Bio-PEPA: a framework for the modelling and analysis of biological systems

Federica Ciocchetta [a,*] and Jane Hillston [a,b]

[a] *Laboratory for Foundations of Computer Science, The University of Edinburgh, Edinburgh EH9 3JZ, Scotland*

[b] *Centre for Systems Biology at Edinburgh (CSBE)* [1]

Abstract

In this work we present Bio-PEPA, a process algebra for the modelling and the analysis of biochemical networks. It is a modification of PEPA, originally defined for the performance analysis of computer systems, in order to handle some features of biological models, such as stoichiometry and the use of general kinetic laws. The domain of application is the one of biochemical networks. Bio-PEPA may be seen as an intermediate, formal, compositional representation of biological systems, on which different kinds of analysis can be carried out. Bio-PEPA is enriched with some notions of equivalence. Specifically, the isomorphism and strong bisimulation for PEPA have been considered and extended to our language. Finally, we show the translation of a biological model into the new language and we report some analysis results.

Key words: Process Algebras, Biochemical Networks, Modelling, Analysis

1 Introduction

In recent years there has been increasing interest in the application of process algebras in the modelling and analysis of biological systems [44,22,26,43,13,40,8]. Process algebras have some interesting properties that make them particularly useful in this context. First of all, biological systems can be abstracted by concurrent

* Corresponding author.
Email addresses: fciocche@inf.ed.ac.uk (Federica Ciocchetta), jeh@inf.ed.ac.uk (Jane Hillston).

[1] The Centre for Systems Biology at Edinburgh is a Centre for Integrative Systems Biology (CISB) funded by the BBSRC and EPSRC in 2006

systems in a straightforward way: species may be seen as processes that can interact with each other and reactions may be modelled using actions. Secondly, process algebras give a formal representation of the system avoiding ambiguity. Thirdly, they offer *compositionality*, i.e. the possibility of defining the whole system starting from the definition of its subcomponents. Finally, different kinds of analysis can be performed on a process algebra model. These analyses provide conceptual tools which are complementary to established techniques: it is possible to detect and correct potential inaccuracies, to validate the model and to predict its possible behaviours.

The process algebra PEPA, originally defined for the performance analysis of computer systems, has been recently applied in the context of signalling pathways [8,9]. Two approaches have been proposed: one based on reagents (the so-called *reagent-centric view*) and another based on pathways (*pathway-centric view*). In both cases the species concentrations are discretised into levels, each level abstracting an interval of concentration values. In the reagent-centric view the PEPA sequential components represent various concentration levels of the species. The abstraction is "*processes as species*". This is different from the abstractions generally adopted in the application of other process algebras in systems biology, such as "*processes as molecules*" or "*processes as interactions*". The former is the most widely-used abstraction in this context and it has been chosen in a lot of case studies involving the π-calculus and Beta-binders [44,43,21,22]. The latter has been proposed in [5] for the modelling of biological systems by means of the *stochastic Concurrent Constraint Programming (sCCP)*. In the pathway-centric approach of PEPA we have a more abstract view: the processes represent sub-pathways. Here multiple copies of components represent levels of concentration. The two views of PEPA have been shown to be equivalent [8].

Even though PEPA has proved useful in studying signalling pathways, it does not allow us to represent all the features of biological networks. The main difficulties are the definition of *stoichiometric coefficients* (i.e. the coefficients used to show the quantitative relationships of the reactants and products in a biochemical reaction) and the representation of *kinetic laws*. Indeed, stoichiometry is not represented explicitly and the reactions are assumed to be elementary. The problem of extending to the domain of kinetic laws beyond basic mass-action (hereafter called *general kinetic laws*) is particularly relevant, as these kinds of reactions are frequently found in the literature as abstractions of complex situations whose details are unknown. Reducing all reactions to the elementary steps is complex and often impractical. This problem impacts also on other process algebras. Generally they rely on Gillespie's stochastic simulation which considers only elementary reactions. Some recent works have extended the approach of Gillespie to deal with complex reactions [1,11] but these extensions are yet to be reflected in the work using process algebras. Previous work concerning the use of general kinetic laws in process algebras and formal methods was presented in [5,14]. These are discussed in Section 3.1.

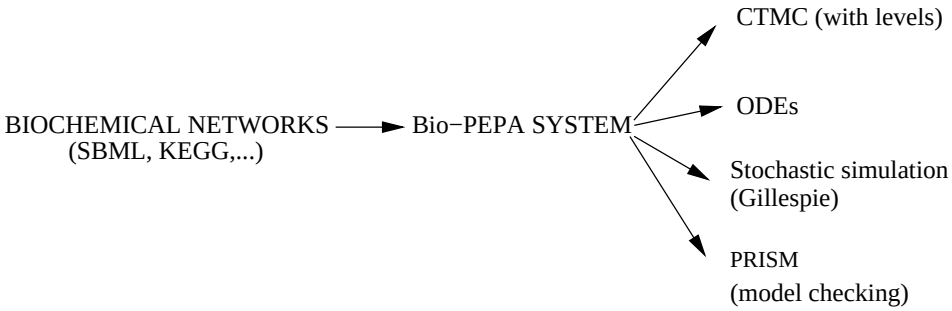

Fig. 1. Schema of the Bio-PEPA framework

In this paper we present Bio-PEPA, a language for the modelling and the analysis of biochemical networks. A preliminary version of the language has been proposed in [16]. Here we describe the final version of the language, we introduce new definitions and more details about our approach.

A major feature of Bio-PEPA is the possibility to represent explicitly some features of biochemical models, such as stoichiometry and the role of the species in a given reaction. Furthermore functional rates are introduced to express general kinetic laws. Each action type represents a reaction and is associated with a functional rate. Bio-PEPA is equipped with an operational semantics and a stochastic labelled system based on discrete levels of concentration. In this respect our language follows a similar approach to the reagent-centric view of PEPA. The representation in terms of discrete levels of concentration is also reflected in the definition of the continuous time Markov chains (CTMC) derived from the system. Hereafter we call this Markov chain *CTMC with levels*. We enrich Bio-PEPA with some notions of *equivalence*. We extend the definition of isomorphism and *strong bisimulation* proposed for PEPA in [36] to Bio-PEPA.

The idea underlying our work is represented schematically in the diagram in Fig. 1. The context of application is biochemical networks. Broadly speaking, biochemical networks consist of some biochemical species, which interact with each other through reactions. The reaction dynamics are described in terms of kinetic laws. The biochemical networks can be obtained from databases such as *KEGG* [38,37] and *BioModels Database* [42]. From the biological model, we develop the Bio-PEPA specification of the system. This is an *intermediate, formal, compositional* representation of the biological model. At this point we can apply different kinds of analysis, including stochastic simulation [32], analysis based on ordinary differential equations (ODEs), numerical solution of CTMC and stochastic model checking using PRISM [45,35]. The choice of one or more methods depends on the context of application [47].

It is worth noting that the use of various kinds of analysis can help in understanding the system. We can use two or more analyses to investigate different but related aspects of the model. Furthermore, when they overlap, the results obtained can provide a further confirmation of the behaviour of the system. These aspects

were considered in [9,17]. The work in [9] concerns the comparison of the results obtained using implicit numerical differentiation formulae to those obtained using approximate stochastic simulation in the case of a signalling pathway. This reveals the flaw in the use of the differentiation procedure producing misleading results. In [17] we presented an approach that uses stochastic simulation and the PRISM probabilistic model checker in tandem in order to investigate the properties of biological systems.

There exist some relations between the different kinds of analysis. It is well-known that the ODEs solution tends to the results of stochastic simulations when the number of elements is relatively high. Similarly, it is shown in [31] that the numerical solution of the CTMC with levels (derived from the PEPA pathway-centric view) tends to the solution of the ODEs when the number of levels increases. An analogous result has been recently proved for Bio-PEPA [18]. We showed that the set of ODEs derived from Bio-PEPA is able to capture the limiting behaviour of the CTMC with levels obtained from the same system. Furthermore we proposed an empirical methodology to find the granularity of the Bio-PEPA system for which the ODE model and the CTMC with levels are in a good agreement. The proposed definition is based on a notion of distance between the two models: the granularity of the system, expressed in terms of the step size of the concentration levels, is chosen in order to minimise this distance. In this way we are able to define an ODE model and a CTMC model that represent the same biological system and we use different analysis techniques from the two representations to investigate various properties of it.

The paper is structured as follows. In the next section a description of biochemical networks is reported. Section 3 describes PEPA and reports the application of PEPA to the modelling of some signalling pathways. Furthermore, some related works concerning the application of process algebras in systems biology are discussed. After that, in Section 4, we define Bio-PEPA. The semantics of Bio-PEPA in terms of a labelled stochastic transition system is presented in Section 5. Section 6 reports some auxiliary definitions, used in the following Section 7, where some equivalences for Bio-PEPA are presented. In Section 8 we discuss the main kinds of analysis that can be used from a Bio-PEPA model. The translation of a biological model into Bio-PEPA and its subsequent analysis is described in Section 9. Finally, Section 10 reports some final observations and future investigations.

2 Biochemical networks

We focus on biochemical networks, such as those collected in the Biomodels Database [42] and KEGG [38]. A biochemical system $\mathcal{M}$ is composed of:

(1) a set of *compartments* $\mathbf{C}$. These represent the locations of the various species;

(2) *a set of chemical species* **S**. These species may be genes, proteins, etc. For each species an initial concentration is given;
(3) *a set of (irreversible) reactions* **R**. The general form of an irreversible reaction j is given by:

$$\kappa_{1j}A_{1j} + \kappa_{2j}A_{2j} + ... + \kappa_{n_jj}A_{n_jj} \xrightarrow{E_{1j},E_{2j},...I_{1j},I_{2j},...;f_j} \kappa'_{1j}B_{1j} + \kappa'_{2j}B_{2j} + ... + \kappa'_{m_jj}B_{m_jj}$$

where A_{hj}, $h = 1, ..., n_j$, are the reactants, B_{lj}, $l = 1, ..., m_j$, are the products, E_{vj} are the enzymes and I_{uj}, the inhibitors. All these species belong to the set **S**. Enzymes and inhibitors are represented differently from the reactants and products. Their role is to enhance or inhibit the reaction, respectively. We call species that are involved in a reaction without changing their concentration (i.e. enzymes/activators and inhibitors) *modifiers*. The parameters κ_{hj} and κ'_{lj} are the stoichiometry coefficients. These express the degree to which species participate in a reaction. The dynamics is described by a kinetic law f_j. Reversible reactions can be regarded as a pair of forward and inverse reactions.

The best known kinetic law is *mass-action*: the rate of the reaction is proportional to the product of the reactants' concentrations. In published models it is common to find *general kinetic laws*, which describe approximations of sequences of reactions [46]. They are useful when it is difficult to derive certain information from the experiments, e.g. the reaction rates of elementary steps, or when there are different time-scales for the reactions. General kinetic laws are valid under some conditions, such as the *quasi-steady-state assumption (QSSA)*. This describes the situation where one or more reaction steps may be considered faster than the others and so the quantity of intermediate elements can be considered to be constant.

3 PEPA and biological systems

PEPA was originally defined for the performance modelling of systems with concurrent behaviour [36]. Systems are represented as the composition of components which undertake actions. In PEPA each action is assumed to have a duration, which is represented by a random variable with a negative exponential distribution. PEPA has a set of combinators that allows the system description to be built up as the concurrent interaction of simple sequential components.

We informally introduce the syntax of the language below. For more details see [36].

Prefix The basic term is the *prefix combinator* $(\alpha, r).P$. It denotes a component which has action of type α and an exponentially distributed duration with parameter r (mean duration $1/r$), and it subsequently behaves as P.
Choice The component $P + Q$ represents a system which may behave either as P

or as Q. The activities of both P and Q are enabled. The first activity to complete distinguishes one of them and the other is discarded.

Constant Constants are components whose meaning is given by a defining equation $C \stackrel{def}{=} P$. They allow us to assign names to patterns of behaviour associated with components.

Hiding In $P/\mathcal{H}$ the set $\mathcal{H}$ identifies those activities which can be considered internal or private to the component P.

Cooperation The term $P \underset{\mathcal{L}}{\bowtie} Q$ denotes cooperation between P and Q over the cooperation set $\mathcal{L}$, that determines those activities on which the cooperands are forced to synchronise. PEPA supports *multiway synchronisation* between components: the result of synchronising on an activity α is thus another α, available for further synchronisation. For action types not in $\mathcal{L}$, the components proceed independently and concurrently with their enabled activities. In the context of performance evaluation the rate for the synchronised activities is the minimum of the rates of the synchronising activities.

PEPA has a structured operational semantics which generates a labelled transition system and from this a continuous time Markov chain (CTMC) is derived.

Recently, PEPA has been applied to the modelling and analysis of signalling pathways. A first study concerns the influence of the Raf Kinase Inhibitor Protein (RKIP) on the Extracellular signal Regulated Kinase (ERK) [8], whereas in [9] the PEPA system for Schoeberl's model [29] involving the MAP kinase and EFG receptors is reported. In [8] two modelling styles have been proposed, one based on the *reagent-centric view* and the other on the *pathway-centric view*. The former focuses on the variation in the reagent concentrations: the concentrations are discretised in levels, each level representing an interval of concentration values. The level l can assume values between 0 and N_{max} (maximum level). The pathway-centric style provides a more abstract view and focuses on the subpathways. The two representations were shown to be equivalent [8]. In addition to the standard analysis offered by process algebras, in [7] a mapping from reagent-centric PEPA models to a system of ordinary differential equations (ODEs), has been proposed.

From these works PEPA has been shown to be appropriate for the modelling of biological systems: it offers a high level of abstraction and focuses on compositionality and on the interactions. By using PEPA as a modelling language it is possible to apply different kinds of analysis, not only stochastic simulation, but also differential equations and study by means of model checking. However, not all the features of biochemical networks can be expressed using the present version of PEPA: the various kinetic laws are not considered and stoichiometry is added by hand in the conversion of PEPA into ODEs. With a few exceptions (e.g. [5]) and a few cases, these features cannot be represented in other process algebras either.

Other process algebras have been considered in the context of biological systems. Initial work focused upon the π-*calculus* and its biochemical stochastic extension [44]. Several case studies have been considered, e.g. [22,40]. The translation of biochemical models into this language is based on the abstraction "*processes as single molecules*": molecules are represented by processes and the biological interactions are abstracted by communications between processes.

Beta-binders [43] is an extension of the π-calculus inspired by biological phenomena. This calculus is based on the concept of *bio-process*, a box with some sites (*beta-binders*) to express the interaction capabilities, in which π-like processes (*pi-processes*) are encapsulated. Beta-binders enrich the standard π-calculus with some constructs that allow the modeller to represent biological features, such as the join between two bio-processes, the split of one bio-process into two, the change of the bio-process interface. In both π-calculus and Beta-binders it is not possible to represent all the features that are present in the biochemical networks proposed in this paper. The kinetic law is assumed to be mass-action and reactions can have at most two reactants. In order to represent multiple-reactant multiple-product reactions transactions are considered [19,20]. Finally, in both cases the analysis of the model is based on stochastic simulation using Gillespie's algorithm [32].

Another language for the modelling of biological systems is the κ-calculus [23,24], based on the description of protein interactions. Processes describe proteins and their compounds, a set of processes model solutions and protein behaviour is given by a set of rewriting rules, driven by suitable side-conditions. The two main rules concern activation and complexation. The calculus is supported by a graphical notation in terms of boxes. A stochastic simulator for κ-calculus is described in [25]. A few applications are reported, as in [24].

Previous works concerning the use of general kinetic laws and stoichiometry in process algebras and formal methods have been proposed in [5,14]. The authors of [5] present a stochastic extension of *Concurrent Constraint Programming* (CCP) and show how to apply it in the case of biological systems. Here each species is represented by a variable and the reactions are expressed by constraints on these variables. The domain of application is extended to any kind of reactions and the rate can be expressed by a generic function. *BIOCHAM* [14] is a programming environment for modelling biochemical systems, making simulations and querying the model in temporal logic. In its current version BIOCHAM is based on a rule-based language for modelling biochemical systems, in which species are expressed by objects and reactions by reaction rules. The rates are expressed by using some functions, whose definition is similar to the one proposed in our work. This language permits the evaluation of temporal logic queries using the *NuSMV* model checker [41]. Functional rates has been recently considered in *Blenx* [27,28], a lan-

guage inspired by Beta-binders [43] for the modelling and analysis of biological systems.

4 Bio-PEPA

The aim of this work is to define a new process algebra in order to model some of the features of biochemical networks that are not possible to represent in PEPA. We will show that the new language is able to represent all the reactions in a straightforward way and it deals with stoichiometry and general kinetic laws.

We adopt a high level of abstraction similar to the one proposed in formalisms such as *SBML* [3]. Furthermore we have made the following assumptions:

(1) compartments are *static*, i.e. compartments are not actively involved in the reactions —they are simply containers.
(2) Reactions are *irreversible reactions*.

The first assumption reflects the current information about locations that can be found in the literature and in the databases of biochemical networks [42]. The current information about compartments is poor and most models are based on some limitations. The assumption of static compartments for Bio-PEPA allows us to keep the language simple and at the same time to represent most of the features of the biochemical networks. For instance, the transport of a species from one compartment to another is modelled by introducing two distinct components representing the species. The translocation is abstracted by a transformation of one species into the other. Compartments must be considered in the definition of a Bio-PEPA system because in the analysis it can be necessary to have the size of the compartments (for instance for Gillespie's algorithm [32]).

Note that the second assumption is not restrictive as a reversible reaction can be split into two irreversible reactions, representing the forward and the inverse direction.

4.1 Discrete concentrations and granularity

The definition of the transition system for Bio-PEPA and the CTMC derived from it is based on the abstraction of discrete levels of concentration within a species: each component represents a species and it is parametric in terms of concentration levels. Some advantages of this view are:

- it deals with incomplete information in the exact number of elements;

- the focus is on the concentration levels not on the number of elements: this leads to a reduction of the state space as there are less states for each component.

This view was originally defined in [8] for PEPA and then used in [10]. In both the cases the authors focused on the case of reactions with mass-action kinetics and stoichiometry equal to one for all the reactants and products. Furthermore they considered the same step size H and the same maximum level N for all the species. In the following we adapt this approach to general kinetic laws, stoichiometry greater than one and different numbers of levels for the species. The granularity is defined in terms of the step size H of the concentration intervals. We define the same step size H for all the species. This is motivated by the fact that, following the *law of conservation of mass*, there must be a "balance" between the concentrations consumed (reactants) and the ones created (products). There are few exceptions to this case. For instance, when a species can be only a modifier in the model we can assigned to it a different step size, as its concentration does not vary. In the case the stoichiometry is greater than one we need to consider concentration quantities proportional to stoichiometric coefficients. Given a species i, we can assume that it has a maximum finite concentration M_i. This is to ensure a finite state space and therefore to make analysis conducted by numerical solution feasible. Each species can assume the discrete concentration levels from 0 (null concentration) to N_i (maximum concentration). We have the following relations:

- The number of levels for the species i is given by $N_i + 1$ where $N_i = \lceil M_i/H \rceil$ (i.e. the integer value greater than or equal to M_i/H).
- $l_i = \lceil x_i/H \rceil$, where l_i is the concentration level and x_i is the concentration for the species i. When initial values are considered we have $l_{i,0} = \lceil x_{i,0}/H \rceil$.

4.2 The syntax

The syntax is designed in order to collect the biological information we need:

$$S ::= (\alpha, \kappa)\ \mathsf{op}\ S \mid S + S \mid C \qquad P ::= P \underset{\mathcal{L}}{\bowtie} P \mid S(x)$$

where $\mathsf{op} = \downarrow \mid \uparrow \mid \oplus \mid \ominus \mid \odot$.

The component S is called *sequential component* (or *species component*) and represents a species. The component P, called a *model component*, describes the system and the interactions among components. The element C is the constant as in PEPA. We assume a countable set of model components C and a countable set of action types $\mathcal{A}$. The element x is a positive real-valued parameter, usually interpreted as a concentration. We consider concentrations in the specification of the system as from them we can derive both the number of molecules and the number of levels in a straightforward way. Furthermore this information is generally given in the biochemical networks and from experiments.

The prefix term in PEPA is replaced by a new one, (α, κ) op S, containing information about the role of the species in the reaction associated with α:

- (α, κ) is the prefix, where $\alpha \in \mathcal{A}$ is the *action type* and κ is the *stoichiometry coefficient* of the species in that reaction;
- the *prefix combinator* "op" represents the role of the element in the reaction. Specifically, $\downarrow$ indicates a *reactant*, $\uparrow$ a *product*, $\oplus$ an *activator*, $\ominus$ an *inhibitor* and $\odot$ a generic *modifier*.

The choice operator, cooperation and definition of constant are unchanged. As in PEPA, we have $\mathcal{L} \subseteq \mathcal{A}$. In contrast to PEPA the hiding operator is omitted, as it is not necessary for our purposes.

In order to fully describe a biochemical network in Bio-PEPA we need to define structures that collect information about the compartments, the species, the constant parameters and the functional rates. In the following the function *name* returns the names of the elements of a given Bio-PEPA component.

Definition 1 *Each compartment is described by "V: v unit", where V is the compartment name, "v" is a positive real number expressing the compartment size and the (optional) "unit" denotes the unit associated with the compartment size. The set of compartments is denoted* $\mathcal{V}$.

The list of compartments is composed of at least one compartment. When no information about compartments is available we add a default compartment whose size is 1 and the unit of which depends on the model.

For each species represented in the system we can add some details that can then be used for the analysis. In the definition below the symbol "_" denotes the empty string.

Definition 2 *For each species we define the element "*$C : H = value\_H, N = value\_N, M = value\_M, V = value\_V, unit = value\_u$*", where:*

- *C is the species component name,*
- H *is the step size and* $value\_H \in \mathbb{R}^+$,
- N *is the maximum level and* $value\_N \in \mathbb{N}$,
- M *is the maximum concentration and* $value\_M \in \mathbb{R}^+ \cup \{\_\}$,
- V *is the name of the enclosing compartment and* $value\_V \in name(\mathcal{V}) \cup \{\_\}$,
- *value_u represents the unit for concentration.*

The set of all the elements described above is denoted $\mathcal{N}$.

All the elements described above can be optional and their use depends on the kind of analysis we aim to perform. If only the compartment name is given, we can use the system for stochastic simulation and we can map it to ODEs, whereas if we are

interested in the CTMC with levels or model checking with PRISM we also need the values for H and N (or, equivalently, H and M).

In order to collect the information about the dynamics of the system, we associate a functional rate f_{α_j} with each action α_j. This function represents the kinetic law of the associated reaction. For the definition of functional rates we consider mathematical expressions with simple operations and operators involving constant parameters and components. All the kinetic laws proposed in the book by Segel [46] can be defined in this way. In addition, for convenience, we include some predefined functions to express the most commonly used kinetic laws.

Definition 3 *The functional rates are expressed by the following grammar:*

$$
\begin{aligned}
f\_rate &::= f_\alpha(\bar{k}, \bar{C}) = sk \mid f_\alpha(\bar{k}) = sk2 \\
sk &::= int \mid float \mid name \mid sk + sk \mid sk \times sk \mid sk/sk \mid sk - sk \mid sk^{sk} \mid \\
&\quad exp(x) \mid log(sk) \mid sin(sk) \mid cos(sk) \\
sk2 &::= fMA(sk) \mid fMM(sk, sk) \mid fH(sk, sk, int)
\end{aligned}
$$

The set of functional rates is denoted $\mathcal{F}_R$.

The mathematical expressions are defined by some mathematical operators (*sk*) and the predefined functions (*sk2*). The general expression for the functional rate contains the names of the parameters and the names of the species components involved in the associated reaction. The predefined kinetic laws considered here are mass-action (*fMA*), Michaelis-Menten (*fMM*) and Hill kinetics (*fH*). They depend only on some parameters; the components/species are derived from the context. The functional rates are defined externally to the components and are evaluated when the system is derived. They are used to derive the transition rates of the system. In the functional rates some parameter constants can be used. These must be defined in the model by means of the set of parameter definitions $\mathcal{K}$.

Definition 4 *Each parameter is defined by "k_{name} = value unit", where "$k_{name} \notin C$" is the parameter name, "value" denotes a positive real number and the (optional) "unit" denotes the unit associated with the parameter. The set of the parameters is denoted $\mathcal{K}$.*

Finally, we have the following definition for the set of sequential components:

Definition 5 *The set Comp of sequential components is defined as*

$$Comp ::= \{C \stackrel{\text{def}}{=} S, \ \textit{where } S \textit{ is a sequential component } \}$$

We can define the Bio-PEPA system in the following way:

Definition 6 *A Bio-PEPA system $\mathcal{P}$ is a 6-tuple $\langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}_R, Comp, P \rangle$, where:*

- *$\mathcal{V}$ is the set of compartments;*
- *$\mathcal{N}$ is the set of quantities describing each species;*
- *$\mathcal{K}$ is the set of parameter definitions;*
- *$\mathcal{F}_R$ is the set of functional rate definitions;*
- *$Comp$ is the set of definitions of sequential components;*
- *P is the model component describing the system.*

In a *well-defined* Bio-PEPA system each element has to satisfy some conditions. The list $\mathcal{N}$ has to contain all the species components and, for each of them, at least its compartment must be defined. The optional elements of $\mathcal{N}$ must satisfy some simple conditions, for instance $value\_H > 0$ and $value\_H \in \mathbb{R}^+$ and $value\_N \in \mathbb{N}$ with $value\_N \geq 1$. Concerning the functional rates, they are well-defined if each variable in their definition refers to the name of a species component in the list $\mathcal{N}$ or a constant parameter in the list $\mathcal{K}$. For the definition of the species components, we have that each component $C \in Comp$ must have subterms of the form "(α, κ) op C" and the action types in each single component must be all distinct. Finally, the model component P must be defined in terms of the species components defined in $Comp$ and, for each cooperation set $\mathcal{L}_j$ in P, $\mathcal{L}_j \subseteq \mathcal{A}(P)$. Moreover, the initial value for each species must be a non-negative real number less than or equal to the maximum value, when given. See [15] for further details.

In the following we consider only well-defined Bio-PEPA systems. We denote by $\tilde{\mathcal{P}}$ the set of well-defined Bio-PEPA systems.

A Bio-PEPA system, with species components parametrised by concentration, is an implicitly continuous-state based representation. However an objective is to allow analysis of Bio-PEPA models by a variety of techniques, some of which are based on discrete-state representation. In particular we derive the CTMC with levels via a structured operational semantics and labelled transition system, after the continuous concentration parameter has been discretised into levels. We define a function *levels* over $\tilde{\mathcal{P}}$, which, given a Bio-PEPA system $\mathcal{P}$, derives the Bio-PEPA system $\mathcal{P}_l$, where the initial concentrations are replaced by the initial levels in the model component. This is possible only if the set $\mathcal{N}$ contains the step size for all the species. $\mathcal{P}_l$ is called a "*Bio-PEPA system with levels*"; it is well-defined if $\mathcal{P}$ is well-defined and if the levels for each each species are less than or equal to the maximum ones. In the following we omit the subscript "l" when it is clear from the context.

4.3 From biochemical networks to Bio-PEPA

The translation tr_BM_BP of a biochemical network $\mathcal{M}$ into a Bio-PEPA system $\mathcal{P} = \langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}_R, Comp, P \rangle$ is based on the following abstractions:

(1) Each compartment is defined in the set $\mathcal{V}$ in terms of a name and an associated volume. In this version of Bio-PEPA compartments are not involved actively in the reactions and therefore are not represented by processes.
(2) Each species i in the network is described by a species component $C_i \in Comp$. The constant component C_i is defined by the "sum" of *elementary components* (i.e. prefixes term) describing the interaction capabilities of the species. We suppose that there is at most one elementary component in each species component with an action of type α. A single definition can express the behaviour of the species at any level.
(3) Each reaction j is associated with an action type α_j and its dynamics is described by a specific function $f_{\alpha_j} \in \mathcal{F}_R$. The constant parameters used in the function can be defined in $\mathcal{K}$.
(4) The model P is defined as the cooperation of the different components C_i.

4.4 Some examples

The following examples show how some biochemical situations can be specified in Bio-PEPA.

4.4.1 Example 1: Mass-action kinetics

Consider the reaction $2X + Y \xrightarrow{f_M} 3Z$, described by the mass-action kinetic law $f_M = r \times X^2 \times Y$. The three species can be specified by the syntax:

$$X \stackrel{def}{=} (\alpha, 2)\downarrow X \quad Y \stackrel{def}{=} (\alpha, 1)\downarrow Y \quad Z \stackrel{def}{=} (\alpha, 3)\uparrow Z$$

The system is described by $(X(x_0) \underset{\{\alpha\}}{\bowtie} Y(y_0)) \underset{\{\alpha\}}{\bowtie} Z(z_o)$, where x_0, y_0 and z_0 denote the initial concentrations of the three components. The functional rate is $f_\alpha = fMA(r)$.

4.4.2 Example 2: Michaelis-Menten kinetics

One of the most commonly used kinetic laws is Michaelis-Menten. It describes a basic enzymatic reaction from the substrate S to the product P and is written as $S \xrightarrow{E;f_E} P$, where E is the enzyme involved in the reaction and $f_E = \frac{v_M \times E \times S}{(K_M + S)}$. For more details about this kinetic law see [46].

The three species can be specified in Bio-PEPA by the following components:

$$S \stackrel{def}{=} (\alpha, 1)\downarrow S \quad P \stackrel{def}{=} (\alpha, 1)\uparrow P \quad E \stackrel{def}{=} (\alpha, 1) \oplus E$$

The system is described by $(S(x_{S,0}) \underset{\{\alpha\}}{\bowtie} E((x_{E,0})) \underset{\{\alpha\}}{\bowtie} P(x_{P,0})$, where $x_{S,0}$, $x_{E,0}$ and $x_{P,0}$ are the initial concentration of the three species and the functional rate is f_α =

$fMM(v_M, K_M)$.

5 Definition of the labelled transition system with levels for Bio-PEPA

Bio-PEPA is given an operational semantics, similar to the one for PEPA. In this context we consider the abstraction for the species in terms of levels of concentration. Therefore, we have to consider the Bio-PEPA sytem with levels, as described at the end of Section 4.2. For the rest of this section we consider Bio-PEPA systems with levels, i.e the model component ha discrete parameters.

We define two relations over the processes. The former, called the *capability relation*, supports the derivation of quantitative information and it is auxiliary to the latter which is called the *stochastic relation*. The stochastic relation gives us the rates associated with each action. The rates are obtained by evaluating the functional rate associated with the action, divided by the step size, and by using the quantitative information derived from the capability relation. The use of two relations makes the definition of the semantics rules straightforward. This allows us to associate the rate with the last step of the derivation representing a given reaction. If only one relation were considered it could be complicated to derive the rate in the appropriate way, especially in the case of general kinetic laws different from mass-action.

The capability relation is $\rightarrow_c \subseteq C \times \Theta \times C$, where the label $\theta \in \Theta$ contains the quantitative information needed for the evaluation of the functional rate. We define the labels θ as $\theta := (\alpha, w)$, where w is a list recording the species that participate in the reaction and is defined as $[S : op(l, \kappa)] \mid w :: w$, with $S \in C$, l the level and κ the stoichiometry coefficient. The order of the components is not important. The relation $\rightarrow_c$ is the minimum relation satisfying the rules reported in Table 1.

The first three axioms describe the behaviour of the three different prefix terms. In the case of a reactant, the level decreases, in the case of a product the level increases whereas in the case of a modifier the level remains the same. Concerning the reactants and the products, the number of levels that changes depends on the stoichiometric coefficient κ. This expresses the degree to which a species (reactant or product) participates in a reaction. Some side conditions concerning the present concentration level must be added to the rules. Specifically, for the reactants the level has to be greater than or equal to κ, whereas for the products the level has to be less than or equal to $(N - \kappa)$, where N is the maximum level. For the modifiers, we have different side conditions according to the specific role of the species. When enzymes are considered, the level has to be greater than zero and less than or equal to the maximum level whereas, in the other cases, the level can be also equal to zero. Indeed if the enzyme is null the rate of the enzymatic reaction with Michaelis-Menten kinetics is zero and the reaction is not possible. This does not happen for

`prefixReac` $((\alpha,\kappa)\downarrow S)(l) \xrightarrow{(\alpha,[S:\downarrow(l,\kappa)])}_c S(l-\kappa) \quad \kappa \le l \le N$

`prefixProd` $((\alpha,\kappa)\uparrow S)(l) \xrightarrow{(\alpha,[S:\uparrow(l,\kappa)])}_c S(l+\kappa) \quad 0 \le l \le (N-\kappa)$

`prefixMod` $((\alpha,\kappa)\, op\, S)(l) \xrightarrow{(\alpha,[S:op(l,\kappa)])}_c S(l)$ with $op = \odot, \oplus, \ominus$ and

$0 < l \le N$ if $op = \oplus$, $0 \le l \le N$ otherwise

`choice1` $$\frac{S_1(l) \xrightarrow{(\alpha,w)}_c S_1'(l')}{(S_1+S_2)(l) \xrightarrow{(\alpha,w)}_c S_1'(l')}$$ `choice2` $$\frac{S_2(l) \xrightarrow{(\alpha,w)}_c S_2'(l')}{(S_1+S_2)(l) \xrightarrow{(\alpha,w)}_c S_2'(l')}$$

`constant` $$\frac{S(l) \xrightarrow{(\alpha,S:[op(l,\kappa)])}_c S'(l')}{C(l) \xrightarrow{(\alpha,C:[op(l,\kappa)])}_c S'(l')} \text{ with } C \stackrel{def}{=} S$$

`coop1` $$\frac{P_1 \xrightarrow{(\alpha,w)}_c P_1'}{P_1 \underset{\mathcal{L}}{\bowtie} P_2 \xrightarrow{(\alpha,w)}_c P_1' \underset{\mathcal{L}}{\bowtie} P_2} \text{ with } \alpha \notin \mathcal{L}$$

`coop2` $$\frac{P_2 \xrightarrow{(\alpha,w)}_c P_2'}{P_1 \underset{\mathcal{L}}{\bowtie} P_2 \xrightarrow{(\alpha,w)}_c P_1 \underset{\mathcal{L}}{\bowtie} P_2'} \text{ with } \alpha \notin \mathcal{L}$$

`coop3` $$\frac{P_1 \xrightarrow{(\alpha,w_1)}_c P_1' \quad P_2 \xrightarrow{(\alpha,w_2)}_c P_2'}{P_1 \underset{\mathcal{L}}{\bowtie} P_2 \xrightarrow{(\alpha,w_1::w_2)}_c P_1' \underset{\mathcal{L}}{\bowtie} P_2'} \text{ with } \alpha \in \mathcal{L}$$

Table 1
Axioms and rules for Bio-PEPA.

reactions representing inhibition. The rules `choice1` and `choice2` have the usual meaning. The rule `constant` is used to define the behaviour of the constant term, defined by one or more prefix terms in summation. The label contains the information about the level and the stoichiometric coefficient related to the action α. The last three rules report the case of cooperation. The rules `coop1` and `coop2` concern the case when the action enabled does not belong to the cooperation set. In this case the label in the conclusion contains only the information about the component that fires the action. The rule `coop3` describes the case in which the two components synchronise and the label reports the information from both the components.

In order to associate the rates with the transitions we introduce the stochastic relation $\rightarrow_s \subseteq \tilde{\mathcal{P}} \times \Gamma \times \tilde{\mathcal{P}}$, where the label $\gamma \in \Gamma$ is defined as $\gamma := (\alpha, r_\alpha)$, with $r_\alpha \in \mathbb{R}^+$.

In this definition r_α represents the parameter of a negative exponential distribution. The dynamic behaviour of processes is determined by a *race condition*: all activities enabled attempt to proceed but only the fastest succeeds. The relation $\rightarrow_s$ is defined as the minimal relation satisfying the rule

$$\texttt{Final} \quad \frac{P \xrightarrow{(\alpha_j, w)}_c P'}{\langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}, Comp, P \rangle \xrightarrow{(\alpha_j, r_\alpha[w, \mathcal{N}, \mathcal{K}])}_s \langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}, Comp, P' \rangle}$$

The second element in the label of the conclusion represents the transition. The rate is calculated from the functional rate f_α in the following way:

$$r_\alpha[w, \mathcal{N}, \mathcal{K}] = \frac{f_\alpha[w, \mathcal{N}, \mathcal{K}]}{H}$$

where H is the step size for the species involved in the reaction and the notation $f_\alpha[w, \mathcal{N}, \mathcal{K}]$ means that the function f_α is evaluated over w, $\mathcal{N}$ and $\mathcal{K}$. In detail, for each component C_i we derive the concentration as $l_i \times H$. Then we replace each free occurrence of C_i with $(l_i \times H)^{\kappa_{ij}}$, where κ_{ij} is the stoichiometric coefficient of the species i with respect to the reaction R_j. Some observations about the derivation of the rate are reported in Subsection 5.1.

A *Stochastic Labelled Transition System* can be defined for a Bio-PEPA system.

Definition 7 *The Stochastic Labelled Transition System (*SLTS*) for a Bio-PEPA system is* $(\tilde{\mathcal{P}}, \Gamma, \rightarrow_s)$*, where* $\rightarrow_s$ *is the minimal relation satisfying the rule* `Final`*.*

The states of the *SLTS* are defined in terms of the concentration levels of the system components and the transitions from one state to another represent reactions that cause changes in the concentration levels of some components.

Note that using the relation $\rightarrow_c$ it is possible to define another labelled transition system (*LTS*) as $(C, \Theta, \rightarrow_c)$.

5.1 Derivation of rates

In the *SLTS* the states represent *levels of concentration* and the transitions cause a change in these levels for one or more species. The number of levels depends on the stoichiometric coefficients of the species involved.

In [10] it was shown how to derive the transition rates in some specific cases. In the following we extend this approach to Bio-PEPA. The derivation is valid even when species have different numbers of levels and maximum concentrations.

Let us consider a reaction j described by a *kinetic law* f_j and with all stoichiometric

coefficients equal to one. Following [10], we can define the transition rate as $(\Delta t)^{-1}$, where Δt is the time to have a variation in the concentration of one step for both the reactants and the products of the reaction. Let y be a variable describing one product of the reaction. We can consider the rate equation for that species with respect to the given reaction. This is $dy/dt = f_j(\bar{x}(t))$, where $\bar{x}$ is the set (or a subset) of the reactants/modifiers of the reaction. We can apply the *Taylor expansion* up to the second term and we obtain

$$y_{n+1} \approx y_n + f(\bar{x}_n) \times (t_{n+1} - t_n)$$

Now we can fix $y_{n+1} - y_n = H$ and then derived the time interval $(t_{n+1} - t_n) = \Delta t$ as $\Delta t \approx H/f(\bar{x}_n)$. From this we obtain the transition rate as $f(\bar{x}_n)/H$.

When the reaction has stoichiometric coefficients different from one, we can consider an approach similar to the one above. Let y be a product of the reaction. The approximation gives:

$$y_{n+1} \approx y_n + r \times \kappa \times \prod_{i=1}^{n_r} x_{i,n}^{\kappa_i} \times (t_{n+1} - t_n)$$

where r is the reaction constant rate, κ is stoichiometric coefficient of the product y, x_i $i = 1, ..., n_r$ are the reactants of the reaction, κ_i $i = 1, ..., n_r$ are the associated stoichiometric coefficients, n_r is the number of distinct reactants.

Now we can fix $y_{n+1} - y_n = \kappa \times H$ and then derive the respective $(t_{n+1} - t_n) = \Delta t$ as $\Delta t \approx H/(r \times \prod_{i=1}^{n_r} x_{i,n}^{\kappa_i})$. From this expression we can derive the rate as usual.

Some observations follow. First of all, this approach is based on an *approximation*; it depends on the time/concentration steps. Secondly, we assume that the species can vary by one step size H at a time interval. Reactants are assumed to decrease until 0 and products increase until a given value. This implies that the kinetic law has to be *non-decreasing* in terms of the reactant concentrations. Mass-action, Hill-kinetics and Michaelis-Menten are all non-decreasing, as are many other kinetic laws.

5.2 Some examples (continued)

In the following we show the transition rates for the examples in Section 4.4.

5.2.1 Example 1: Mass-action kinetics

In the case of levels of concentration, the the model component is described by $(X(l_{X0}) \underset{\{\alpha\}}{\bowtie} Y(l_{Y0})) \underset{\{\alpha\}}{\bowtie} Z(l_{Z0})$, where l_{X0}, l_{Y0} and l_{Z0} denote the initial levels of the three components and are derived from the initial concentrations.

The rate associated with a transition is given by:

$$r_\alpha = \frac{r \times (l_X \times H)^2 \times (l_Y \times H)}{H}$$

where l_X, l_Y are the concentration levels for the species X and Y in a given state and H is the step size of all the species.

5.2.2 Example 2: Michaelis-Menten kinetics

The model component with levels of concentration is described by $(S(l_{S0}) \underset{\{\alpha\}}{\bowtie} E(l_{E0})) \underset{\{\alpha\}}{\bowtie} P(l_{P0})$.

The transition rate is given by:

$$r_\alpha = \frac{v_M \times (l_S \times H) \times (l_E \times h)}{(K_M + l_S \times H)} \times \frac{1}{H}$$

where l_S, l_E are the concentration levels for the species S and E in a given state and H is the step size of all the species.

6 Auxiliary definitions

In this section we report some auxiliary definitions. First of all we define the set of action types enabled in a species or model component.

Definition 8 *The set of current action types enabled in the model component P, denoted $\mathcal{A}(P)$, is defined as:*

$$\mathcal{A}((\alpha, \kappa)\ \mathit{op}\ S) = \{\alpha\}$$

$$\mathcal{A}(S_1 + S_2) = \mathcal{A}(S_1) \cup \mathcal{A}(S_2)$$

$$\mathcal{A}(C) = \mathcal{A}(S) \text{ where } C \stackrel{\text{def}}{=} S$$

$$\mathcal{A}(S(l)) = \mathcal{A}(S)$$

$$\mathcal{A}(P_1 \underset{\mathcal{L}}{\bowtie} P_2) = \mathcal{A}(P_1) \backslash \mathcal{L} \cup \mathcal{A}(P_2) \backslash \mathcal{L} \cup (\mathcal{A}(P_1) \cap \mathcal{A}(P_2) \cap \mathcal{L})$$

If $\mathcal{P}$ is a Bio-PEPA system with model component P, the set of current action types enabled in $\mathcal{P}$ is $\mathcal{A}(\mathcal{P}) = \mathcal{A}(P)$.

The following definitions concern the derivative of a component, the derivative set and the derivative graph. We refer to the relation $\rightarrow_s$. The case of $\rightarrow_c$ is analogous, the only differences are in the label and in the fact that the stochastic relation refers to Bio-PEPA systems and the capability relation refers to model components.

Definition 9 *If $\mathcal{P} \xrightarrow{(\alpha,r)}_s \mathcal{P}'$ then $\mathcal{P}'$ is a one-step $\rightarrow_s$ system derivative of $\mathcal{P}$.*
If $\mathcal{P} \xrightarrow{(\alpha_1,r_1)}_s \mathcal{P}_1 \xrightarrow{(\alpha_2,r_2)}_s .... \xrightarrow{(\alpha_n,r_n)}_s \mathcal{P}'$ then $\mathcal{P}'$ is a system derivative of $\mathcal{P}$.

We can indicate the sequence $\xrightarrow{\gamma_1}_s \xrightarrow{\gamma_2}_s .... \xrightarrow{\gamma_n}_s$ with $\xrightarrow{\mu}_s$, where μ denotes the sequence $\gamma_1\gamma_2, ...\gamma_n$ (possibly empty).

Definition 10 *A system α-derivative of $\mathcal{P}$ is a system $\mathcal{P}'$ such that $\mathcal{P} \xrightarrow{(\alpha,r)}_s \mathcal{P}'$. For each $\alpha \in \mathcal{A}$ we have at most one system α-derivative of a system $\mathcal{P}$.*

Definition 11 *The system derivative set $ds(\mathcal{P})$ is the smallest set such that:*

- $\mathcal{P} \in ds(\mathcal{P})$;
- *if $\mathcal{P}' \in ds(\mathcal{P})$ and there exists $\alpha \in \mathcal{A}(\mathcal{P}')$ such that $\mathcal{P}' \xrightarrow{(\alpha,r)}_s \mathcal{P}''$ then $\mathcal{P}'' \in ds(\mathcal{P})$.*

Definition 12 *The system derivative graph $\mathcal{D}(\mathcal{P})$ is the labelled directed multi-graph whose set of nodes is $ds(\mathcal{P})$ and whose multi-set of arcs are elements in $ds(\mathcal{P}) \times ds(\mathcal{P}) \times \Gamma$.*

It is worth noting that in the case of well-defined Bio-PEPA components the multiplicity of $\langle \mathcal{P}_i, \mathcal{P}_j, \gamma \rangle$ is always one.

The definitions above refer to Bio-PEPA systems with levels. The only element of the system $\mathcal{P} = \langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}, Comp, P \rangle$ that evolves is the model component P. The other elements collect information about the compartments, the species, the rates and report the definition of the species components. They remain unchanged in the evolution of the system. In some cases it can be useful (and simpler) to focus on the model component instead of considering the whole system and use the other components for the derivation of the rates. We define a function $\pi_P(\mathcal{P}) = P$, that, given a Bio-PEPA system returns the model component. Then we define a *(component) derivative* of P by considering the model component P' of the system derivative of $\mathcal{P}$. Similarly, we define a *(component) α-derivative* of P, *(component) derivative set* $ds(P)$ and the *(component) derivative graph* $\mathcal{D}(P)$ starting from the definitions for the associated system $\mathcal{P}$.

In the derivation of the CTMC (see Section 8.1) we need to identify the actions describing the transitions from one state to another.

Definition 13 *Let $\mathcal{P}$ be a Bio-PEPA system and let $P = \pi_P(\mathcal{P})$. Let P_u, P_v be two derivatives of a model component P with P_v a one-step derivative of P_u. The set of action types associated with the transitions from the process P_u to the process P_v is*

denoted $\mathcal{A}(P_u|P_v)$.

The next definition concerns the *complete action type set* of a system $\mathcal{P}$ and of a component P.

Definition 14 *The complete action type set of a system* $\mathcal{P}$ *is defined as:*

$$\bar{\mathcal{A}} = \cup_{\mathcal{P}_i \in ds(\mathcal{P})} \mathcal{A}(\mathcal{P}_i)$$

The complete action type set of a component P is defined similarly.

Other useful definitions are the ones concerning the exit rate and transition rates. In the following we report the definition for the model components, but a similar definition can be used for Bio-PEPA systems.

Definition 15 *Let us consider a Bio-PEPA system* $\mathcal{P} = \langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}, Comp, P \rangle$ *and let* $P_1, P_2 \in ds(P)$*. The exit rate of a process* P_1 *is defined as:*

$$rate(P_1) = \sum_{\{\alpha | \exists \mathcal{P}_2. \mathcal{P}_1 \xrightarrow{(\alpha, r_\alpha[w,\mathcal{N},\mathcal{K}])}_s \mathcal{P}_2,\ P_1 = \pi_P(\mathcal{P}_1)\}} r_\alpha[w, \mathcal{N}, \mathcal{K}]$$

Similarly, the transition rate is defined as:

$$rate(P_1 \mid P_2) = \sum_{\{\alpha | \mathcal{P}_1 \xrightarrow{(\alpha, r_\alpha[w,\mathcal{N},\mathcal{K}])}_s \mathcal{P}_2,\ P_1 = \pi_P(\mathcal{P}_1),\ P_2 = \pi_P(\mathcal{P}_2)\}} r_\alpha[w, \mathcal{N}, \mathcal{K}]$$

Given the transition labels it can be useful to define some functions to extract information from them. For the label θ in the capability relation, the function $action(\theta) = \alpha$ extracts the former element of the pair (i.e. the action type) and $list(\theta) = w$ returns the second element (i.e. the vector of quantitative information). Furthermore, the functions $reacts(\theta)$, $prods(\theta)$ $mods(\theta)$, $enzs(\theta)$, $inhibs(\theta)$, $totMods(\theta)$ return the sets of component names that are indicated as reactants, products, generic modifiers, enzymes, inhibitors and any of the last three possibilities from the vector w, respectively. The functions $\#reacts$, $\#prods$,... return the number of elements involved as reactants, products and so on. For the label γ in the stochastic relation, the function $action(\gamma) = \alpha$ extracts the first element of the pair (i.e. the action type) and the function $rate(\gamma) = r \in \mathbb{R}$ returns the second element (i.e. the rate).

7 Equivalences

It is sometimes useful to consider *equivalences* between models in order to determine whether the systems represented are in some sense the "same". In this

section we present some notions of equivalence for Bio-PEPA with levels in the model component. Some characteristics of the language impact on the definitions of equivalence and we start by highlighting those. Firstly, there is no hiding operator or τ actions. Therefore, in Bio-PEPA we do not have weaker forms of equivalence based on abstracting τ actions. Secondly, in well-defined systems we have at most one action of a given type in each sequential term and each component describes the behaviour of a single species. So we cannot have processes of the form "$S + S$" or terms such as "$A = a.C$" (where A and C differ). Thirdly, if we have two transitions between the processes P and P', they involve different action types and they represent similar reactions that differ only in the kind/number of modifiers. Finally, we have defined two relations within the semantics. In one case the labels contain the information about the action type and about the elements involved. This is used as an auxiliary relation for the derivation of the second one, in which the labels contain the information about the action type and the rate (similarly to PEPA activity). Thus we have a choice of which relation on which to base each notion of equivalence.

In the case of Bio-PEPA we need to define equivalences both for systems and model components. It is worth noting that the only element that changes in the transitions of a Bio-PEPA system is the model component. All the other elements remain unchanged. We define equivalences for the Bio-PEPA systems in terms of equivalences for the model components. Specifically, we say that two Bio-PEPA systems $\mathcal{P}_1$ and $\mathcal{P}_2$ are equivalent if their respective model components are equivalent.

In the following we use the same symbol to denote equivalences for both the system and the corresponding model component. In this section we present definitions of isomorphism and strong bisimulation which are similar to the relations defined for PEPA in [36]. Furthermore we show some relationships between the defined equivalences.

7.1 Isomorphism

Isomorphism is a strong notion of equivalence based on the derivation graph of the components (systems). Broadly speaking, two components (systems) are isomorphic if they generate derivation graphs with the same structure and capable of carrying out exactly the same activities.

We have the following definition of isomorphism based on the capability relation:

Definition 16 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q, respectively. A function $\mathcal{F} : ds(P) \to ds(Q)$ is a component isomorphism between P and Q, with respect to $\to_c$, if $\mathcal{F}$ is an injective function and for any component $P' \in ds(P)$, $\mathcal{A}(P') = \mathcal{A}(\mathcal{F}(P'))$, with $r_\alpha[w, \mathcal{N}, \mathcal{K}] = r'_\alpha[w', \mathcal{N}', \mathcal{K}']$ for each $\alpha \in \mathcal{A}(P)$, where w and w' are the two second components of the relation*

label, associated with P' and $\mathcal{F}(P')$, and for all $\alpha \in \mathcal{A}$ the set of α-derivatives of $\mathcal{F}(P')$ is the same as the set of $\mathcal{F}-$images of the α-derivatives of P', with respect to $\rightarrow_c$.

The definition of isomorphism based on the capability relation is very strong since the labels in the derivative graph contain a lot of information. Formally, we can define isomorphic components in the following way:

Definition 17 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q. P and Q are isomorphic with respect to $\rightarrow_c$ (denoted $P =_c Q$), if there exists a component isomorphism $\mathcal{F}$ between them such that $\mathcal{D}(\mathcal{F}(P)) = \mathcal{D}(Q)$, where $\mathcal{D}$ denotes the derivative graph.*

We can now define when two Bio-PEPA systems are isomorphic.

Definition 18 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q. $\mathcal{P}_1$ and $\mathcal{P}_2$ are isomorphic with respect to $\rightarrow_c$ (denoted $\mathcal{P}_1 =_c \mathcal{P}_2$), if $P =_c Q$.*

For the stochastic relation we have the following three definitions.

Definition 19 *A function $\mathcal{F} : ds(\mathcal{P}_1) \rightarrow ds(\mathcal{P}_2)$ is a system isomorphism between $\mathcal{P}_1$ and $\mathcal{P}_2$, with respect to $\rightarrow_s$, if $\mathcal{F}$ is an injective function and for any system $\mathcal{P}'_1 \in ds(\mathcal{P}_1)$, $\mathcal{A}(\mathcal{P}'_1) = \mathcal{A}(\mathcal{F}(\mathcal{P}'_1))$, and for all $\alpha \in \mathcal{A}$, the set of system α-derivatives of $\mathcal{F}(\mathcal{P}'_1)$ is the same as the set of $\mathcal{F}-$images of the system α-derivatives of $\mathcal{P}'_1$, with respect to $\rightarrow_s$.*

Definition 20 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q. P and Q are isomorphic with respect to $\rightarrow_s$ (denoted $P =_s Q$), if there exists a system isomorphism $\mathcal{F}$ between $\mathcal{P}_1$ and $\mathcal{P}_2$ such that $\mathcal{D}(\mathcal{F}(\mathcal{P}_1)) = \mathcal{D}(\mathcal{P}_1)$.*

Definition 21 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q. $\mathcal{P}_1$ and $\mathcal{P}_2$ are isomorphic with respect to $\rightarrow_s$ (denoted $\mathcal{P}_1 =_s \mathcal{P}_2$), if $P =_s Q$.*

The next proposition reports some properties of the two notions of isomorphism.

Proposition 1 *The following properties hold.*

(1) Both $=_c$ and $=_s$ are equivalence relations.
(2) Both $=_c$ and $=_s$ are congruences.
(3) Isomorphic components ($=_c$ or $=_s$) generate identical Markov processes.
(4) $=_c \subset =_s$.

The proof of the first three points is analogous to the case of isomorphism for PEPA in [36]. The last point follows from the fact that in the former isomorphism we take into account the information in the vector w on the label of the capability relation,

in addition to the rate and the action type. Thus isomorphism $=_c$ is more strict.

7.1.1 Equational laws

In the following the symbol "=" denotes either $=_c$ or $=_s$. The proof follows the definition of isomorphism and the semantic rules.

Choice The laws for choice are:

1) $P + Q = Q + P$

2) $P + (Q + R) = (P + Q) + R$

Cooperation The laws for cooperation are:

(1) $P \underset{\mathcal{L}}{\bowtie} Q = Q \underset{\mathcal{L}}{\bowtie} P$

(2) $P \underset{\mathcal{L}}{\bowtie} (Q \underset{\mathcal{L}}{\bowtie} R) = (P \underset{L}{\bowtie} Q) \underset{\mathcal{L}}{\bowtie} R$

(3) $P \underset{\mathcal{K}}{\bowtie} Q = P \underset{\mathcal{L}}{\bowtie} Q \quad$ if $\mathcal{K} \cap (\bar{\mathcal{A}}(P) \cup \bar{\mathcal{A}}(Q)) = \mathcal{L}$

(4) $(P \underset{\mathcal{L}}{\bowtie} Q) \underset{K}{\bowtie} R = \begin{cases} P \underset{\mathcal{L}}{\bowtie} (Q \underset{\mathcal{K}}{\bowtie} R) \text{ if } \bar{\mathcal{A}}(R) \cap (\mathcal{L} \backslash \mathcal{K}) = \emptyset \wedge \bar{\mathcal{A}}(P) \cap (\mathcal{K} \backslash \mathcal{L}) = \emptyset \\ Q \underset{\mathcal{L}}{\bowtie} (P \underset{\mathcal{K}}{\bowtie} R) \text{ if } \bar{\mathcal{A}}(R) \cap (\mathcal{L} \backslash \mathcal{K}) = \emptyset \wedge \bar{\mathcal{A}}(Q) \cap (\mathcal{K} \backslash \mathcal{L}) = \emptyset \end{cases}$

Constant The law for constant is: If $A \stackrel{def}{=} P$ then $A = P$

In the case of Bio-PEPA systems we have the following law, that follows directly from the definition.

Bio-PEPA systems The law for Bio-PEPA systems is:

Let $\mathcal{P}_1$ and $\mathcal{P}_2$ be two Bio-PEPA systems, with $P = \pi_P(\mathcal{P}_1)$ and $Q = \pi_P(\mathcal{P}_2)$.

If $P = Q$ then $\mathcal{P}_1 = \mathcal{P}_2$.

7.2 Strong bisimulation

The definition of bisimulation is based on the *labelled transition system*. Strong bisimulation captures the idea that bisimilar components (systems) are able to perform the same actions with same rates resulting in derivatives that are themselves bisimilar. This makes the components (systems) indistinguishable to an external observer. We give two definitions according to the two relations.

In the case of the capability relation the label contains a lot of information. We can define different relations according to the information we want to consider. In the following we report two possible relations.

Definition 22 *A binary relation $\mathcal{R} \subseteq C \times C$ is a strong capability bisimulation if $(P, Q) \in \mathcal{R}$ implies for all $\alpha \in \mathcal{A}$:*

- *if $P \xrightarrow{\theta_1}_c P'$ then, for some Q' and θ_2, $Q \xrightarrow{\theta_2}_c Q'$ with $(P', Q') \in \mathcal{R}$ and*

(1) $action(\theta_1) = action(\theta_2) = \alpha$;

(2) $\#reacts(list(\theta_1)) = \#reacts(list(\theta_2))$, $\#prods(list(\theta_1)) = \#prods(list(\theta_2))$, $\#enzs(list(\theta_1)) = \#enzs(list(\theta_2))$,$\#inhibs(list(\theta_1)) = \#inhibs(list(\theta_2))$;

- *the symmetric definition with Q replacing P.*

Definition 23 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q, respectively. P and Q are strong capability bisimilar, written $P \sim_c Q$, if $(P, Q) \in \mathcal{R}$ for some strong capability bisimulation $\mathcal{R}$ and $r_\alpha[w, \mathcal{N}, \mathcal{K}] = r'_\alpha[w', \mathcal{N}', \mathcal{K}']$ for all $\alpha \in \mathcal{A}$, where w and w' are the two second components of the relation label, associated with P' and $\mathcal{F}(P')$.*

A condition concerning the transition rate is added. In the case of Bio-PEPA systems we have the following definition.

Definition 24 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q, respectively. $\mathcal{P}_1$, $\mathcal{P}_2$ are strong capability bisimilar, written $\mathcal{P}_1 \sim_c \mathcal{P}_2$, if $P \sim_c Q$.*

We can relax the second point in the Def. 22 omitting it entirely. In this way we obtain a weaker form of strong capability bisimulation. We denote this $P \sim_c^2 Q$ in the case of model components and $\mathcal{P}_1 \sim_c^2 \mathcal{P}_2$ in the case of systems.

The definition of *strong stochastic bisimulation* is reported below.

Definition 25 *A binary relation $\mathcal{R} \subseteq \tilde{\mathcal{P}} \times \tilde{\mathcal{P}}$ is a strong stochastic bisimulation, if $(\mathcal{P}_1, \mathcal{P}_2) \in \mathcal{R}$ implies for all $\alpha \in \mathcal{A}$:*

- *if $\mathcal{P}_1 \xrightarrow{\gamma_1}_s \mathcal{P}'_1$ then, for some $\mathcal{P}'_2$ and γ_2, $\mathcal{P}_2 \xrightarrow{\gamma_2}_s \mathcal{P}'_2$ with $(\mathcal{P}'_1, \mathcal{P}'_2) \in \mathcal{R}$ and*

(1) $action(\gamma_1) = action(\gamma_2) = \alpha$

(2) $rate(\gamma_1) = rate(\gamma_2)$

- *the symmetric definition with $\mathcal{P}_2$ replacing $\mathcal{P}_1$.*

Definition 26 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q, respectively. P and Q are strong stochastic bisimilar, written $P \sim_s Q$, if $(\mathcal{P}_1, \mathcal{P}_2) \in \mathcal{R}$ for some strong stochastic bisimulation $\mathcal{R}$.*

Definition 27 *Let $\mathcal{P}_1$, $\mathcal{P}_2$ be two Bio-PEPA systems whose model components are P and Q, respectively. $\mathcal{P}_1$, $\mathcal{P}_2$ are strong stochastic bisimilar, written $\mathcal{P}_1 \sim_s \mathcal{P}_2$, if $P \sim_s Q$.*

Some facts about the strong bisimulation relations are reported in the following proposition.

Proposition 2 *The following facts hold:*

(1) the bisimulations $\sim_c$, $\sim_c^2$ *and* $\sim_s$ *are all equivalences and congruences;*
(2) $\sim_c \subset \sim_c^2$;
(3) $\sim_s = \sim_c^2$;
(4) $=_c \subset \sim_c$ *and* $=_s \subset \sim_s$

The last point reports that two components that are isomorphic are also strong bisimilar. The proof is identical to the case for PEPA. From this some equational laws are defined for the bisimulation relation too.

7.2.1 Example

Consider the following systems representing two biological systems. The former corresponds to the example with Michaelis-Menten presented in Section 4.4, the other is a variant of it. The former system $\mathcal{P}_1$ represents a system described by an enzymatic reaction with kinetic law $\dfrac{v_1 \times E \times S}{K_1 + S}$, where S is the substrate and E the enzyme. We have that the set $\mathcal{N}_1$ is defined as "$S : H = h, N = N_S;\ P : H = h, N = N_P; E : H = 1, N = 1;$ " for some values of the step sizes and number of levels. The component and the model components are defined as:

$$S \stackrel{def}{=} (\alpha, 1)\downarrow S \quad E \stackrel{def}{=} (\alpha, 1) \oplus E \quad P \stackrel{def}{=} (\alpha, 1)\uparrow P$$

The model component P_1 is $(S(x_{S,0}) \underset{\{\alpha\}}{\bowtie} E(x_{E,0})) \underset{\{\alpha\}}{\bowtie} P(x_{P,0})$. The functional rate is $f_\alpha = fMM(v_1, K_1)$.

The second system $\mathcal{P}_2$ describes an enzymatic reaction where the enzyme is left implicit (it is constant). The rate is given by $\dfrac{v_1 \times S'}{K_1 + S'}$, where S' is the substrate.

We have that the set $\mathcal{N}_2$ is defined as "$S' : H = h, N = N_{S'};\ P' : H = h, N = N_{P'};$ ".

The components are defined as $S' \stackrel{def}{=} (\alpha, 1)\downarrow S'$ and $P' \stackrel{def}{=} (\alpha, 1)\uparrow P'$ and the model component P_2 is $S'(x_{S0}) \underset{\{\alpha\}}{\bowtie} P'(x_{P0})$. In this case $f_\alpha = \dfrac{v_1 \times S'}{K_1 + S'}$ and the component S' and P' have the same number of levels and the step size of S and P.

We have that $P_1 \sim_s P_2$, but $P_1 \nsim_c P_2$, because the number of enzymes is different. The same relations are valid if the systems rather than the model components are considered.

8 Analysis

A Bio-PEPA system is an *intermediate, formal, compositional* representation of the biological model. Based on this representation we can perform different kinds of analysis. In this section we discuss briefly how to use a Bio-PEPA system to derive a *CTMC with levels*, a set of *Ordinary Differential Equations* (*ODEs*), a *Gillespie simulation* and a *PRISM* model.

8.1 From Bio-PEPA to CTMC with levels

As for the reagent-centric view of PEPA, the CTMC associated with the system refers to the concentration levels of the species components. Specifically, the states of the CTMC are defined in terms of concentration levels and the transitions from one state to the other describe some variations in these levels. In the following we refer to Bio-PEPA systems with levels.

Theorem 1 *For any finite Bio-PEPA system* $\mathcal{P} = \langle \mathcal{V}, \mathcal{N}, \mathcal{K}, \mathcal{F}_R, Comp, P \rangle$*, if we define the stochastic process X(t) such that* $X(t) = P_i$ *indicates that the system behaves as the component* P_i *at time t, then X(t) is a Markov Process.*

The proof is not reproduced here but it is analogous the one presented for PEPA [36]. Instead of the PEPA activity we consider the label γ and the rate is obtained by evaluating the functional rate in the system. We consider finite models to ensure that a solution for the CTMC is feasible. This is equivalent to supposing that each species in the model has a maximum level of concentration.

Theorem 2 *Given* $(\tilde{\mathcal{P}}, \Gamma, \rightarrow_s)$*, let* $\mathcal{P}$ *be a Bio-PEPA system, with model component P. Let* $n_c = |ds(P)|$*, where* $ds(P)$ *is the derivative set of P. Then the infinitesimal generator matrix of the CTMC for* $\mathcal{P}$ *is a square matrix Q* $(n_c \times n_c)$ *whose elements* $q_{u,v}$ *are defined as*

$$q_{u,v} = \sum_{\alpha_j \in \mathcal{A}(P_u|P_v)} r_{\alpha_j}[w_u, \mathcal{N}, \mathcal{K}] \quad \text{if } u \neq v \qquad q_{u,u} = -\sum_{u \neq v} q_{u,v} \quad \text{otherwise.}$$

where P_u, P_v *are two derivatives of P.*

It is worth noting that the states of the CTMC are defined in terms of the derivatives of the model component. These derivatives are uniquely identified by the levels of species components in the system, so we can give the following definition of the CTMC states:

Definition 28 *The CTMC states derived from a Bio-PEPA system can be defined as vectors of levels* $\sigma = (l_1, l_2, ..., l_n)$*, where* l_i*, for* $i = 1, 2, ..., n$*, is the level of the*

species i and n is the total number of species.

This leads to the following proposition.

Proposition 3 *Let $\mathcal{P}$ be a Bio-PEPA system with model component P. Let P_u and P_v be two derivatives of P such that the latter is one-step derivative of the former. If there exist two action types α_1 and α_2 that belong to $\mathcal{A}(P_u|P_v)$ then:*

(1) $\alpha_1 \neq \alpha_2$*;*
(2) the two action types refer to two transitions/biological reactions that differ only in the modifiers.

If two transitions are possible between a pair of states, the actions involved are different and they represent reactions that differ only in the modifiers and/or the number of enzymes used. The former point follows from the definition of well-defined Bio-PEPA system. The second point follows because the only possibility to have two transitions between two given states is that the associated reactions have the same reactants and products. We can see this by observing that the states depend on the levels and the reactions cause some changes in these levels. The only elements involved that do not change during a reaction are the modifiers.

As mentioned earlier the CTMC with levels is an approximation of the continuous view of the system captured by ODEs. Its advantage is that the state space of the CTMC with levels can be considerably smaller than that generated by the molecular view of the system. This means that a variety of different analysis techniques such as passage time analysis and probabilistic model-checking are accessible. The CTMC with levels can also be regarded as an approximation of the CTMC underlying the mapping to a stochastic simulation model based on molecules. In this case the levels represent aggregations of molecules.

8.2 From Bio-PEPA to ODEs

The translation into ODEs is similar to the method proposed for PEPA (reagent-centric view) [7]. It is based on the syntactic presentation of the model and on the derivation of the stoichiometry matrix $D = \{d_{ij}\}$ from the definition of the components. The entries of the matrix are the stoichiometric coefficients and are obtained in the following way: for each component C_i consider the prefix subterms C_{ij} representing the contribution of the species i to the reaction j. If the term represents a reactant we write the corresponding stoichiometry κ_{ij} as $-\kappa_{ij}$ in the entry d_{ij}. For a product we write $+\kappa_{ij}$ in the entry d_{ij}. All other cases are null.

The derivation of ODEs from the Bio-PEPA system $\mathcal{P}$, hereafter called t_{ODE}, is based on the following steps:

(1) definition of the stoichiometry $(n \times m)$ matrix D, where n is the number of species and m is the number of reactions;
(2) definition of the *kinetic law vector* $(m \times 1)$ v_{KL} containing the kinetic law of each reaction;
(3) association of the variable x_i with each component C_i and definition of the vector $(n \times 1)$ $\bar{x}$.

The ODE system is then obtained as:

$$\frac{d\bar{x}}{dt} = D \times v_{KL}$$

with initial concentrations x_{i0}, for $i = 1, ..., n$.

8.3 From Bio-PEPA to stochastic simulation

Gillespie's stochastic simulation algorithm [32] is a widely-used method for the simulation of biochemical reactions. It deals with homogeneous, well-stirred systems in thermal equilibrium and constant volume, composed of n different species that interact through m reactions. Broadly speaking, the goal is to describe the evolution of the system $\mathbf{X}(t)$, in terms of the number of molecules of each species, starting from an initial state. Every reaction is characterised by a stochastic rate constant c_j, termed the *basal rate* (derived from the constant rate r by means of some simple relations proposed in [32,47]). Using this it is possible to calculate the *actual rate* $a_j(\mathbf{X}(t))$ of the reaction, that is the probability of the reaction R_j occurring in time $(t, t + \Delta t)$ given that the system is in a specific state.

The translation of a Bio-PEPA model for Gillespie simulation is similar to the approach proposed for ODEs. The initial number of molecules for the species i can be calculated easily from the concentration as $X_{i,0} = x_{i,0} \times v \times N_A$, where v is the volume of the compartment of the species and N_A is the Avogadro number, i.e the number of molecules in a mole of a substance. For details see [15]. The main drawbacks are the definition of the rates and the correctness of the approach for general kinetic laws and reactions with more than two reactants. Indeed Gillespie's stochastic simulation algorithm supposes elementary reactions with at most two reactants and constant rates (mass-action kinetics). If the model contains only this kind of reactions the translation is straightforward. If there are non-elementary reactions and general kinetic laws, it is a widely-used approach to consider them translated directly into a stochastic context. This is not always valid and some counterexamples have been demonstrated [6]. The authors of [6] showed that when Gillespie's algorithm is applied to Hill kinetics in the context of the transcription initiation of autoregulated genes, the magnitude of fluctuations is overestimated. The application of Gillespie's algorithm in the case of general kinetics laws is discussed by several authors [1,11]. Rao and Arkin [1] show that this approach is valid in the case

of some specific kinetic laws, such as Michaelis-Menten and inhibition. However, it is important to remember that these laws are approximations and that specific conditions (such as "$S \gg E$" in the case of Michaelis-Menten) hold. The derivation of Gillespie's rates for reactions with more than two reactants is presented in [47] and these reactions are supported by various stochastic simulators (for instance [4]). Here we follow the same approach proposed in [39]: we apply Gillespie's algorithm, but particular attention must be paid to the interpretation of the simulation results and to their validity.

8.4 From Bio-PEPA to PRISM

PRISM [45] is a probabilistic model checker, a tool for the formal modelling and analysis of systems which exhibit random or probabilistic behaviour. PRISM has been used to analyse systems from a various application domains. Models are described using the PRISM language, a simple state-based language and it is possible to specify quantitative properties of the system using a temporal logic, called *CSL* [2,12] (Continuous Stochastic Logic). For our purposes the underlying mathematical model of a PRISM model is a CTMC with levels. However we present the translation separately as the models are specified in the PRISM language.

The PRISM language is composed of *modules* and *variables*. A model is composed of a number of modules which can interact with each other. A module contains a number of local variables. The values of these variables at any given time constitute the state of the module. The global state of the whole model is determined by the local state of all modules. The behaviour of each module is described by a set of commands. Each update describes a transition which the module can make if the guard is true. A transition is specified by giving the new values of the variables in the module, possibly as a function of other variables. Each update is also assigned a probability (or in some cases a rate) which will be assigned to the corresponding transition.

We map Bio-PEPA systems to PRISM models where the variables express levels of concentration. Alternatively, it is possible to derive PRISM models where molecules are counted instead of levels. The two mappings are similar, the only differences are in the definition of the possible values for the species and the rates. Specifically, the values for the species are given in terms of levels or molecules and the rates must be chosen in order to take the interpretation into account. When levels are considered we use the rates defined above whereas in the case of molecules the rates are the ones for Gillespie's simulation. The maximum level/concentration for each species must be given in the specification of Bio-PEPA system and, if necessary, the maximum number of molecules can be derived from it.

In the following we focus on the definition of the the PRISM model in terms of

concentration levels. We have the following correspondences:

- The model is defined as **stochastic** (this term is used in PRISM for CTMC).
- Each element in the set of parameters $\mathcal{K}$ is defined as a *global constant*.
- The concentration step, the maximum number of levels and the volume size for each species are defined as *global constants*.
- Each species component is represented by a *PRISM module*. The species component concentration is represented by a *local variable* and it can (generally) assume values between 0 and N_i. For each sub-term (i.e. reaction where the species is involved) we have a definition of a *command*. The name of the command is related to the action α (and then to the associated reaction). The guards and the change in levels are defined according to whether the element is a reactant, a product or a modifier of the reactions.
- The functional rates are defined inside an auxiliary module.
- In PRISM the rate associated with an action is the *product* of the rates of the commands in the different modules that cooperate. For each reaction, we give the value "1" to the rate of each command involved in the reaction, with the exception of the command in the module containing the functional rates. In this case the rate is the functional rate f, expressing the kinetic law. The rate associated with a reaction is given by $1 \times 1 \times ... \times f = f$, as desired.

9 Example: a simple genetic network

In order to show how to model genetic networks in Bio-PEPA, we consider a model from [6]. The model describes a general genetic network with a negative feedback through dimers, such as the one representing the control circuit for the λ repressor protein CI of λ-phage in *E.Coli*.

In the present work the stochastic and deterministic simulations are obtained exporting the Bio-PEPA system by means of the maps described above.

9.1 The biological model

A schema of the model is reported in Figure 2. The model is composed of three biological entities that interact with each other through five reactions (of which one is reversible). The biological entities are the mRNA molecule (M), the protein in monomer form (P) and the protein in dimeric form ($P2$). The first reaction (1) is the transcription of the mRNA (M) from the genes/DNA (not considered explicitly). The protein P in the dimer form ($P2$), which is the final result of the network, has an inhibitory effect on this process. The second reaction (2) is the translation of the protein P from M. Other two reactions represent the degradation of M (3) and the

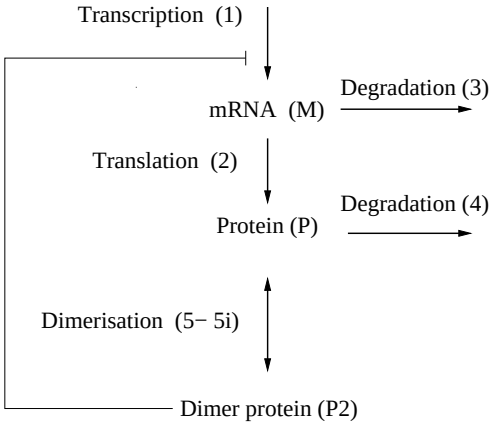

Fig. 2. Genetic network model

degradation of P (4). Finally there is the dimerization of P and its inverse process (5,5i). All the reactions are described by mass-action kinetics with the exception of the first reaction, which has a Michaelis-Menten kinetics.

9.2 The Bio-PEPA system

The translation of the model in Bio-PEPA is based on the following steps.

- *Definition of compartments.* The only compartment is defined as $v_{cell} : \ 1\ (nM)^{-1}$.
- *Definition of the set* $\mathcal{N}$.

$$M : \ H = 1, N = 1, V = v_{cell}, unit = nM;$$

$$P : \ H = 30, N = 2, V = v_{cell}, unit = nM;$$

$$P2 : \ H = 30, N = 6, V = v_{cell}, unit = nM;$$

  We consider $N = 2$ for P since the stoichiometry of P in the dimerization reaction is 2. For illustrative purposes we consider a minimal number of levels in order to keep the state space small.
- *Definition of the set of functional rates* $\mathcal{F}_R$.

$$f_{\alpha_1} = \tfrac{v}{K_M + P2};$$

$$f_{\alpha_2} = fMA(k2); \qquad f_{\alpha_3} = fMA(k3); \qquad f_{\alpha_4} = fMA(k4);$$

$$f_{\alpha_5} = fMA(k5); \qquad f_{\alpha_{5i}} = fMA(k5i)$$

  where the suffix of the action type α refers to the number of the reaction as reported in Fig. 2.
- *Definition of the set of parameters.* The parameter values are

$$K_M = 356\ nM; \quad v = 2.19\ s^{-1}; \quad k_2 = 0.043\ s^{-1}; \quad k_3 = 0.0039\ s^{-1};$$

$$k_4 = 0.0007\ s^{-1}; \quad k_5 = 0.025\ s^{-1} nM^{-1}; \quad k_{5i} = 0.5\ s^{-1}$$

- *Definition of the set of species components and of the model component.*

$$
\begin{aligned}
M &\stackrel{def}{=} (\alpha_2, 1) \oplus M + (\alpha_3, 1)\downarrow M + (\alpha_1, 1)\uparrow M \\
P &\stackrel{def}{=} (\alpha_4, 1)\downarrow P + (\alpha_5, 2)\downarrow P + (\alpha_{5i}, 2)\uparrow P) + (\alpha_2, 1)\uparrow P \\
P2 &\stackrel{def}{=} (\alpha_1, 1) \ominus P2 + (\alpha_{5i}, 1)\downarrow P2 + (\alpha_5, 1)\uparrow P2 \\
&(M(0) \underset{\{\alpha_2\}}{\bowtie} P(0)) \underset{\{\alpha_5, \alpha_{5i}\}}{\bowtie} P2(0)
\end{aligned}
$$

9.3 Analysis

The model is amenable to a number of different analyses as we report in the following paragraphs.

First of all, from the Bio-PEPA system we can derive the SLTS and the CTMC. Remember that both consider levels of concentration. The transition system consists of 42 states and 108 transitions, in the case we consider the information about species listed above. The states are described by the levels of the single components. Specifically, we can define a state using a vector $(M(l_M), P(l_P), P2(l_{P2}))$, where l_i, for $i = M, P, P2$, represents the level of each component. The parameter l_i can assume the values 0 and 1 in the case of M, the values 0, 1, 2 for P and values between 0 and 6 for $P2$. The labels γ_t of the stochastic transition system contain the action type α_j and the rate r_{α_j}, calculated by applying the associated function f_{α_j} to the quantitative information collected in the labels of the capability relation and dividing this by the step size of the reactants/products involved in the reaction. These rates are the ones associated with the CTMC transitions.

A second kind of analysis concerns differential equations. The stoichiometry matrix D associated with the system is

	R1	R2	R3	R4	R5	R5i	
M	+1	0	-1	0	0	0	x_1
P	0	+1	0	-1	-2	+2	x_2
P2	0	0	0	0	+1	-1	x_3

The kinetic-law vector is $v_{KL}^T = v/(K + x_3); k_2 \times x_1; k_3 \times x_1; k_4 \times x_2; k_5 \times x_2^2; k_{5i} \times x_3)$. The system of ODEs is obtained as $d\bar{x}/dt = D \times v_{KL}$:

$$\frac{dx_1}{dt} = \frac{v}{K + x_3} - k_3 \times x_1$$
$$\frac{dx_2}{dt} = k_2 \times x_1 - k_4 \times x_2 - 2 \times k_5 \times x_2^2 + 2 \times k_{5i} \times x_3$$
$$\frac{dx_2}{dt} = k_5 \times x_2^2 - k_{5i} \times x_3$$

The derivation of the Gillespie's simulation model is straightforward and not reported here.

The simulation results are depicted in Figure 3. We consider both deterministic and stochastic simulation. The two simulation graphs show the same behaviour (with the exception of some noise in the Gillespie's simulation), as expected.

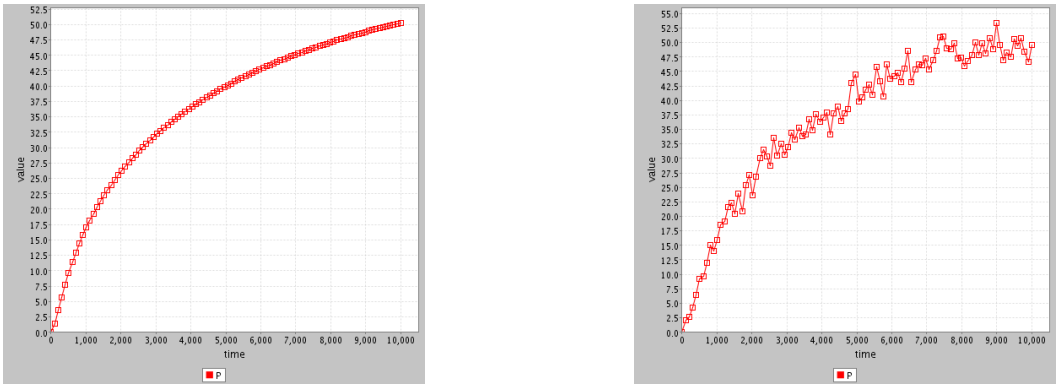

Fig. 3. ODE and Gillespie simulation results. In the case of Gillespie we consider 10 runs.

Finally, we consider the analysis by means of PRISM with levels. The full translation of the model into PRISM is reported in [15]. Each species is represented by a PRISM module and the reactions in which it is involved are captured by commands. In the following we report the definition of the modules representing the protein in the monomer and dimer form respectively.

module p

$p : [0..Np]$ **init** 0;

$[a2]\ p < Np \rightarrow (p' = p + 1);$

$[a4]\ p > 0 \rightarrow (p' = p - 1);$

$[a5]\ p > 0 \rightarrow (p' = p - 2);$

$[a5i]\ p < Np \rightarrow (p' = p + 2);$

endmodule

module pd

$p2 : [0..Np2]$ **init** 0;

$[a5i]\ p2 > 0 \rightarrow (p2' = p2 - 1);$

$[a5]\ p2 < Np2 \rightarrow (p2' = p2 + 1);$

endmodule

The variables p and $p2$ are *local* with respect to each of the two modules and represent the species "protein in monomer form" and "protein in dimer form", re-

spectively. The possible values are $[0..Np]$ for p and $[0..Np2]$ for $p2$, while the initial values are 0. The monomer P is involved in four reactions while the dimer form $P2$ in just two. We have an additional module with the functional rates.

Properties of the system can be expressed formally in *CSL* and analysed against the constructed model. Two simple examples of possible queries are considered below. A first query considers the probability that the monomer is at level i at time T. The property is expressed by the form "$P =?[...]$", that returns a numerical value representing the probability of the proposition inside the square brackets. In our case the query is $P =?[true\ U[T,T]\ p = i]$, where U is *the bounded until operator* and $[T,T]$ indicates a single time instant. A property of the form "*prop1* U[*time*] *prop2*" is true for a path if *time* defines an interval of real values and the path is such that *prop2* becomes true at a time instant which falls within the interval and *prop1* is true in all time instants up to that point. The second query concerns the proportion of the protein in monomer form (P) relative to the total quantity of the protein (i.e. $P + P2$). In order to define this property, we need a reward structure. State rewards can be specified using multiple reward items, each of the form "*guard:reward*;", where guard is a predicate and reward is an expression. States of the model which satisfy the predicate in the guard are assigned the corresponding reward. Specifically, in our case we define the reward:

rewards

$true : \frac{p}{(p+p2)}$;

endrewards

This reward assigns the value $\frac{p}{(p+p2)}$ to each state of the system. We can ask for the frequency of P by using the query $R =?[I = T]$. This is an *instantaneous reward property*, i.e. it refers to the reward of a model at a particular instant in time T. The property "$I = T$" associates with a path the reward in the state of that path when exactly T time units have elapsed. The letter "R" indicates that the property refers to a reward structure. The results of the two queries are reported in Fig. 4.

10 Conclusions

In this paper we have presented Bio-PEPA, a modification of the process algebra PEPA for the modelling and the analysis of biochemical networks. Bio-PEPA allows us to represent explicitly some features of biochemical networks, such as stoichiometry and general kinetic laws. Thus not only elementary reactions with constant rates, but also complex reactions described by general kinetic laws can be considered. The potential to consider various kinds of kinetic laws permits us to model a vast number of biochemical networks. Indeed complex reactions are

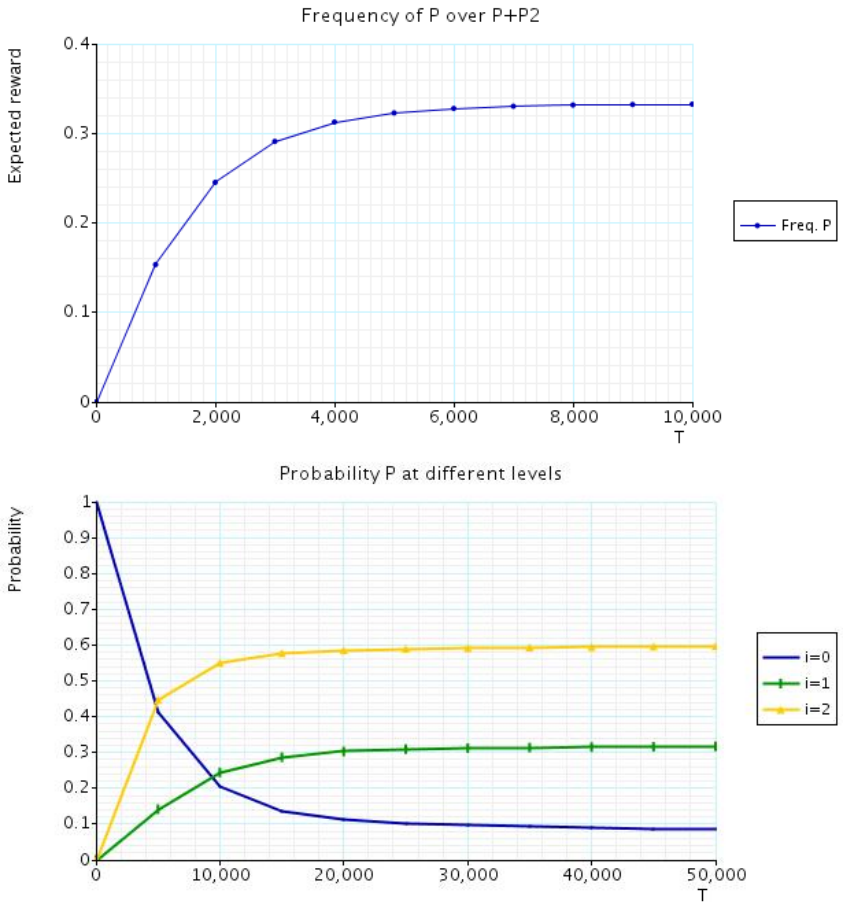

Fig. 4. PRISM query results. The figure at the top reports the graph of the proportion of monomer P over the total protein with respect to time.Below it is depicted the probability that the monomer protein is at levels 0, 1 and 2, with respect to time.

frequently found in models as abstractions of sequences of elementary steps and reducing to elementary reactions is often impossible and undesirable.

Bio-PEPA is in enriched with some notions of equivalence. We have presented definitions of isomorphism and strong bisimulation which are similar to the relations defined for PEPA in [36]. These equivalences are quite strict. A further investigation concerns the definition of other forms of equivalence, more appropriate for studying biological systems.

A principal feature of Bio-PEPA is the possibility of mapping the system to different kinds of analysis. In this work we have shown how to derive a CTMC with levels from a Bio-PEPA system and we have discussed the derivation of ODEs, stochastic simulation and PRISM models. Indeed Bio-PEPA has been defined as an intermediate language for the formal representation of the model. We have extended the definition of CTMC with levels, defined in [10], to the case of general kinetic laws and to different levels for the species. The main benefit of this approach with respect to stochastic simulation is the reduction in state space which leaves models

amenable to numerical solution and model checking. Compared with ODE-based analysis the important stochastic aspect of behaviour is retained. The approach is based on some assumptions.

First of all, all the species must have a finite maximum concentration. This is to ensure a finite state space in the corresponding CTMC, making numerical solution feasible. However, we can have a species without a limiting value. In these cases we can consider a maximum level for the values greater than a certain (high) value. A second point concerns the assumption that all the species have the same step size. This may be a problem when the species can have maximum concentrations belonging to different concentration scales; some species can have only few levels whereas others can have many. Furthermore, some species (for instance genes) are present in the system only in few copies and in this case the representation in terms of continuous concentration is wrong. In order to handle this situation, Bio-PEPA could be enriched with discrete variables. The possibility to consider different step sizes and discrete variables is a topic for future work.

The different kinds of analysis proposed for Bio-PEPA are strongly related. An area for further work will concern a deeper study of these relationships, in particular for the CTMC with levels and Gillespie models. An outstanding problem is the application of Gillespie's stochastic simulation with general kinetic laws. Indeed the original definition of the algorithm in [32] is based on the assumption of elementary reactions. However recently there have been some extensions to handle reactions with general kinetic laws and with more that two reactants. The approach proposed in this work is to use Gillespie simulation also in the general context, but to be careful about the interpretation of the results. The validation of the model against experimental data and prior knowledge is extremely important in this situation. In particular, if we obtain results different from the ones expected, the problem could be in the application of Gillespie's algorithm with the reactions present in the model.

In Bio-PEPA compartments are assumed to be static and are simply represented by names. This choice is motivated by the fact that, even though compartments play an important role in biological systems, at the present the quantitative information about them is poor. Most biochemical networks in the literature and databases (see for instance [42]) describe static compartments and often are based on strong assumptions. For example, all compartments are assumed to have the same volume or all the species are well-mixed when in reality they are not. In the present work we fix our attention to these kinds of network and Bio-PEPA is able to represent most features of them. In the future, we plan to extend the language in order model more complex definition of compartments, based on general assumptions.

Finally, a tool for the analysis of biochemical networks using Bio-PEPA is under implementation and a translation from SBML into Bio-PEPA is planned.

References

[1] A.P. Arkin, C.V. Rao, Stochastic chemical kinetics and the quasi-steady-state assumption: application to the Gillespie algorithm, *Journal of Chemical Physics* 11 (2003) 4999–5010.

[2] A. Aziz, K. Kanwal, V. Singhal, V. Brayton, Verifying continuous time Markov chains, Proc. *8th International Conference on Computer Aided Verification (CAV'96)*, LNCS 1102 (1996) 269–276, Springer.

[3] B.J. Bornstein, J.C. Doyle, A. Finney, A. Funahashi, M. Hucka, S.M. Keating, H. Kitano, B.L. Kovitz, J. Matthews, B.E. Shapiro and M.J. Schilstra, Evolving a Lingua Franca and Associated Software Infrastructure for Computational Systems Biology: The Systems Biology Markup Language (SBML) Project, *Systems Biology* 1 (2004) 41–53.

[4] S. Ramsey, D. Orrell and H. Bolouri. Dizzy: stochastic simulation of large-scale genetic regulatory networks. *J. Bioinf. Comp. Biol.* 3(2):415–436, 2005.

[5] L. Bortolussi, A. Policriti, Modeling Biological Systems in Stochastic Concurrent Constraint Programming,*Constraints* 13:1 (2008), also in Proc. of *WCB 2006*, 2006.

[6] R. Bundschuh, F. Hayot, C. Jayaprakash. Fluctuations and Slow Variables in Genetic Networks, *Biophys. J.* 84 (2003) 1606–1615.

[7] M. Calder, S. Gilmore, J. Hillston, Automatically deriving ODEs from process algebra models of signalling pathways, Proc. of *CMSB'05* , 204–215, 2005.

[8] M. Calder, S. Gilmore, J. Hillston, Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA. Proc. of Bionconcur'04. Extended version in *T. Comp. Sys. Biology*, VII, LNCS 4230 (2006) 1–23 Springer.

[9] M. Calder, A. Duguid, S. Gilmore and J. Hillston. Stronger computational modelling of signalling pathways using both continuous and discrete-space methods. Proc. of *CMSB'06*, LNCS 4210 (2006) 63–77.

[10] M. Calder, V. Vyshemirsky, D. Gilbert, and R. Orton, Analysis of Signalling Pathways using Continuous Time Markov Chains, *T. Comp. Sys. Biology*, VI, LNCS 4220 (2006) 44–67, Springer.

[11] Y. Cao, D.T. Gillespie and L. Petzold, Accelerated Stochastic Simulation of the Stiff Enzyme-Substrate Reaction. *J. Chem. Phys.* 123:14 (2005) 144917–144929.

[12] C. Baier, J.-P. Katoen and H. Hermanns, Approximate Symbolic Model Checking of Continuous-Time Markov Chains, *Proceedings of CONCUR'99*, LNCS 1664 (1999), 146–161.

[13] L. Cardelli, E.M. Panina, A. Regev, E. Shapiro and W. Silverman, BioAmbients: An Abstraction for Biological Compartments. *Theoretical Computer Science* 325:1 (2004) 141–167, Elsevier.

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

[15] F. Ciocchetta, and J. Hillston, Bio-PEPA: a framework for the modelling and analysis of biological systems, Technical report EDI-INF-RR-1231, University of Edinburgh, 2008.

[16] F. Ciocchetta and J. Hillston, Bio-PEPA: an extension of the process algebra PEPA for biochemical networks, Proc. of *FBTC 2007*, *Electronic Notes in Computer Science* 194:3 (2008) 103–117.

[17] F. Ciocchetta, S. Gilmore, M. L. Guerriero and J. Hillston, Stochastic Simulation and Probabilistic Model-Checking for the Analysis of Biochemical Systems, submitted to *CMSB 2008*.

[18] F. Ciocchetta, A. Degasperi, J. Hillston, M. Calder, Some investigations concerning the CTMC and the ODE model derived from Bio-PEPA, submitted to *FBTC 2008*.

[19] F. Ciocchetta, and C. Priami, Biological transactions for quantitative models, Proc. of *MeCBIC 2006*, ENTCS 171:2 (2007) 55–67.

[20] F. Ciocchetta, and C. Priami, Beta-binders with Biological Transactions, Technical report TR-10-2006, The Microsoft Research-University of Trento Centre for Computational and Systems Biology, 2006.

[21] F. Ciocchetta, C. Priami and P. Quaglia, Modeling Kohn Interaction Maps with Beta-Binders: An Example, T. Comp. Sys. Biology, III (2005), 33–48, Springer.

[22] G. Costantin, C. Laudanna, P. Lecca, C. Priami, P. Quaglia and B. Rossi, Language modeling and simulation of autoreactive lymphocytes recruitment in inflamed brain vessels, *SIMULATION: Transactions of The Society for Modeling and Simulation International* 80 (2003) 273–288.

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

[24] V. Danos, J. Feret, W. Fontana, R. Harmer and J. Krivine. Ruled-based modelling of cellular signalling. Proc. of *CONCUR'07*, LNCS, 4703 (2007).

[25] V. Danos, J. Feret, W. Fontana and J. Krivine. Scalable simulation of cellular signalling networks. Proc. of *APLAS'07*, 2007.

[26] V. Danos and J. Krivine, Formal molecular biology done in CCS-R, Proc. of *Workshop on Concurrent Models in Molecular Biology (BioConcur'03)*, 2003.

[27] L. Dematté, C. Priami, A. Romanel, Modelling and simulation of biological processes in BlenX, *SIGMETRICS Performance Evaluation Review* 35:4 (2008) 32–39.

[28] L. Dematté, C. Priami, A. Romanel, The BlenX language: A Tutorial, Chapter for the tutorial of SFM-08:Bio, LNCS, 5015 (2008), 313–365.

[29] C. Eichler-Jonsson, E.D. Gilles, G. Muller and B. Schoeberl, Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors, *Nature Biotechnology* 20 (2002) 370–375.

[30] T.S. Gardner, M. Dolnik, and J.J. Collins. A theory for controlling cell cycle dynamics using a reversibly binding inhibitor. *Proc. Nat. Acad. Sci. USA* 95 (1998) 14190–14195.

[31] N. Geisweiller, J. Hillston and M. Stenico, Relating continuous and discrete PEPA models of signalling pathways, *Theoretical Computer Science*, in press. Available online at www.science-direct.com, *doi* : 10.1016/*j.tcs*.2008.04.012

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

[33] A. Goldbeter, A Minimal Cascade Model for the Mitotic Oscillator Involving Cyclin and Cdc2 kinase, *Proc. Nat. Acad. Sci.* 8 (1991) 9107–9111.

[34] E.L. Haseltine and J.B. Rawlings, Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics, *J. Chem. Phys.* 117 (2006) 6959–6969.

[35] J. Heath, M. Kwiatkowska, G. Norman, D. Parker and O. Tymchyshyn, Probabilistic Model Checking of Complex Biological Pathways, Theoretical Computer Science (Special Issue on Converging Sciences: Informatics and Biology) 391 (2008) 239–257.

[36] J. Hillston, *A Compositional Approach to Performance Modelling*, Cambridge University Press, 1996.

[37] M. Kanehisa, A database for post-genome analysis, *Trends Genet.* 13 (1997) 375–376.

[38] *KEGG home page*, available at http://sbml.org/kegg2sbml.html.

[39] A.M. Kierzek and J. Puchalka, Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks, *BIOPHYS J.* volume 86 (2004) 1357–1372.

[40] C. Kuttler and J. Niehren, Gene regulation in the π-calculus: simulating cooperativity at the lambda switch, *Transactions on Computational Systems Biology* VII, LNCS 4230 (2006) 24–55, Springer.

[41] *NuMSV model checker*, available at http://nusmv.irst.itc.it.

[42] N. Le Novére, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J.L. Snoep, and M. Hucka, BioModels Database: a Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems, *Nucleic Acids Research* 34 (2006) D689–D691.

[43] C. Priami and P. Quaglia, Beta-binders for biological interactions, Proc. of *CMSB'04*, LNCS 3082 (2005) 20–33, Springer.

[44] C. Priami, A. Regev, W. Silverman and E. Shapiro, Application of a stochastic name-passing calculus to representation and simulation of molecular processes, *Information Processing Letters* 80 (2001) 25–31.

[45] Prism web site. http://www.prismmodelchecker.org/

[46] I.H. Segel, *Enzyme Kinetics: Behaviour and Analysis of Rapid Equilibrium and Steady-State Enzyme Systems*, Wiley-Interscience, New-York, 1993.

[47] O. Wolkenhauer, M. Ullah, W. Kolch and K. H. Cho, Modelling and Simulation of IntraCellular Dynamics: Choosing an Appropriate Framework, *IEEE Transactions on NanoBioScience* 3 (2004) 200–207.