques, such as τ-leaping, could not take the full advantage of ODEs since in such an oscillating scenario the propensities of the fast rules change rapidly and repeatedly over the time intervals and the length of leap is bound by the frequency of the oscillations.

A comparison of the computational time needed to perform 100 runs using the hybrid method versus the stochastic simulation technique provided a dramatic improvement on the computational effort. The relative speed-up, measured as the ratio of the computational run time of the stochastic simulations to that of hybrid simulations, was about 80.

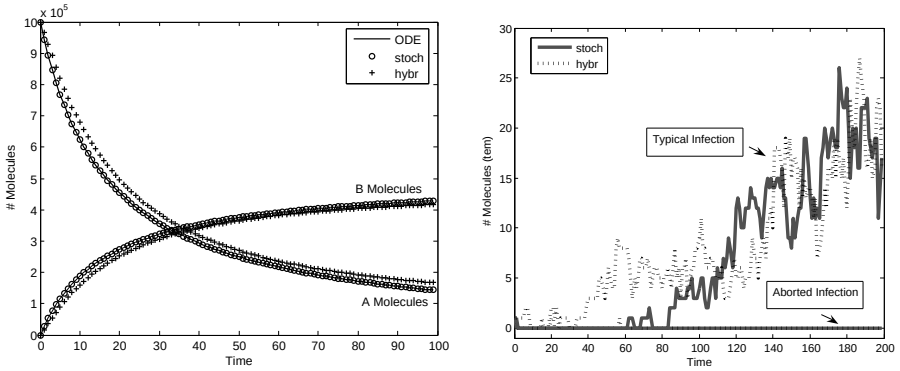
4.2. Further Examples

In this section the results of the hybrid algorithm, when compared with the stochastic simulation algorithm, are presented on two benchmark models collected from the literature. The two models are a simple crystallisation system [16, 30, 17] and a model of intracellular viral infection [32, 30, 17] whose quantitative behaviours have been accurately reproduced by the hybrid simulation.

The first benchmark is a simplified model for the crystallisation of species A, consisting of two reactions, one occurring many more times than the other. The rules and rates are taken from [16], Table 3. Our partitioning scheme, setting the thresholds $\phi = 6 \cdot 10^{-4}$ and $\psi = 100$, allow to classify the fast reaction as continuous and the other as discrete.

Figure 9(a) compares the averages of species A and B computed by 100 runs of the hybrid and pure stochastic simulation algorithms. In this example an implicit step (instead of an explicit one) for the ODEs solver could further improve the approximation.

The second benchmark is a general model of the infection of a cell by a virus. This model includes the genomic and template viral nucleic acids (*gen* and *tem* respectively) and the viral structural protein (*struct*). This model has two interesting features: (i) the three components of the model exhibit fluctuations that vary by several orders of magnitude; and (ii) the model solution exhibits a bimodal distribution either a "typical" infection in



(a) Simple Crystallisation results (b) Virus infection results

Figure 9: Comparison of the hybrid and pure stochastic simulations on the examples. Fig. (a) shows the average of 100 runs of the hybrid and pure stochastic simulations compared with the ODE solution. Fig. (b) shows two exemplificative solutions computed by the hybrid and pure stochastic simulations.

which all species become populated, or an "aborted" infection in which all species are eliminated from the cell. The rules and rates are taken from [32], Table 1.

Figure 9(b) shows two exemplificative solutions computed by the hybrid (using $\phi = 10$ and $\psi = 25$) and pure stochastic simulations corresponding to a "typical" and "aborted" infection of species *tem*.

Based on 100 runs of the hybrid and pure stochastic simulation algorithms the speedup was around 500% for the first benchmark and around 200% for the second benchmark. Note, however, that the presence of data structures to handle compartmentalisation introduces in our simulator an overhead which can reduce the efficiency differences between the two algorithms with respect to other hybrid simulators.

5. A Real Model of Different Cellular Fate

To assess the soundness and efficiency of our hybrid approach on a real biological problem we decided to apply it to a well known system where stochastic effects play a fundamental role in determining its development: the HIV-1 transactivation mechanism.

After a cell has been infected, the retrotransposed DNA of the virus is integrated in the host genome and it begins its transcription into *mRNA* and then the translation to yield viral proteins; the initial speed of this

mechanism, however, is fairly slow. The speedup of the viral production process is determined by a regulation system driven by the viral protein *TAT*: this protein is capable of binding cellular factors of the host to produce the *pTEFb* complex which in its acetylated form is able to bind to the integrated viral genome and speed up the transcription machinery, thus ending in more viral proteins and, therefore, more *TAT*, determining a positive loop.

The time scale during which this loop is triggered is affected by several factors e.g. the initial low *TAT* production and the rate of its degradation, the equilibrium between the active (acetylated) and inactive form of *pTEFb*. As a consequence, the stochastic fluctuations in this events are considered pivotal in determining when viral proteins are produced in a sufficient quantity to determine cellular lysis and viral spreading. Since HIV is known to stay dormant and inactive in some types of cells and since the time between the infection and the high viral production rate related to the active phase of AIDS is variable, this transactivation mechanism is of great interest. Viral latency is also believed to be the cause of the persistent low level viremia observed in patients treated with antiretroviral therapies, therefore understanding its molecular bases is fundamental in order to be able to circumvent it and hopefully find a way to completely eradicate the virus avoiding lifelong therapies [33].

We decided to follow the direction taken in a previous study about this system (see [34]), in which an experimental setting is developed where a fluorescent protein, *GFP*, is the only one encoded by an engineered viral genome, along with *TAT*. In [34] they were able to identify different evolutions in the *GFP* level over time: cellular clones with exactly the same genome showed two different behaviours, one produced a high quantity of *GFP* (they called it "bright") and the other one with very little *GFP* ("off"). This work also reported that a purely stochastic simulation was able to individuate this bistability; a later work (see [17]) confirmed these results performing purely stochastic and mixed deterministic-stochastic simulations.

Since CWC systems are able to represent compartments, we slightly modified the original set of rules used in these works to explicitly represent the cytoplasm and the nucleus of an infected cell; all the kinetic rates were maintained, the one for *TAT* nuclear import has been determined from the literature (see [35]). The set of rules we adopted is given in Figure 10, where we refer to the cytoplasm as the $\top$ compartment while η is the label used for the nucleus. As regards the rules: (B_1) represents the slow basal rate of viral *mRNA* transcription; (N_1) describes the *mRNA* export from

$$
\begin{array}{ll}
(B_1) & \eta : LTR \overset{10^{-9}}{\longmapsto} LTR\ \ mRNA \\
(N_1) & \top : (x \rfloor mRNA\ \ X)^{\eta} \overset{7.2 \cdot 10^{-4}}{\longmapsto} (x \rfloor X)^{\eta}\ \ mRNA \\
(B_2) & \top : mRNA \overset{0.5}{\longmapsto} mRNA\ \ GFP \\
(B_3) & \top : mRNA \overset{1.32 \cdot 10^{-3}}{\longmapsto} mRNA\ \ TAT \\
(N_2) & \top : (x \rfloor X)^{\eta}\ \ TAT \overset{8.5 \cdot 10^{-3}}{\longmapsto} (x \rfloor TAT\ \ X)^{\eta} \\
(N_3) & \top : (x \rfloor TAT\ \ X)^{\eta} \overset{7.2 \cdot 10^{-4}}{\longmapsto} (x \rfloor X)^{\eta}\ \ TAT \\
(B_4) & \eta : TAT\ \ LTR \overset{1.5 \cdot 10^{-4}}{\longmapsto} pTEFb \\
(B_5) & \eta : pTEFb \overset{17 \cdot 10^{-3}}{\longmapsto} TAT\ \ LTR \\
(B_6) & \eta : pTEFb \overset{10^{-3}}{\longmapsto} pTEFb\_ac \\
(B_7) & \eta : pTEFb\_ac \overset{0.13}{\longmapsto} pTEFb \\
(B_8) & \eta : pTEFb\_ac \overset{0.1}{\mapsto} LTR\ \ TAT\ \ mRNA \\
(B_9) & \top : GFP \overset{3.01 \cdot 10^{-6}}{\longmapsto} \bullet \\
(B_{10}) & \top : TAT \overset{4.3 \cdot 10^{-5}}{\longmapsto} \bullet \\
(B_{11}) & \top, \eta : mRNA \overset{4.8 \cdot 10^{-5}}{\longmapsto} \bullet
\end{array}
$$

Figure 10: CWC rules for the TAT transactivation system

the nucleus to the cytoplasm; (B_2) and (B_3) express the translations of this $mRNA$ into GFP and TAT proteins, respectively; (N_2) and (N_3) represent the nuclear import and export of TAT; (B_4) and (B_5) model the binding and unbinding of TAT with (not represented here) host cellular factors and the viral genome portion LTR that forms $pTEFb$ which, when acetylated (by rule (B_6)) determines an higher transcriptional activity, which is represented in (B_8) by the unbinding that releases LTR and TAT and creates an $mRNA$ molecule (note the higher rate with respect to (B_1)); (B_7) represents the $pTEFb$ deacetylation and (B_9), (B_{10}) and (B_{11}) model the degradation processes of the proteins and the $mRNA$ (note that $mRNA$ degrades both in the nucleus and in the cytoplasm, the other proteins only degrade in the cytoplasm; also note how the compartment labelling mechanism allows to express this fact in a simple and elegant way).

We performed 200 purely stochastic simulations (i.e. setting $\phi = +\infty$) and 200 hybrid simulations (using $\phi = 0.5$ and $\psi = 10$). The initial term of

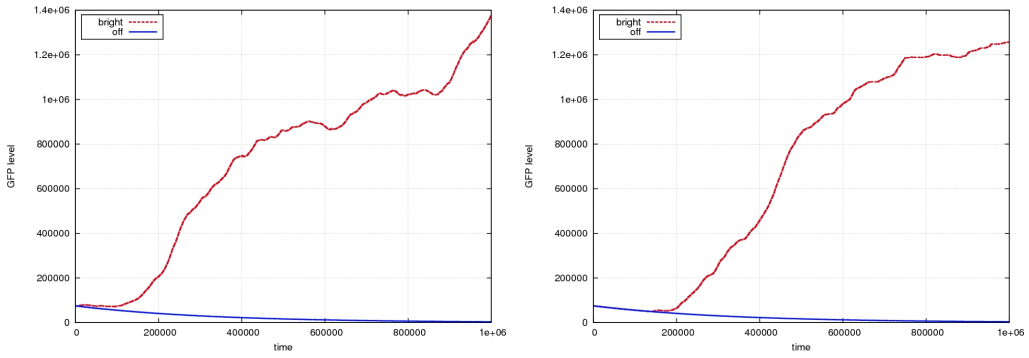


Figure 11: Two different simulations pure stochastic (on the left) and hybrid (on the right) started with the same parameters: "bright" and "off" behaviour

our simulations is represented by the CWC term

$$75000\,GFP\ 5\,TAT\ (\bullet \rfloor\, LTR)^{\eta},$$

while the time interval of our simulations has been fixed to 10^6 seconds (the same parameters are used in [34, 17]). Both our stochastic and hybrid simulations clearly showed the two possible evolutions of the system which correspond to the "bright" and the "off" cellular populations (in order to display the double destiny, almost all the biochemical rewrite rules have to be simulated with the stochastic approach). Figure 11 shows two exemplificative runs of "bright" and "off" behaviour resulting from pure stochastic and hybrid simulations. Figure 12 reports the centroids of two clusters obtained using the k-means algorithm [36] on the 200 runs performed using the pure stochastic and hybrid simulations using a sampling step size $\Delta t = 10$. In both cases a 4% of simulations showed an "off" behaviour and this results confirm the statistical analysis reported in [17]. The stochasticity of the centroids corresponding to the "off" behaviour is due to the few points belonging to the cluster. As could be seen in Figures 11 and 12, the hybrid simulations are comparable to the purely stochastic ones and, even with the relatively high thresholds used in this particular case, the hybrid simulations were computationally more efficient (almost 40% faster).[6]

[6]Comparisons are made using the same stochastic engine, in both cases with no particular optimisation.

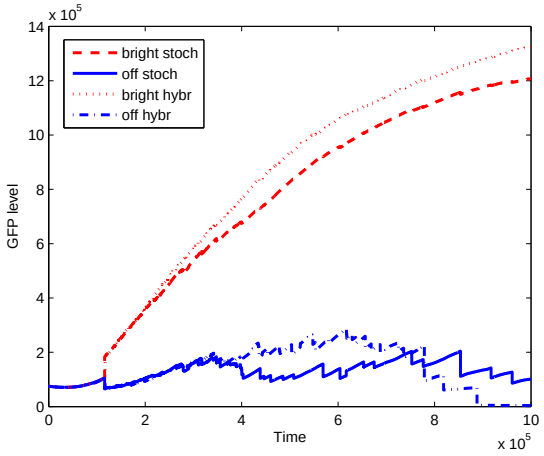


Figure 12: Centroids comparison of the two clustered results obtained from 200 simulations pure stochastic and hybrid showing the "bright" and "off" behaviour

6. Related Work

In this section we will put our paper in the framework of qualitative, stochastic and hybrid models for the description and analysis of biological systems.

6.1. Qualitative Models

In the last few years many formalisms originally developed by computer scientists to model systems of interacting components have been applied to Biology. Among these, there are Petri Nets [4], Hybrid Systems [3], and the π-calculus [8, 37, 38]. Moreover, new formalisms have been defined for describing biomolecular and membrane interactions [7, 9, 39, 5, 10, 40]. Others, such as P systems [6], have been proposed as biologically inspired computational models and have been later applied to the description of biological systems.

The π-calculus and new calculi based on it [10, 40] have been particularly successful in the description of biological systems, as they allow describing systems in a compositional manner. Interactions of biological components are modelled as communications on channels whose names can be passed; sharing names of private channels allows describing biological compartments.

These calculi offer very low-level interaction primitives, but may cause the models to become very large and difficult to read. Calculi such as those

proposed in [9, 39, 5] give a more abstract description of systems and offer special biologically motivated operators. However, they are often specialised to the description of some particular kinds of phenomena such as membrane interactions or protein interactions.

P systems [6] have a simple notation and are not specialised to the description of a particular class of systems, but they are still not completely general. For instance, it is possible to describe biological membranes and the movement of molecules across membranes, and there are some variants able to describe also more complex membrane activities. However, the formalism is not so flexible to allow describing easily new activities observed on membranes without extending the formalism to model such activities.

CWC can describe situations that cannot be easily captured by the previously mentioned formalisms, which consider membranes as atomic objects (extensions of P systems with objects on membranes can be found in [41, 42]). Representing the membrane structure as a multiset of the elements of interest allows the definition of different functionalities depending on the type and the number of elements on the membrane itself.

Danos and Laneve [5] proposed the κ-calculus. This formalism is based on graph rewriting where the behaviour of processes (compounds) and of set of processes (solutions) is given by a set of rewrite rules which account for, e.g., activation, synthesis and complexation by explicitly modelling the binding sites of a protein.

CLS [7] has no explicit way to model protein domains (however they can be encoded, and a variant with explicit binding has been defined in [43, 44]), but accounts for an explicit mechanism (the *looping sequences*) to deal with compartments and membranes. Thus, while the κ-calculus seems more suitable to model protein interactions, CLS allows for a more natural description of membrane interactions. Another feature lacking in many other formalisms is the capacity to express ordered sequences of elements. While we might encode ordered structures in CWC with nested compartments, CLS offers such a feature in an explicit way, thus allowing to naturally operate over proteins or DNA fragments which should be frequently defined as ordered sequences of elements.

6.2. Stochastic Models

Our stochastic semantics is defined in terms of the collision-based paradigm introduced by Gillespie. A similar approach is taken in the quantitative variant of the κ-calculus ([45]) and in BioSPi ([8]). Motivated by the law of

mass action, here we need to count the number of the reactants present in a system in order to compute the exact rate of a reaction. In [11], a stochastic semantics for bigraphs has been developed. An application in the field of systems biology has been provided by modelling a process of membrane budding. Compartmentalised stochastic simulations, addressing the problem of dynamic structure, have also been investigated in the domain of P systems, see, e.g., [46, 47].

In the following, we would like to compare our work with two closer ones, namely [12] and [48].

A stochastic semantics for CLS (SCLS) has been defined in [12]. Such a semantics computes the transition rates by resorting to a complete counting mechanism to detect all the possible occurrences of patterns within a term. Our rules, similar to what happens in [47] for P systems, in Bio-PEPA [49], in [48] for a variant of the ambient calculus and in [27] for CLS, are equipped with rate functions, rather than with rate constants. Such functions may allow the definition of kinetics that are more complex than the standard mass-action ones. In particular, equipping the rewrite rules of our calculus with a function leads to the definition of a stochastic semantics that can abstract from the classical one based on collision analysis (based on constant rates and practical for a very low level analysis, for example chemical interactions), and allows defining more complex rules (for higher simulation levels, for example cellular or tissue interactions) which might follow different probability distributions.

Indeed, CWC has been originally proposed in [14] as a variant of (S)CLS with the aim of strongly simplifying the development of efficient automatic tools for the analysis of biological systems, while keeping the same expressiveness. The main simplification consists in the removal of the sequencing operator, thus lightening the formal treatment of the patterns to be matched in a term (whose complexity in SCLS is strongly affected by the variables matching in the sequences). Then, in [23] we extended CWC with compartment labels, a feature that is not present in (S)CLS.

BioAmbients [40], is a calculus in which biological systems are modelled using a variant of the ambient calculus. In BioAmbients both membranes and elements are modelled by ambients, and activities by capabilities (enter, exit, expel, etc.). In [48], BioAmbients are extended by allowing the rates associated with rules to be context dependent. Dependency is realised by associating to a rule a function which is evaluated when applying the rule, and depends on the context of the application. The context contains the

state of the sibling ambients, that is the ambients in parallel in the innermost enclosing ambient (membrane). The property of the context used to determine the value of the function is its volume that synthesises (with a real number) the elements present in the context.

MGS [50, 51], is a domain specific language for simulation of biological processes. The state of a dynamical system is represented by a collection. The elements in the collection represent either entities (a subsystem or an atomic part of the dynamical system) or messages (signal, command, information, action, etc.) addressed to an entity. The dynamics is defined by rewrite rules specifying the collection to be substituted through a pattern language based on the neighbourhood relationship induced by the topology of the collection. It is possible to specify stochastic rewrite strategies. In [52], this feature is used to provide the description of various models of the genetic switch of the λ phage, from a very simple biochemical description of the process to an individual-based model on a Delaunay graph topology.

Finally, we would like to mention the recent framework proposed by Oury and Plotkin [24] which is based on stochastic multi-level multiset rewriting and is similar to our CWC. Their models, constructed from species and agents (representing, respectively, atoms and compartments in CWC) evolve according to a stochastic semantics associating rates to rewrite rules. As pointed out in [24], the main difference with respect to CWC is that the analysis is strongly term rewriting oriented and compartment wrappings may not be specified explicitly (an encoding based on nested agents is shown to do the work).

6.3. Hybrid Models

Stochastic formulation of chemical kinetics is mainly based on Gillespie's algorithm [15], which explicitly accounts for the individual reactive collisions among the molecules. However, it is problematic to use exact simulation methods to study systems containing a large number of molecules affected by fast reactions due to the computational cost of accounting for individual molecular collisions.

The problem of efficient simulation of systems involving reactions varying across multiple scales of time and molecular concentrations employing a mixed stochastic–deterministic method to approximate system dynamics is not new, and has been already addressed by several recent studies. Gillespie, in [20], presented the "τ–leap" method, an approximate technique for accelerating stochastic simulation, in which the occurrence of some fast reactions

can be eliminated by taking time steps that are larger than a single reaction. An improved procedure for selecting the value of τ has also been presented in [53].

Haseltine and Rawlings [30] partition the system into the subsets of *slow* and *fast* reactions, and approximate the fast reactions either deterministically or as Langevin equations. In the method of Rao and Arkin [54], some of the reactions are explicitly simulated with Gillespie's algorithm whereas others are described by random variables distributed according to the probability density functions at quasi-stationary state. The last two methods require direct intervention of the modeller to partition the system into reaction sets covering different time and concentration regimes. Similarly to our hybrid approach Salis and Kaznessis [16] proposed two parameters to define how many reactions occur within a time step and how fine grained the species must be to appear continuous-valued. They approximate fast reactions using a chemical Langevin equation. Our method, instead, simplifies the stochastic integration ignoring the fluctuations in fast reaction dynamics using ODEs.

Bortolussi and Policriti [55] provide a Stochastic Concurrent Constraint Programming (sCCP) algebra with a semantics based on hybrid automata combining discrete and continuous steps. A similar technique is developed in [56] for the Stochastic π-Calculus. In [57], a hybrid analysis technique, combining stochastic simulations with ODEs is presented in the context of PEPA precess algebra.

In [58], the HYPE process algebra, developed to model hybrid systems in which the continuous behaviour of a subsystem does not need to be understood in advance of the modelling process, is used to model the repressilator genetic regulatory network.

In [59] a hybrid technique, computing ODE with the Runge-Kutta numerical approximation method, is adapted in the Real-Time Maude rewriting logic.

The state of the art approaches, methods and tools in hybrid modelling cut down the computational cost of large stochastic exact simulations. However, they introduced new forms of complexities. In particular: (i) when different scales are taken into consideration, a deep consistency study on the system under analysis should be carried out, (ii) new parameters have to be defined to control the degree of the approximation, (iii) there is no commonly understood policy to partition the set of rules.

7. Conclusions

In this paper we have defined a hybrid simulation technique for systems described in CWC, which combines the stochastic approach with the deterministic one obtained through ODEs. The method alternates discrete transitions, computed probabilistically according to the stochastic method, and continuous transitions, computed in a deterministic way by a set of ODEs. Our technique turns out to accurately capture the dynamics of systems that exhibit stochastic effects and takes advantage, whenever the deterministic approach is applicable, of the efficiency of the ODEs integration method.

The running example used to make comparisons between the different quantitative simulation methods, and the HIV-1 transactivation mechanism are challenging tests for our hybrid methodology. On the one hand, the modified Lotka-Volterra systems allows us to take a full advantage of the computational gain possible with the hybrid method. On the other hand, the simulation of the HIV-1 transactivation mechanism follows a simulation which is *almost* purely stochastic: only a few rules pass the threshold condition, thus the computational gain of the deterministic approach is, in this particular case, very limited (though still sensible).

Compartment labels introduced in this paper are a novelty with respect to the original CWC calculus presented in [14]. As we have seen, these labels are necessary when building a system of ODEs for a compartment of type ℓ. In future work, we plan to exploit these labels as an intrinsical information about the properties of a compartment. For example, assuming that compartments of the same type have approximatively the same volume, we might use the compartment type to define a set of biochemical rules whose kinetics incorporate the information about the volume of the compartment on which the rule could be applied. Suppose, in practice, to analyse a system in which two different kinds of cells may interact. Let us call ℓ_1 and ℓ_2 the compartment types of the two kinds of cells. Suppose, then, that particles a and b are free to float between these cells and the top level interspace hosting all the cells. Finally, particles a and b may interact by complexation and produce the particle c. If it holds that the top level interspace on which the different cells float has around 100x the volume of a cell of type ℓ_1 and if a cell of type ℓ_1 has around 3x the volume of a cell of type ℓ_2, we can express the different speeds of the $a - b$ complexation in the different compartments (according to their volumes) with the three following rules:

$$\top : a\ b \stackrel{k}{\longmapsto} c, \qquad \ell_1 : a\ b \stackrel{k \cdot 100}{\longmapsto} c, \qquad \ell_2 : a\ b \stackrel{k \cdot 300}{\longmapsto} c.$$

Actually, it is crucial to consider in detail the volumes of the involved compartments and to consider adequate kinetics for the biochemical rules used to simulate the system behaviour. We notice, in particular, that the approach based on ODEs directly translates chemical reactions into mathematical equations and computes the concentrations over time of the involved species (usually the *molar concentration*, which denotes the number of moles of a given substance per litre). Models based on the stochastic approach, instead, simulate the activity of each single individual involved in the evolution of the system. Such a delicate difference between the two methods should be carefully taken into account when developing the set of rules to be simulated with the hybrid approach.

In Appendix A we provide a preliminary analysis about a possible method to automatically and dynamically partition the set of reactions between deterministic and stochastic calculations. Such a technique could also be used to estimate the approximation error of our method. As a future work we plan to investigate more deeply these kinds of methods.

Acknowledgements

We gratefully acknowledge the helpful comments and suggestions received from the anonymous reviewers of MeCBIC 2010 and TCS. We thank Marco Aldinucci for the useful discussions on the simulator implementation issues and for providing us computing resources (Intel workstations with 2 quad-core Xeon). The authors also wish to thank Sergio Rabellino and the ICT staff of the Computer Science Department of the University of Torino for providing technical support and assistance in running the simulations. Finally, we thank Prof. Nello Balossino (University of Torino) who made us available the computing resources of the Laboratory of Signals and Images "G. Tamburelli".

References

[1] M. Elowitz, A. Levine, E. Siggia, P. Swain, Stochastic gene expression in a single cell, Science 297 (5584).

[2] A. Regev, E. Shapiro, Cells as computation, Nature 419 (2002) 343.

[3] R. Alur, C. Belta, F. Ivancic, Hybrid Modeling and Simulation of Biomolecular Networks, in: HSCC, Vol. 2034 of LNCS, Springer, 2001, pp. 19–32.

[4] H. Matsuno, A. Doi, M. Nagasaki, S. Miyano, Hybrid Petri net representation of gene regulatory network, in: Prooceedings of Pacific Symposium on Biocomputing, World Scientific Press, 2000, pp. 341–352.

[5] V. Danos, C. Laneve, Formal molecular biology, Theor. Comput. Sci. 325 (1) (2004) 69–110.

[6] G. Păun, Membrane computing. An introduction, Springer, 2002.

[7] R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, A. Troina, A Calculus of Looping Sequences for Modelling Microbiological Systems, Fundam. Inform. 72 (1-3) (2006) 21–35.

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

[9] L. Cardelli, Brane Calculi, in: CMSB, Vol. 3082 of LNCS, Springer, 2004, pp. 257–278.

[10] P. Degano, D. Prandi, C. Priami, P. Quaglia, Beta-binders for Biological Quantitative Experiments, Electr. Notes Theor. Comput. Sci. 164 (3) (2006) 101–117.

[11] J. Krivine, R. Milner, A. Troina, Stochastic Bigraphs, Electron. Notes Theor. Comput. Sci. 218 (2008) 73–96. doi:http://dx.doi.org/10.1016/j.entcs.2008.10.006.

[12] R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, P. Tiberi, A. Troina, Stochastic Calculus of Looping Sequences for the Modelling and Simulation of Cellular Pathways, Transactions on Computational Systems Biology IX (2008) 86–113.

[13] L. Dematté, C. Priami, A. Romanel, The Beta Workbench: a computational tool to study the dynamics of biological systems, Briefings in Bioinformatics 9 (5) (2008) 437–449.

[14] M. Coppo, F. Damiani, M. Drocco, E. Grassi, A. Troina, Stochastic Calculus of Wrapped Compartments, in: 8th Workshop on Quantitative Aspects of Programming Languages (QAPL'10), Vol. 28, EPTCS, 2010, pp. 82–98.

[15] D. Gillespie, Exact stochastic simulation of coupled chemical reactions, J. Phys. Chem. 81 (1977) 2340–2361.

[16] H. Salis, Y. Kaznessis, Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions, J Chem Phys 122(5):54103.

[17] M. Griffith, T. Courtney, J. Peccoud, W. Sanders, Dynamic partitioning for hybrid simulation of the bistable HIV-1 transactivation network, Bioinformatics 22 (2006) 2782–2789.

[18] A. Crudu, A. Debussche, O. Radulescu, Hybrid stochastic simplifications for multiscale gene networks, BMC Syst Biol 3:89.

[19] D. T. Gillespie, The chemical Langevin equation, Journal of Chemical Physics 113 (1) (2000) 297.

[20] D. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems, The Journal of Chemical Physics 115 (2001) 1716.

[21] Y. Cao, D. T. Gillespie, L. Petzold, Efficient Stepsize Selection for the Tau-Leaping Method, Journal of Chemical Physics 58.

[22] M. Aldinucci, M. Coppo, F. Damiani, M. Drocco, E. Giovannetti, E. Grassi, E. Sciacca, S. Spinella, A. Troina, CWC Simulator, Dipartimento di Informatica, Università di Torino, http://sourceforge.net/projects/cwcsimulator/ (2010).

[23] M. Coppo, F. Damiani, M. Drocco, E. Grassi, E. Sciacca, S. Spinella, A. Troina, Hybrid Calculus of Wrapped Compartments, in: 4th International Meeting on Membrane Computing and Biologically Inspired Process Calculi (MeCBIC'10), Vol. 40, EPTCS, 2010, pp. 102–120.

[24] N. Oury, G. Plotkin, Multi-Level Modelling via Stochastic Multi-Level Multiset Rewriting, draft submitted to MSCS (2011).

[25] S. Singer, G. Nicolson, The fluid mosaic model of the structure of cell membranes, Science 175 (1972) 720–731.

[26] E. Parzen, Stochastic Processes, Holden-Day, 1962.

[27] M. Dezani-Ciancaglini, P. Giannini, A. Troina, A Type System for a Stochastic CLS, in: MeCBIC'09, Vol. 11, EPTCS, 2009, pp. 91–105.

[28] F. Horn, R. Jackson, General mass action kinetics, Archive for Rational Mechanics and Analysis 47 (1972) 81–116, 10.1007/BF00251225.
URL http://dx.doi.org/10.1007/BF00251225

[29] E. Haseltine, J. Rawlings, On the origins of approximations for stochastic chemical kinetics, The Journal of chemical physics 123 (2005) 164115.

[30] E. Haseltine, J. Rawlings, Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics, The Journal of Chemical Physics 117 (2002) 6959.

[31] K. Vasudeva, U. Bhalla, Adaptive stochastic-deterministic chemical kinetic simulations, Bioinformatics 20 (1) (2004) 78.

[32] R. Srivastava, L. You, J. Summers, J. Yin, Stochastic vs. deterministic modeling of intracellular viral kinetics, Journal of Theoretical Biology 218 (3) (2002) 309–321.

[33] V. Dahl, L. Josefsson, S. Palmer, HIV reservoirs, latency, and reactivation: Prospects for eradication, Antiviral Research 85 (2010) 286–294.

[34] L. Weinberger, J. Burnett, J. Toettcher, A. Arkin, D. Schaffer, Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity, Cell 122 (2005) 169–182.

[35] N. Nitin, L. LaConte, W. Rhee, G. Bao, Tat peptide is capable of importing large nanoparticles across nuclear membrane in digitonin permeabilized cells, Ann Biomed Eng 37 (2009) 2018–2027.

[36] J. Hartigan, M. Wong, A k-means clustering algorithm, JR Stat. Soc., Ser. C 28 (1979) 100–108.

[37] M. Curti, P. Degano, C. Priami, C. T. Baldari, Modelling biochemical pathways through enhanced pi-calculus, Theor. Comput. Sci. 325 (1) (2004) 111–140.

[38] A. Regev, W. Silverman, E. Y. Shapiro, Representation and Simulation of Biochemical Processes Using the pi-Calculus Process Algebra, in: Pacific Symposium on Biocomputing, 2001, pp. 459–470.

[39] N. Chabrier-Rivier, M. Chiaverini, V. Danos, F. Fages, V. Schächter, Modeling and querying biomolecular interaction networks, Theor. Comput. Sci. 325 (1) (2004) 25–44.

[40] A. Regev, E. M. Panina, W. Silverman, L. Cardelli, E. Y. Shapiro, BioAmbients: an abstraction for biological compartments, Theor. Comput. Sci. 325 (1) (2004) 141–167.

[41] R. Brijder, M. Cavaliere, A. Riscos-Núñez, G. Rozenberg, D. Sburlan, Membrane systems with proteins embedded in membranes, Theor. Comput. Sci. 404 (1-2) (2008) 26–39.

[42] M. Cavaliere, S. Sedwards, Decision problems in membrane systems with peripheral proteins, transport and evolution, Theor. Comput. Sci. 404 (1-2) (2008) 40–51.

[43] R. Barbuti, A. Maggiolo-Schettini, P. Milazzo, Extending the Calculus of Looping Sequences to Model Protein Interaction at the Domain Level, in: ISBRA, Vol. 4463 of LNCS, Springer, 2007, pp. 638–649.

[44] B. Aman, M. Dezani-Ciancaglini, A. Troina, Type Disciplines for Analysing Biologically Relevant Properties, Electr. Notes Theor. Comput. Sci. 227 (2009) 97–111.

[45] V. Danos, J. Feret, W. Fontana, J. Krivine, Scalable Simulation of Cellular Signaling Networks, in: APLAS, Vol. 4807 of LNCS, Springer, 2007, pp. 139–157.

[46] D. Pescini, D. Besozzi, G. Mauri, C. Zandron, Dynamical probabilistic P systems, Int. J. Found. Comput. Sci. 17 (1) (2006) 183–204.

[47] A. Spicher, O. Michel, M. Cieslak, J.-L. Giavitto, P. Prusinkiewicz, Stochastic P systems and the simulation of biochemical processes with dynamic compartments, Biosystems 91 (3) (2008) 458–472.

[48] L. Bortolussi, M. G. Vigliotti, CoBiC: Context-dependent Bioambient Calculus, Electr. Notes Theor. Comput. Sci. 253 (3) (2009) 187–201.

[49] F. Ciocchetta, J. Hillston, Bio-PEPA: An Extension of the Process Algebra PEPA for Biochemical Networks, ENTCS 194 (3) (2008) 103–117.

[50] J.-L. Giavitto, Invited Talk: Topological Collections, Transformations and Their Application to the Modeling and the Simulation of Dynamical Systems, in: RTA, Vol. 2706 of Lecture Notes in Computer Science, Springer, 2003, pp. 208–233.

[51] O. Michel, J.-L. Giavitto, J. Cohen, A. Spicher, MGS home page, `http://mgs.spatial-computing.org/`.

[52] O. Michel, A. Spicher, J.-L. Giavitto, Rule-based programming for integrative biological modeling, Natural Computing 8 (4) (2009) 865–889.

[53] Y. Cao, D. Gillespie, L. Petzold, Efficient step size selection for the tau-leaping simulation method, The Journal of chemical physics 124 (2006) 044109.

[54] C. Rao, A. Arkin, Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm, The Journal of Chemical Physics 118 (2003) 4999.

[55] L. Bortolussi, A. Policriti, Hybrid dynamics of stochastic programs, Theor. Comput. Sci. 411 (20) (2010) 2052–2077.

[56] L. Bortolussi, A. Policriti, Hybrid Dynamics of Stochastic pi-Calculus, Mathematics in Computer Science 2 (3) (2009) 465–491.

[57] A. Stefanek, R. Hayden, J. T. Bradley, Hybrid analysis of large scale PEPA models, in: PASTA'10, 2010.

[58] V. Galpin, J. Hillston, L. Bortolussi, HYPE Applied to the Modelling of Hybrid Biological Systems, Electr. Notes Theor. Comput. Sci. 218 (2008) 33–51.

[59] M. Fadlisyah, P. Ölveczky, E. Abrahám, Formal Modeling and Analysis of Hybrid Systems in Rewriting Logic using Higher-Order Numerical Methods and Discrete-Event Detection, in: CSSE'11, IEEE Computer Society, 2011.

[60] D. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, Journal of computational physics 22 (4) (1976) 403–434.

Appendix A. Dynamical Partition of Deterministic and Stochastic Rules

A basic point of our hybrid approach is the criterion to determine, at each computation step, the reductions to compute in the stochastic or in the deterministic way. In this paper we have chosen simply to put a threshold on the number of possible reagents (ψ) and on the propensity of a reaction (ϕ). The thresholds are set by our accuracy requirements following a consistency check on quantities and probabilities [31]. We now present a criterion for the choice of the threshold ϕ of our algorithm to divide the reactions in fast and slow by satisfying the conditions explained in Section 4.

Assume a single reaction r_i:

$$\ell : \bar{a}_i \xmapsto{k_i} \bar{b}_i \tag{A.1}$$

where, for simplicity, we assume that $\bar{a}_i$ does not contain duplicated atoms. The probability that reaction r_i does not occur in a time τ is given by

$$e^{-\pi_{r_i}\cdot\tau} \text{ where } \pi_{r_i} = k_i \cdot [a_{i_1}]_\iota \cdot \ldots \cdot [a_{i_{r_i}}]_\iota.$$

In order to bind the propensity of a reaction to the probability of that reaction to not occur in a given time τ, and to correlate this quantities, we apply the power series approximation:

$$e^{-x} = \sum_{n=0}^{\infty} \frac{(-x)^n}{n!} \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \ldots.$$

When $x \ll 1$, for instance when $x < 0.2$, the approximation error goes as $o(x^2)$ and is under 5%, and can be truncated in the second order term (namely $e^{-x} \approx 1 - x$). Thus, given $x = \pi_{r_i} \cdot \tau$ we get the following:

$$e^{-\pi_{r_i}\cdot\tau} \approx 1 - \pi_{r_i} \cdot \tau \quad \text{if} \quad \pi_{r_i} \cdot \tau < 0.2 \tag{A.2}$$

On the contrary, the probability that reaction r_i occurs within the time τ is given by:

$$1 - e^{-\pi_{r_i}\cdot\tau} \approx \pi_{r_i} \cdot \tau$$

The above discussion suggests a possible schema to estimate at each step of our simulation algorithm the threshold ϕ. According to [60], given a set

of reactions $\{r_i\}_i$, the probability that no reaction at all occurs in a time $[t, t+\tau]$ is:

$$p(\tau) = e^{-\tau \cdot \sum_{r_i} \pi_{r_i}} \tag{A.3}$$

Let us consider the expectation time τ_m for the probability density A.3:

$$\tau_m = E[\tau] = \sum_{r_i} \pi_{r_i} \int_0^\infty \tau \cdot e^{-\tau \cdot \sum_{r_i} \pi_{r_i}} d\tau = \frac{1}{\sum_{r_i} \pi_{r_i}} \tag{A.4}$$

Introducing τ_m into the approximation A.2, the probability that a certain reaction r_i occurs in time τ_m is approximated as:

$$\pi_{r_i} \cdot \tau_m \tag{A.5}$$

Therefore we could consider as *slow* a reaction whose probability to take place before the expected time of the first reaction in the system is low, say lower than 0.2.

On the contrary we might consider as fast a reaction r_i, and compute it deterministically, when $\pi_{r_i} \cdot \tau_m > 0.2$, which gives

$$\pi_{r_i} \geq \frac{0.2}{\tau_m} = 0.2 \cdot \sum_{r_i} \pi_{r_i}$$

thus, for any $\phi \geq 0.2 \cdot \sum_{r_i} \pi_{r_i}$ we obtain the required accuracy level.

This result could be exploited to improve the consistency of our hybrid algorithm by providing a mean to compute at runtime the value for the threshold ϕ, this increases the robustness of the computational framework. Also note that, in this case, the threshold ϕ could be parametrized by the desired level of accuracy. The thresholds ϕ used in the examples in Sections 4.1 and 5 have been chosen according to the above method.

The reason for considering the threshold ψ on the number of possible reagents is explained by situations, such as the one described in the following example. In this case a reaction of kind A.1 contains at least one substrate j with a small number $[a_{i_j}]_\iota$ of molecules but the other substrates have a number of molecules large enough to make $\pi_{r_i} > \phi$.

Example B. *Let us consider the reaction:*

$$(r) \qquad A\ B \overset{10}{\mapsto} C$$

and assume that $[A] = 1000$, $[B] = 1$. *We therefore have a reaction propensity relative to a time step* τ *:*

$$\pi_r \cdot \tau = 10 \cdot 1000 \cdot 1 \cdot \tau = \phi \cdot \tau = 10000 \cdot \tau$$

We can exploit Euler first-order criterion (the error term is the square of the probability of an individual molecule reacting). In order to achieve an error approximation of 5% *we have:*

$$\left(\frac{\pi_r}{[B]}\right)^2 \leq 0.05$$

that is

$$\pi_r \cdot \tau \leq \sqrt{0.05} \quad \Rightarrow \quad \tau \leq 2.24 * 10^{-5}$$

Therefore, deterministic computation is clearly inappropriate for a selected time step $\tau \sim 2.24 * 10^{-5}$ *(even if* $\phi = 10000$ *is large), because stochastic phenomena involving species* B *would be neglected. Furthermore, since small numbers of molecules are involved, we are very far from the thermodynamic limit. This situation shows that we must set a cut-off number of molecules below which all calculations involving that molecule should be done stochastically. We should specify a cut-off* ψ *on molecular number, which is sufficiently large so that the change in standard deviation is tolerable.*

An additional refinement to the use of sharp thresholds ϕ and ψ is to consider two intervals $[\phi_{\text{low}}, \phi_{\text{high}}]$ e $[\psi_{\text{low}}, \psi_{\text{high}}]$ in order to use additional information on the rate and population size, such as their gradient, in uncertain situations of oscillatory behaviour.