Qualitative and Quantitative Analysis of a Bio-PEPA Model of the Gp130/JAK/STAT Signalling Pathway

Maria Luisa Guerriero

Laboratory for Foundations of Computer Science, The University of Edinburgh, UK

Abstract. Computational modelling of complex biochemical systems has grown in importance over recent years as a tool for supporting biological studies. Consequently, several formal languages have been recently proposed as modelling languages for biology. Among these, process algebras have been proved capable of providing researchers with new hypotheses on the behaviour of biochemical systems.
Bio-PEPA is a process algebra recently defined for the modelling and analysis of biochemical systems, which provides modellers with a wide range of analysis techniques: models can be analysed by stochastic simulation, model-checking, and mathematical methods based on ordinary differential equations.
In this work, we use Bio-PEPA for modelling the gp130/JAK/STAT signalling pathway, and we use both stochastic simulation and model-checking to analyse several qualitative and quantitative aspects of the system.

1 Introduction

Several modelling approaches have been used over recent years to analyse complex biological systems such as signaling pathways, ranging from traditional mathematical methods based on differential equations to computational methods based on stochastic simulation and model-checking. Each of these techniques can be more suitable than others in some context or to study some particular features of biological systems.

Process algebras are formal languages traditionally used to model distributed systems of concurrent computing devices. Starting from the biochemical π-calculus [1], several other process algebras have been recently adapted in order to model biochemical systems [2–5], following the "molecules as processes" paradigm introduced in the landmark paper [6]: molecules are modelled as concurrent processes, and biochemical reactions are represented by actions performed by synchronising processes.

Bio-PEPA [7, 8] is a process algebra specifically defined to model and analyse biochemical networks. Compared to other process algebras, Bio-PEPA uses a more abstract view of biochemical systems, the so-called "species as processes" abstraction: processes represent molecular species instead of single molecules, and multi-way synchronisations of processes represent changes in the amounts of molecular species resulting from biochemical reactions. Such an abstract view enables modellers to deal with analysis techniques which are computationally infeasible when considering the "molecules as processes" abstraction.

The main feature of Bio-PEPA is that it integrates several kinds of analysis techniques. Both discrete stochastic and continuous deterministic models can be automatically generated from Bio-PEPA models, thus allowing modellers to perform time-series

analysis via stochastic simulation, Markovian analysis and ordinary differential equations (ODEs); in addition, system properties can be verified through model-checking and mathematical techniques such as bifurcation, stability and continuation analysis. Moreover, as for the other process algebras, Bio-PEPA is equipped with an operational semantics which supports various kinds of formal analysis (e.g. causality, equivalence, and reachability analysis).

In this work, we define a Bio-PEPA model of the gp130/JAK/STAT signalling pathway, a well-studied system which plays a major role in several biological processes both in human and other organisms. A lot of experimental data is available about the molecules in the pathway, and some mathematical and computational models have been already developed. For these reasons, the gp130/JAK/STAT pathway represents a good case study for exploiting some of the possible Bio-PEPA analysis methods in order to study different aspects (both qualitative and quantitative) of the system, and compare them with existing models.

The rest of the paper is structured as follows. First, the Bio-PEPA language is introduced in Sec. 2, while the pathway and the Bio-PEPA model are described in Sec. 3 and Sec. 4, respectively. The following three sections are devoted to the analysis of the model: in Sec. 5 several qualitative properties are analysed via model-checking, in Sec. 6 we present some stochastic simulation results, and in Sec. 7 model-checking is employed for quantitative analysis. Finally, Sec. 8 is an overview of the related work and Sec. 9 contains some concluding remarks.

2 Bio-PEPA

Bio-PEPA [7, 8] is a process algebra which has been recently defined for the modelling and analysis of biochemical networks. It is a biologically-inspired language based on PEPA [9] and, differently from PEPA and other process algebras, it is able to explicitly represent details such as stoichiometric coefficients and the roles of species in reactions, and it supports the definition of general kinetic laws. Bio-PEPA models can be analysed by different techniques (stochastic simulation, analysis based on ODEs, numerical solution of the continuous-time Markov chain (CTMC), and probabilistic model-checking), since the mappings of Bio-PEPA models into specifications for those approaches have been defined [10].

The Bio-PEPA language is based on discrete levels of parameterised species: each component represents a species and its parameter may be interpreted as the number of molecules or discrete levels of concentration depending on the type of analysis to be applied. Parametric levels are considered for the definition of the transition system and for the derivation of a CTMC whose states represent the concentration levels of the species.

The syntax of Bio-PEPA is defined as:

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

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

The component S is called a *species component* and abstracts a molecular species, whereas the component P, called a *model component*, describes the system and the interactions among components. The prefix term $(\alpha, \kappa)\ \mathsf{op}\ S$ contains information about

the role of the species in the reaction associated with the action type α: κ is the *stoichiometric coefficient* of the species and the *prefix combinator* "op" represents its role in the reaction. Specifically, $\downarrow$ indicates a *reactant*, $\uparrow$ a *product*, $\oplus$ an *activator*, $\ominus$ an *inhibitor* and $\odot$ a generic *modifier*. The operator "+" expresses the choice between possible actions and the constant C is defined by an equation $C \stackrel{def}{=} S$. The parameter $x \in \mathbb{R}^+$ in $S(x)$ represents the concentration of S. Finally, the process $P \underset{\mathcal{I}}{\bowtie} Q$ denotes the cooperation between components: the set $\mathcal{I}$ determines those activities on which the operands are forced to synchronise. Reaction rates are defined as *functional rates* associated with actions.

Bio-PEPA supports a modelling style in terms of *concentration levels*: the species amounts are discretised into a number of levels, from level 0 (i.e. species not present) to a maximum level N (which depends on the maximum concentration of the species). Each level represents an interval of concentration and the granularity of the system is expressed in terms of the *step size* H (i.e. the length of the concentration interval).

Definition 1. *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 the species (i.e. H and N), $\mathcal{K}$ is the set of parameter definitions, $\mathcal{F}_R$ is the set of functional rates, Components is the set of definitions of species components, P is the model component describing the system.*

For discrete state space analysis the behaviour of the system is defined in terms of an operational semantics. A *Stochastic Labelled Transition System (SLTS)* is defined for a Bio-PEPA system. From this we can obtain a *Continuous Time Markov Chain (CTMC)*. Both the SLTS and the CTMC derived from Bio-PEPA are defined in terms of levels of concentration, and the generated Markov chain is called *CTMC with levels*. For a full description of the language semantics see [10].

The Bio-PEPA language is supported by software tools such as the Bio-PEPA Workbench [11], which automatically processes Bio-PEPA models and generates other representations in forms suitable for simulation and model-checking. For instance, the generated simulation model can be executed using the Dizzy stochastic simulator [12]. The representation which is used for discrete state space generation and analysis by numerical solution of the underlying CTMC is expressed in the reactive modules language supported by the PRISM model-checker [13]. In addition, the Bio-PEPA Workbench generates reward structures and common CSL [14] formulae used in model-checking.

3 The Gp130/JAK/STAT Signalling Pathway

The gp130/JAK/STAT signalling pathway is a well-studied biological system, of great clinical interest because of its key role in human fertility, neuronal repair and haematological development [15–17]. Much experimental data is available on this pathway, and a few mathematical and computational models [18–21] have been developed.

The signalling cascade in the gp130/JAK/STAT pathway is triggered by members of the family of IL (interleukin)-6-type cytokines binding to plasma membrane receptor complexes containing the common signal transducing receptor chain gp130 (glycoprotein 130). Among the targets of gp130 signal transduction, we consider the transcription

factors of the STAT (signal transducers and activators of transcription) family, in particular STAT3. A key feature of the pathway is the nuclear/cytoplasmic shuttling of STATs: upon activation, STATs can translocate into the nucleus and activate the transcription of downstream gene targets.

Different cytokines signal through the formation of different receptor complexes, all of them containing gp130 and another subunit. We focus here on two different cytokines: LIF (leukaemia inhibitory factor) and OSM (oncostatin M). LIF signals through an heterodimeric receptor complex gp130:LIFR. OSM exhibits the uncommon ability to signal through two different receptor complexes: the type I OSM receptor complex (gp130:LIFR), and the type II OSM receptor complex (gp130:OSMR).

Figure 1 is a graphical representation of the biochemical reactions occurring in the gp130/JAK/STAT pathway. In the inset the different types of receptor complexes are shown.

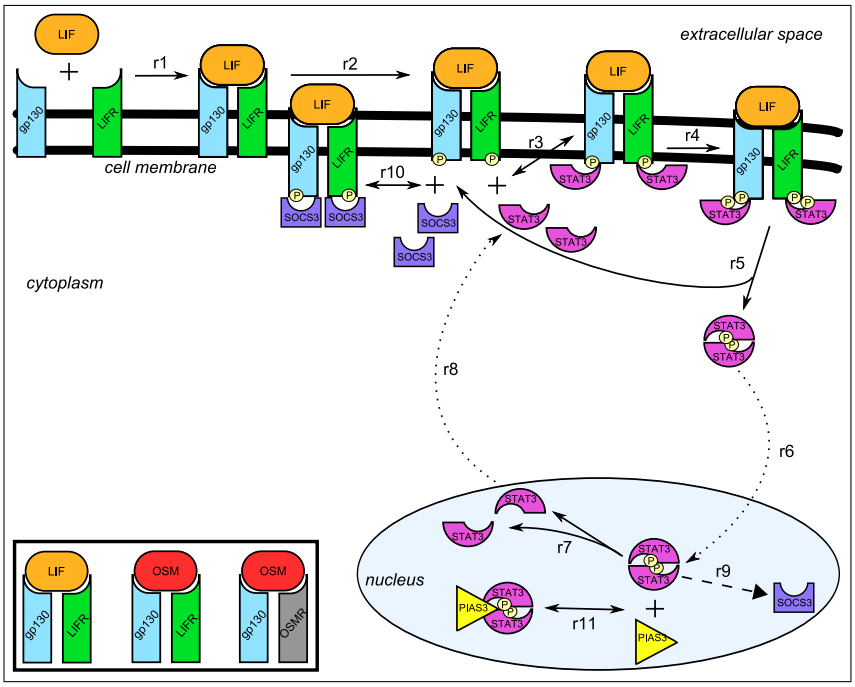

Fig. 1. Gp130/JAK/STAT pathway: graphical representation. Full arrows represent biochemical reactions, dotted arrows represent transports, dashed arrows represent syntheses.

The molecular species we consider in the model are: two ligands (LIF and OSM), three membrane-bound receptors (gp130, LIFR and OSMR), one effector (STAT3), and two inhibitors (SOCS3 and PIAS3). JAK kinase and TC-PTP phosphatase are implicitly modelled.

Four compartments are involved in the system: the exosol (the extracellular space, where the two ligands are located), the cell membrane (location of the receptors), the cytosol (initial location of STAT3), and the nucleus (in which STAT3 can translocate).

Receptors are activated by ligand bindings, and active receptors dimerise to form receptor complexes (gp130:LIFR or gp130:OSMR) (reaction r1 in Fig. 1). Once the receptor dimeric complex is formed, each receptor subunit (gp130, LIFR and OSMR) can undergo JAK-mediated phosphorylation (r2). STAT3 can bind on receptors' phosphorylated sites (r3), and the binding of STAT3 leads to its activation (phosphorylation) (r4).

Once phosphorylated, STAT3 dissociates from the receptor complex, and its phosphorylated site allows STAT3 to homodimerise (r5). When STAT3 is in dimeric form, it can translocate into the nucleus (r6) where it can carry out its specific functions (not modelled here): STAT3 binds to the DNA, thus activating the transcription of downstream gene targets. Nuclear STAT3 dimers are inactivated through TC-PTP -mediated dephosphorylation, which leads to the dimers' dissociation (r7) and to STAT3 export to the cytoplasm (r8), where STAT3 can undergo additional cycles of activation.

The two inhibition mechanisms considered are due to SOCS3 and PIAS3. SOCS3 is synthesised by STAT3 (r9) and it acts by competing with STAT3 in binding to receptors (r10). PIAS3 acts by binding to active nuclear STAT3 (r11).

4 The Bio-PEPA Model

A Bio-PEPA model of the gp130/JAK/STAT pathway has been developed. The full model can be downloaded from [22]. The model and the reaction rates are based on [21], though some differences are present due to the conceptual differences in the used modelling languages (see Sec. 8 for a discussion of such differences). All kinetic laws are assumed to be mass-action (i.e. depending on the amount of reactants and on given kinetic constants).

Each possible form of the molecular species is modelled as a distinct Bio-PEPA species component. For instance, STAT3 is modelled by four distinct species components representing, respectively, the cytoplasmic dephosphorylated monomeric form (*STAT3_c*), the cytoplasmic phosphorylated dimeric form (*STAT3-PD_c*), the nuclear phosphorylated dimeric form (*STAT3-PD_n*), and the nuclear dephosphorylated monomeric form (*STAT3_n*); further species components are defined for each state of each complex containing STAT3.

Reactions and biochemical modifications are represented by reactions over which the involved species components synchronise. For instance, the reaction representing r7 in Fig. 1 is modelled as the reaction $dephospho\_dedimer\_stat_{59}$, which decreases the amount of *STAT3-PD_n* and increases (with stoichiometry coefficient 2) the amount of *STAT3_n*.

As an example, the definitions of the species *STAT3-PD_n* and *STAT3_n* are reported (here we use the simplified syntax of the Bio-PEPA Workbench, in which the trailing S in prefix terms (α, κ) op S can be omitted).

$$STAT3\text{-}PD\_n ::= (reloc\_stat\_cn_{58}, 1)\uparrow \; + \; (synth\_socs_{61}, 1) \oplus + (unbind\_pias_{80}, 1)\uparrow$$
$$+ \; (dephospho\_dedimer\_stat_{59}, 1)\downarrow \; + \; (bind\_pias\_stat_{80}, 1)\downarrow$$
$$STAT3\_n ::= (dephospho\_dedimer\_stat_{59}, 2)\uparrow \; + \; (reloc\_stat\_nc_{60}, 1)\downarrow$$

For each of the involved reactions, a functional rate specifying its kinetic rate law is defined. The ones used in the species definitions for *STAT3-PD_n* and *STAT3_n* are defined as follows.

$$reloc\_stat\_cn_{58} = \left[\frac{0.693}{k_{58}} \cdot STAT3\text{-}PD\_c\right];$$
`//STAT3-PD_c relocation cytoplasm -> nucleus`

$$dephospho\_dedimer\_stat_{59} = [k_{59} \cdot STAT3\text{-}PD\_n];$$
`//STAT3-PD_n dephosphorylation & dedimerisation`

$$reloc\_stat\_nc_{60} = \left[\frac{0.693}{k_{60}} \cdot STAT3\_n\right];$$
`//STAT3-PD_n relocation nucleus -> cytoplasm`

$$synth\_socs_{61} = [k_{61} \cdot STAT3\text{-}PD\_n];$$
`//SOCS3 synthesis by STAT3-PD_n`

$$bind\_pias\_stat_{80} = \left[\frac{k_{80}}{nucleus \cdot N_A} \cdot PIAS3 \cdot STAT3\text{-}PD\_n\right];$$
`//PIAS3/STAT3-PD_n binding`

$$unbind\_pias\_stat_{80} = [k_{-80} \cdot PIAS3{:}STAT3\text{-}PD\_n];$$
`//PIAS3/STAT3-PD_n unbinding`

As mentioned above, the Bio-PEPA Workbench [11] allows us to automatically generate representations of the Bio-PEPA model for different analysis tools. In the following sections we show some of the analyses performed using these generated models. In particular, we consider the PRISM [23, 13] and Dizzy [12, 24] models. We use the PRISM model-checker to verify that some desired properties of the system are satisfied, and the Dizzy simulation tool to perform time-series analysis via stochastic simulation.

5 Model-checking Based Qualitative Analysis

As a first step in the analysis of the model we use the PRISM model-checker [23, 13] to verify a number of qualitative properties of the system. Such properties are intended to be *consistency checks* on the model and they allow us to check for the presence of possible human errors in the modelling process. This kind of checks is particularly useful when modelling complex systems such as the pathway we consider here since, due to the size of the models, trivial typing errors are likely to occur and may be hard to identify.

5.1 PRISM Modelling and Specification Language

PRISM [23, 13] is a probabilistic model-checker, which can be used to verify properties of CTMCs. Models are described using the state-based PRISM language, and it is possible to specify quantitative properties of the system using a property specification language which includes *CSL* (Continuous Stochastic Logic) [25, 26]. The PRISM language is composed of *modules* and *variables*. A model is composed of a number of interacting modules and each module contains a number of local variables, whose values constitute the state of the module. The global state of the model is determined by the local state of all modules. The behaviour of the modules is given by a set of guarded commands, each describing a transition which is enabled when the guard is true. A command includes an update which gives new values to the variables.

PRISM properties are made up of *state properties* ϕ and *path properties* ψ. The syntax of PRISM properties is given by the following grammar.

$$\phi ::= \mathbf{true} \mid \mathbf{false} \mid expr \mid \phi \wedge \phi \mid \phi \vee \phi \mid \neg\phi \mid \phi \Rightarrow \phi \mid \mathcal{P}_{\bowtie p}[\psi] \mid \mathcal{P}_{=?}[\psi] \mid \mathcal{S}_{\bowtie p}[\phi] \mid \mathcal{S}_{=?}[\phi]$$

$$\psi ::= \mathbf{X}\,\phi \mid \phi\,\mathbf{U}^{I}\,\phi \mid \phi\,\mathbf{U}\,\phi \mid \mathbf{F}^{I}\,\phi \mid \mathbf{F}\,\phi \mid \mathbf{G}^{I}\,\phi \mid \mathbf{G}\,\phi$$

Here *expr* is a boolean expression (containing literal values, identifiers and the standard arithmetic and relational operators), $\bowtie \in \{<, \leq, \geq, >\}$ is a relational parameter, $p \in [0, 1]$ is a probability, and I is an interval of $\mathbb{R}^{+}$.

The operators $\mathcal{P}_{\bowtie p}[\psi]$ and $\mathcal{P}_{=?}[\psi]$ are used to express transient properties (i.e. which depend on time) whereas the operators $\mathcal{S}_{\bowtie p}[\phi]$ and $\mathcal{S}_{=?}[\phi]$ are used to express steady state properties (i.e. which hold in the long run). The result of the verification of formulae $\mathcal{P}_{\bowtie p}[\psi]$ (resp. $\mathcal{S}_{\bowtie p}[\phi]$) is one of the boolean values **true** or **false** depending on whether ψ (resp. ϕ) is satisfied. The result of the verification of formulae $\mathcal{P}_{=?}[\psi]$ (resp. $\mathcal{S}_{=?}[\phi]$) is the expected probability with which ψ (resp. ϕ) is satisfied.

The operators **X**, **U**, **F**, and **G** are used to express *neXt*, *Until*, *Finally*, and *Globally* properties, respectively. Time-bounded formulae are indexed by an interval I.

The PRISM language supports the specification and analysis of reward-based properties. Reward structures allow us to associate real values with certain states or transitions of the model. Such values, which can be thought of as "costs" of the specified states/transitions, are taken into account during the solution of the CTMC. In this way it is possible to reason about various quantitative measures such as "expected number of instances of processes", "expected number of occurrences of reactions", "expected time until a condition is satisfied", etc. The PRISM reward language supports the expression of both instantaneous and cumulative rewards.

5.2 Model-checking the Bio-PEPA Model with PRISM

In the PRISM models generated by the Bio-PEPA Workbench, one module is defined for each species, and the module local variables are used to record the current quantity of each species. The transitions correspond to the activities of the Bio-PEPA model and the updates take the stoichiometry into account. Transition rates are specified in an auxiliary module which defines the functional rates corresponding to all the reactions.

Moreover, lower and upper bounds must be defined for each variable (i.e. for the amount of each species). The step size H in the Bio-PEPA model allows us to consider different PRISM models with different *granularity*, leading to systems with different numbers of levels.

As an example, we provide the PRISM definitions relative to the species *STAT3-PD_n* and *STAT3_n*, which are obtained from the corresponding Bio-PEPA species definitions reported in Sec. 4.

First, the lower and upper levels for both species are computed from the defined step size H and the given bounds on species amounts.

$$\begin{aligned} MIN\_STAT3\text{-}PD\_n &= MIN\_STAT3\_n = 0 \\ MAX\_STAT3\text{-}PD\_n &= MAX\_STAT3\_n = 1500 \end{aligned}$$

$$\begin{aligned} N_L\_STAT3\text{-}PD\_n &= \left\lfloor \tfrac{MIN\_STAT3\text{-}PD\_n}{H} \right\rfloor & N_U\_STAT3\text{-}PD\_n &= \left\lfloor \tfrac{MAX\_STAT3\text{-}PD\_n}{H} \right\rfloor \\ N_L\_STAT3\_n &= \left\lfloor \tfrac{MIN\_STAT3\_n}{H} \right\rfloor & N_U\_STAT3\_n &= \left\lfloor \tfrac{MAX\_STAT3\_n}{H} \right\rfloor \end{aligned}$$

The specifications of the behaviour of *STAT3-PD_n* and *STAT3_n* are given by the two following modules. The third module contains the definition of the functional rates for all reactions.

module $STAT3\text{-}PD\_n$

$STAT3\text{-}PD\_n : [N_L\_STAT3\text{-}PD\_n .. N_U\_STAT3\text{-}PD\_n]$ **init** 0;

$[reloc\_stat\_cn_{58}]\,(STAT3\text{-}PD\_n + 1 \leq N_U\_STAT3\text{-}PD\_n) \rightarrow 1 : (STAT3\text{-}PD\_n' = STAT3\text{-}PD\_n + 1);$

$[synth\_socs_{61}]\,(STAT3\text{-}PD\_n + 0 \leq N_U\_STAT3\text{-}PD\_n) \rightarrow 1 : (STAT3\text{-}PD\_n' = STAT3\text{-}PD\_n + 0);$

$[dephospho\_dedimer\_stat_{59}]\,(STAT3\text{-}PD\_n \geq 1 + N_L\_STAT3\text{-}PD\_n) \rightarrow 1 : (STAT3\text{-}PD\_n' = STAT3\text{-}PD\_n - 1);$

$[bind\_pias\_stat_{80}]\,(STAT3\text{-}PD\_n \geq 1 + N_L\_STAT3\text{-}PD\_n) \rightarrow 1 : (STAT3\text{-}PD\_n' = STAT3\text{-}PD\_n - 1);$

$[unbind\_pias\_stat_{80}]\,(STAT3\text{-}PD\_n + 1 \leq N_U\_STAT3\text{-}PD\_n) \rightarrow 1 : (STAT3\text{-}PD\_n' = STAT3\text{-}PD\_n + 1);$

endmodule

module $STAT3\_n$

$STAT3\_n : [N_L\_STAT3\_n .. N_U\_STAT3\_n]$ **init** 0;

$[dephospho\_dedimer\_stat_{59}]\,(STAT3\_n + 2 \leq N_U\_STAT3\_n) \rightarrow 1 : (STAT3\_n' = STAT3\_n + 2);$

$[reloc\_stat\_nc_{60}]\,(STAT3\_n \geq 1 + N_L\_STAT3\_n) \rightarrow 1 : (STAT3\_n' = STAT3\_n - 1);$

endmodule

module *Rates*

$$[reloc\_stat\_cn_{58}]\ \left(\frac{\frac{0.693}{k_{58}}\cdot STAT3\text{-}PD\_c\cdot H}{H} > 0\right) \rightarrow \left(\frac{\frac{0.693}{k_{58}}\cdot STAT3\text{-}PD\_c\cdot H}{H}\right) :\ \textbf{true};$$

$$[dephospho\_dedimer\_stat_{59}]\ \left(\frac{k_{59}\cdot STAT3\text{-}PD\_n\cdot H}{H} > 0\right) \rightarrow \left(\frac{k_{59}\cdot STAT3\text{-}PD\_n\cdot H}{H}\right) :\ \textbf{true};$$

$$[reloc\_stat\_nc_{60}]\ \left(\frac{\frac{0.693}{k_{60}}\cdot STAT3\_n\cdot H}{H} > 0\right) \rightarrow \left(\frac{\frac{0.693}{k_{60}}\cdot STAT3\_n\cdot H}{H}\right) :\ \textbf{true};$$

$$[synth\_socs_{61}]\ \left(\frac{k_{61}\cdot STAT3\text{-}PD\_n\cdot H}{H} > 0\right) \rightarrow \left(\frac{k_{61}\cdot STAT3\text{-}PD\_n\cdot H}{H}\right) :\ \textbf{true};$$

$$[bind\_pias\_stat_{80}]\ \left(\frac{\frac{k_{80}}{nucleus\cdot N_A}\cdot PIAS3\cdot H\cdot STAT3\text{-}PD\_n\cdot H}{H} > 0\right) \rightarrow \left(\frac{\frac{k_{80}}{nucleus\cdot N_A}\cdot PIAS3\cdot H\cdot STAT3\text{-}PD\_n\cdot H}{H}\right) :\ \textbf{true};$$

$$[unbind\_pias\_stat_{80}]\ \left(\frac{k_{-80}\cdot PIAS3\text{:}STAT3\text{-}PD\_n\cdot H}{H} > 0\right) \rightarrow \left(\frac{k_{-80}\cdot PIAS3\text{:}STAT3\text{-}PD\_n\cdot H}{H}\right) :\ \textbf{true};$$

endmodule

The PRISM model generated from the Bio-PEPA model of the gp130/JAK/STAT pathway has 63 species and 118 reactions. Because of the well-known state space explosion problem of model-checking, even if we consider only a few levels for each species, the state space for this model is so huge that it makes the numerical solution of the CTMC nearly unmanageable. To overcome this problem, we consider a subdivision of the pathway into two distinct sub-models in such a way that the analysis of the individual sub-models becomes more feasible.

In order to find an appropriate modularisation, we adopt the approach proposed in [27, 28], based on the identification of sub-systems with no retroactivity. For the considered model of the gp130/JAK/STAT pathway, two modules with low coupling can be easily identified.

In the first sub-model, which refers to the bindings of ligands to receptors and the activation of the receptor dimers, we consider all the distinct combinations of ligand/receptor complexes, and we describe in detail the formation of all possible types of active receptor dimers, considering the fact that different ligand-receptor pairs have different binding affinities.

In the second sub-model, which refers to the downstream signalling pathway, we instead consider as a starting point a single "generic" type of active receptor dimer (referred to as *rcpt-DP*), and we focus on the reactions involving the activation of STAT3 and its cytoplasmic/nuclear shuttling.

These two sub-models refer to sub-systems of the gp130/JAK/STAT pathway which act in a rather sequential way and, as a consequence, it is reasonable to assume that, for the downstream STAT3 signalling to occur, the receptor-complexes must have been already activated. The initial number of active receptor dimers in the second sub-model is defined as the sum of the steady-state quantities of all the active receptor dimers in the first sub-model. This assumption is justified by the fact that the activation of the receptors is fast compared with the following reactions, and therefore the amount of initially inactive receptors is negligible when considering the downstream pathway.

As discussed in [27, 28], the absence of retroactivity ensures that the modularisation has no significant effect on the overall behaviour of the system. This, together with the fact that we use the output of the first sub-model as input of the second sub-model, ensures that the structural qualitative properties verified for the individual sub-models in the rest of this section also hold for the full model. Particular care should be taken when verifying quantitative temporal properties over sub-models. Here we only consider *semi-quantitative* analysis (Sec. 7) as we are interested in relative rather than absolute values. Therefore, in this particular case, the absence of retroactivity ensures the validity, in the full model, of the analysis results obtained in the sub-models. In general, however, the actual reaction rates in the composite model (and therefore the analysis results) might be different from the ones in the sub-models, and more advanced approaches for modularisation should be applied.

In the rest of this section we use $H = 200$ as the step size for the ligands-receptors sub-model, and $H = 300$ for the downstream sub-model. See Sec. 7 for a discussion of the choice of step size values.

Deadlock Detection. *Deadlock states* are the ones in which no transition is enabled. In some cases the presence of deadlock states is (correctly) due to the presence of irreversible reactions which lead to the transformation of all reactants into non-reactive proteins. In other cases deadlocks could be due to the scarcity of one of the reactants of a multimolecular reaction; in our model, for instance, all receptors are consumed (i.e. transformed into different forms, such as dimers) while still ligands are available. In other cases deadlocks could be caused by modelling errors.

PRISM automatically detects deadlock states when building the state space of models, and this feature can be considered the first step in the identification of potential modelling errors.

For instance, in the ligands-receptors sub-model, any state in which ligands are present while all *gp130* receptors have been consumed is a deadlock. This suggests that *gp130* is the bottleneck of the system.

Species Invariants. One simple and yet interesting property that can be verified is the presence of *invariants* in the amount of the involved proteins.

Species invariants are commonly present in biochemical systems because of the existence of basic constraints such as the law of conservation of mass, which states that the amount (i.e. mass) of reactants consumed by a reaction must be equal to the amount of products of the reaction.

For instance, given the conservation of mass and the absence of synthesis and degradation reactions, we expect that the sum of the amounts of *LIFR* receptor present in its various possible forms (free, as *gp130:LIF:LIFR* complex and as *gp130:OSM:LIFR* complex, with one or both of its subunits phosphorylated) is constant (and equal to the *LIFR* initial amount).

The satisfaction of the following properties confirms the existence of the expected invariants on the total amount of ligands and receptors (as an example, we report the ones for *LIF* and *LIFR*).

$\mathcal{P}_{\geq 1}$[**G** (*LIF* + *gp130:LIF:LIFR* + *gp130-P:LIF:LIFR* + *gp130:LIF:LIFR-P* + *gp130-P:LIF:LIFR-P* = N_U*_LIF*)] → **true**

$\mathcal{P}_{\geq 1}$[**G** (*LIFR* + *gp130:LIF:LIFR* + *gp130:OSM:LIFR* + *gp130-P:LIF:LIFR* + *gp130:LIF:LIFR-P* + *gp130-P:LIF:LIFR-P* + *gp130-P:OSM:LIFR* + *gp130:OSM:LIFR-P* + *gp130-P:OSM:LIFR-P* = N_U*_LIFR*)] → **true**

Here, and in the rest of the section, the notation $\mathcal{P}_{\bowtie p}[\psi] \rightarrow$ **true** (resp. **false**) means that ψ is satisfied (resp. is not satisfied), while the notation $\mathcal{P}_{=?}[\psi] \rightarrow p$ (with $p \in \mathbb{R}$) means that the result of ψ is the probability p.

Reachability Analysis. *Reachability* properties allow us to verify whether a given state is eventually reached. States of interest can be, for instance, the ones in which some species reaches a threshold or is totally consumed, or when the amounts of two species coincide.

We consider here the states in which a certain number of receptors are phosphorylated, and the ones in which a certain amount of active nuclear STAT3 (*STAT3-PD_n*) is present.

We consider first the ligands-receptors sub-model. The satisfaction of the first of the following properties guarantees that a state in which one fourth of the total amount of available receptors is phosphorylated is always reached at some time point. On the contrary, the second property, which is not satisfied, proves that we do not necessarily reach a state with one third of receptors phosphorylated.

$\mathcal{P}_{\geq 1}$[**F** (*gp130-P:LIF:LIFR-P* + *gp130-P:OSM:LIFR-P* + *gp130-P:OSM:OSMR-P* > (N_U*_OSMR* + N_U*_LIFR* + N_U*_gp*130) / 4)] → **true**

$\mathcal{P}_{\geq 1}$[**F** (*gp130-P:LIF:LIFR-P* + *gp130-P:OSM:LIFR-P* + *gp130-P:OSM:OSMR-P* > (N_U*_OSMR* + N_U*_LIFR* + N_U*_gp*130) / 3)] → **false**

The next property, instead, guarantees that in general we could reach a system where no *gp130:OSMR* receptor complex is activated.

$\mathcal{P}_{\geq 1}$[**F** (*gp130-P:OSM:OSMR-P* > 0)] → **false**

Regarding the downstream sub-model, we check for the following properties, which guarantee that, at some time point, at least half the initial amount of STAT3 has been transported into the nucleus and activated, but not all of it.

$\mathcal{P}_{\geq 1}$[**F** (*STAT3-PD_n* > N_U*_STAT3_c* / 2)] → **true**

$\mathcal{P}_{\geq 1}$[**F** (*STAT3-PD_n* > N_U*_STAT3_c*)] → **false**

Reversibility. A system is called *reversible* if the initial state is reachable from any other state (i.e. the system is able to self-reinitialise). More generally, a state is called *reversible* if it can be reached again at some later time point.

The following property, if satisfied, guarantees the reversibility of the system: it states that it is always possible to return to the initial state (in the PRISM language "**init**" is a predefined formula which completely specifies the initial state).

$$\mathcal{P}_{=?}[\mathbf{G}\ (\text{“}\mathbf{init}\text{”} \Rightarrow \mathcal{P}_{\geq 1}[\mathbf{X}\ (!\text{“}\mathbf{init}\text{”} \Rightarrow \mathcal{P}_{\geq 1}[\mathbf{F}\ (\text{“}\mathbf{init}\text{”})])])]$$

For the ligands-receptors sub-system the result of this property is 0, since we have considered bindings to be irreversible and, therefore, the system cannot return to the initial state in which all receptors and ligands are free.

The downstream sub-system, instead, is reversible (the result of the property is 1), thanks to the cytoplasmic/nuclear STAT3 shuttling, which enables the system to return to the initial state in which cytoplasmic STAT3 molecules are not phosphorylated and not bound to receptor dimers.

Liveness. The notion of *liveness* of a reaction in a given state refers to the possibility of it occurring in such a state. In particular, it is interesting to know which reactions are live in the initial state.

Since PRISM properties are state-based, it is not possible to explicitly check for the occurrence of a given reaction. However, knowing how each model component is affected by the occurrence of a given reaction, we can verify this kind of property by checking for the expected variations in the involved components.

We are interested, for instance, in verifying that in the initial state the binding reactions between ligands and receptors can occur, leading to the three possible types of ligand/receptor dimers (*gp130:LIF:LIFR*, *gp130:OSM:LIFR*, and *gp130:OSM:OSMR*).

The following three properties are satisfied, confirming that the three known types of complexes can be formed.

$$\mathcal{P}_{\geq 1}[\mathbf{G}\ (\text{“}\mathbf{init}\text{”} \Rightarrow \mathcal{P}_{>0}[\mathbf{X}\ (gp130 = N_U\_gp130 - 1\ \&\ LIF = N_U\_LIF - 1\ \&\ LIFR = N_U\_LIFR - 1)])]\ \rightarrow\ \mathbf{true}$$

$$\mathcal{P}_{\geq 1}[\mathbf{G}\ (\text{“}\mathbf{init}\text{”} \Rightarrow \mathcal{P}_{>0}[\mathbf{X}\ (gp130 = N_U\_gp130 - 1\ \&\ OSM = N_U\_OSM - 1\ \&\ LIFR = N_U\_LIFR - 1)])]\ \rightarrow\ \mathbf{true}$$

$$\mathcal{P}_{\geq 1}[\mathbf{G}\ (\text{“}\mathbf{init}\text{”} \Rightarrow \mathcal{P}_{>0}[\mathbf{X}\ (gp130 = N_U\_gp130 - 1\ \&\ OSM = N_U\_OSM - 1\ \&\ OSMR = N_U\_OSMR - 1)])]\ \rightarrow\ \mathbf{true}$$

The following property, instead, is not satisfied: it states, as desired, that *LIF* cannot bind to receptors to form *gp130:OSMR* dimers.

$$\mathcal{P}_{\geq 1}[\mathbf{G}\ (\text{“}\mathbf{init}\text{”} \Rightarrow \mathcal{P}_{>0}[\mathbf{X}\ (gp130 = N_U\_gp130 - 1\ \&\ LIF = N_U\_LIF - 1\ \&\ OSMR = N_U\_OSMR - 1)])]\ \rightarrow\ \mathbf{false}$$

Causality Analysis. *Causality relations* between given reactions can be expressed and verified by properties which relate the order of "appearance" of relevant molecules. This kind of property can be used, for instance, to verify the order in which intermediate products are formed within a cascade of events.

A form of causality relation can be expressed by using the sequence and consequence relations defined in [29]: specifically, while *sequence* formulae describe ordering relations between events (e.g. "in order to reach a given state, we must first reach another one"), *consequence* formulae describe causal relations (e.g. "if a given state occurs, it is necessarily followed by a second one").

For example, the ordering and causality relations between *STAT3* phosphorylation, homodimerisation and relocation into the nucleus can be verified by the following pairs of properties (assuming at system initialisation all *STAT3* is present in cytoplasmic monomeric form (*STAT3-P_c*).

When the result of the first property is 0, such a property states that it is not possible for a *STAT3-PD_c* molecule to be present if in all previous states we had no *rcpt-DP:STAT3-DP1* (a complex formed by a receptor dimer and a STAT3 molecule). Similarly, the following property (when it evaluates to 0) states that *STAT3-PD_c* must be produced before *STAT3-PD_n* appears.

$$\mathcal{P}_{=?}[(\textit{rcpt-DP:STAT3-DP1} = 0)\ \mathbf{U}\ STAT3\text{-}PD\_c > 0] \rightarrow\ 0$$

$$\mathcal{P}_{=?}[(STAT3\text{-}PD\_c = 0)\ \mathbf{U}\ STAT3\text{-}PD\_n > 0] \rightarrow\ 0$$

The following two properties complement the previous two, stating that if at least one complex *rcpt-DP:STAT3-DP1* is formed, then at least one *STAT3-PD_c* molecule will necessarily be formed.

$$\mathcal{P}_{=?}[\mathbf{G}\ (\textit{rcpt-DP:STAT3-DP1} > 0 \Rightarrow \mathcal{P}_{\geq 1}[\mathbf{F}\ (STAT3\text{-}PD\_c > 0)])] \rightarrow\ 1$$

$$\mathcal{P}_{=?}[\mathbf{G}\ (STAT3\text{-}PD\_c > 0 \Rightarrow \mathcal{P}_{\geq 1}[\mathbf{F}\ (STAT3\text{-}PD\_n > 0)])] \rightarrow\ 1$$

As another example, the following two properties verify that the transport of phosphorylated *STAT3* dimers can only occur from the cytoplasm to the nucleus, but not vice versa. The result of the first property is 0 (i.e. transport of *STAT3-PD* can occur from cytoplasm to nucleus), while the result of the second property is 1 (i.e. transport of *STAT3-PD* cannot occur from nucleus to cytoplasm) for all reachable values of i, j.

$$\mathcal{P}_{=?}[\mathbf{F}\ (STAT3\text{-}PD\_c = i\ \&\ STAT3\text{-}PD\_n = j\ \&\ \mathcal{P}_{\leq 0}[\mathbf{X}\ (STAT3\text{-}PD\_c = i - 1\ \&\ STAT3\text{-}PD\_n = j + 1)])] \rightarrow 0$$

$$\mathcal{P}_{=?}[\mathbf{F}\ (STAT3\text{-}PD\_c = i\ \&\ STAT3\text{-}PD\_n = j\ \&\ \mathcal{P}_{\leq 0}[\mathbf{X}\ (STAT3\text{-}PD\_c = i + 1\ \&\ STAT3\text{-}PD\_n = j - 1)])] \rightarrow 1$$

Conversely, the transport of dephosphorylated *STAT3* monomers can only occur from the nucleus to the cytoplasm.

$$\mathcal{P}_{=?}[\mathbf{F}\ (STAT3\_c = i \,\&\, STAT3\_n = j \,\&\, \mathcal{P}_{\leq 0}[\mathbf{X}\ (STAT3\_c = i-1 \,\&\, STAT3\_n = j+1)])] \rightarrow 1$$

$$\mathcal{P}_{=?}[\mathbf{F}\ (STAT3\_c = i \,\&\, STAT3\_n = j \,\&\, \mathcal{P}_{\leq 0}[\mathbf{X}\ (STAT3\_c = i+1 \,\&\, STAT3\_n = j-1)])] \rightarrow 0$$

6 Simulation Based Time-series Analysis

In the previous section we have used model-checking in order to check for a number of simple formulae which guarantee us that some key properties of the gp130/JAK/STAT model are satisfied. This analysis allows us to be more confident about the absence of modelling errors.

Now we progress our analysis of the model by means of stochastic simulation. We report here some results obtained by simulating the full model (comprising both the ligands-receptors and the downstream sub-systems) using the Gibson-Bruck [30] stochastic simulation engine implemented in Dizzy [12, 24].

Figure 2 shows the time-series evolution produced by the model (Fig. 2(a)) versus the ones in which each of the three inhibitors has been removed (Fig. 2(b)–(d)). Each plot refers to average values computed over 1000 simulation runs, and the amounts of the four different forms of STAT3 are shown (cytoplasmic and nuclear dephosphorylated monomers, and cytoplasmic and nuclear phosphorylated dimers).

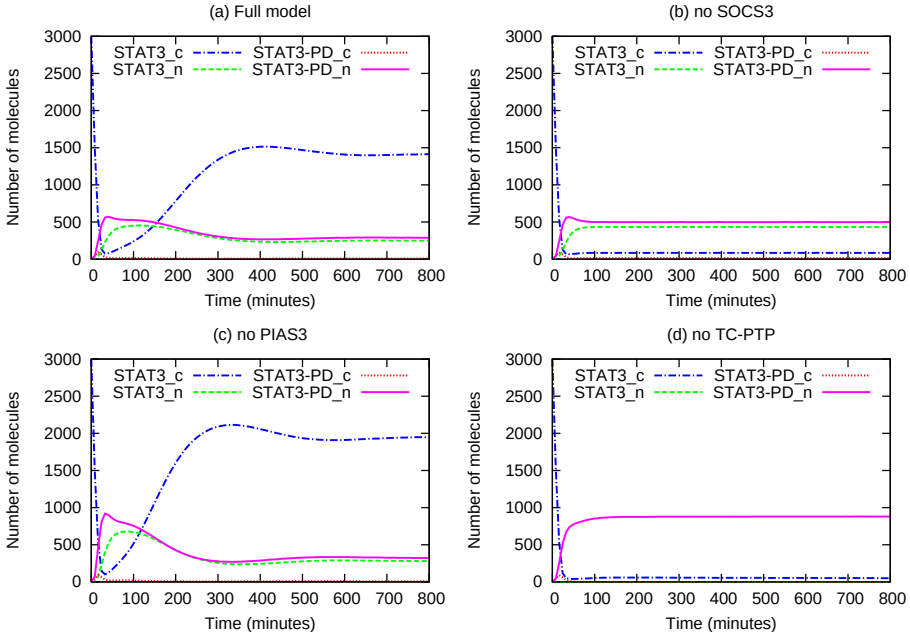

Fig. 2. Simulation results: full model vs. no inhibitors.

In all the performed simulations, at system initialisation STAT3 is only present in cytoplasmic monomeric form. As shown in Fig. 2(a), as time passes, STAT3 is phosphorylated, dimerised, and transported into the nucleus, until the systems reaches a state in which the inhibition of nuclear STAT3 by dephosphorylation and the nuclear/cytoplasmic shuttling lead nuclear and cytoplasmic STAT3 to be in equilibrium.

When the amount of nuclear STAT3 increases significantly, the inhibitory role of SOCS3 (which is under transcription control of STAT3) comes into play (Fig. 2(b)). SOCS3 is responsible for signal attenuation and, hence, after reaching a peak, nuclear STAT3 decreases.

PIAS3 slows down the production of active nuclear STAT3 by binding to it (Fig. 2(c)). Therefore, if PIAS3 is present, part of nuclear STAT3 is bound to it, while, if PIAS3 is knocked down, the amount of available STAT3 increases.

A third inhibitor, TC-PTP, allows nuclear STAT3 to translocate back into the cytoplasm, by dephosphorylating it (Fig. 2(d)). If TC-PTP is present, STAT3 nuclear/cytoplasmic shuttling occurs; instead, if TC-PTP is knocked out (i.e. if nuclear STAT3 is not dephosphorylated), STAT3 accumulates in the nucleus, whilst cytoplasmic STAT3 molecules quickly disappear.

7 Semi-quantitative Analysis of the CTMC with Levels

In Sec. 5 we have shown how model-checking can be used in order to discover modelling errors by checking for some basic properties which guarantee the model to behave as expected. In this section, instead, we use model-checking also for quantitative analysis, with the purpose of completing the simulation-based analysis in order to provide additional insight on the behaviour of the gp130/JAK/STAT pathway.

The main advantage of model-checking with respect to stochastic simulation is the fact that model-checking is exhaustive: it explores all the possible behaviours of the model and it does not require us to compute an average behaviour of a number of stochastic simulation runs.

As mentioned before, the main disadvantage of model-checking is the state space explosion problem, which implies that we cannot deal with too many levels for the model components without inducing an intractable model.

In has been shown (see [10]) that, as the number of levels increases, the behaviour of the CTMC with levels tends to the behaviour of ODEs (when the number of molecules is large enough to average out the randomness of the system); this result guarantees the theoretical correctness of the approach. However, if the number of levels is too small, the error introduced by the discretisation becomes significant and the numerical solution of the generated CTMC fails to reproduce the correct behaviour.

The number of levels for model components is related to the step size H and to the upper N_U and lower N_L bounds for each species. The step size H represents the granularity of the system, and it directly affects the accuracy of the results; the upper and lower bounds are also relevant to the accuracy, since imposing bounds on the numbers of molecules causes a state space truncation which might potentially have impact on the behaviour of the system.

Therefore, when performing CTMC analysis of Bio-PEPA models, the choice of the step size and of the upper and lower bounds is essential: they must be carefully selected so that the number of levels to be used for the model components is a suitable trade-off between accuracy and efficiency.

In the following sections we report some of the results obtained by using the PRISM model-checker to perform quantitative analysis. First we consider reward-based properties which allow us to observe the time-series for some of the species of the system (for comparison with the stochastic simulation), and we discuss the error introduced by discretising and bounding the model; afterwards, we define further properties in order to compute additional (semi-)quantitative measures.

Time-series Analysis Using State Rewards. A reward structure is automatically defined by the Bio-PEPA Workbench for each PRISM component, and it can be referred to either by the component name or by an integer value (implicitly assigned to reward structures based on the order in which they are defined). These reward structures associate an instantaneous reward equal to the current amount of the corresponding molecular species with each state. The evaluation of these reward-based properties corresponds to computing an average behaviour for the species at given time points.

As an example, the following reward is used to observe the time evolution of the receptor dimer *gp130:LIF:LIFR*.

rewards "*gp130:LIF:LIFR*"
  true : *gp130:LIF:LIFR* · H;
endrewards

Figure 3 reports the results obtained by verifying on the ligands-receptors subsystem the reward-based property

$$\mathcal{R}^{i}_{=?}[I = T]$$

for time points $T \leq 30$ minutes, where i is an integer variable used to index the reward structure of interest.

Figure 4, instead, reports the results obtained by verifying the same reward-based property for time points $T \leq 800$ minutes on the downstream sub-system. In this figure, we also report the standard deviation of the number of molecules, which is computed by exploiting reward structures associating the square of the number of molecules of each species with each state: the standard deviation is calculated as the square root of the variance $E(Y)^2 - E(Y^2)$, where Y is the random variable representing a species in the network, whereas $E(Y)$ and $E(Y^2)$ indicate the expected values for the amount of the species Y and for its square value.

Figures 3 and 4 have been obtained by analysing the sub-models with step sizes $H = 200$ and $H = 300$ respectively. In the next section we discuss the considerations which lead us to the choice of such values.

Three kinds of approximation errors could have been introduced by our analysis of the CTMC with levels due to, respectively, the discretisation of the amounts (H), their bounding (N_L and N_U), and the subdivision into modules.

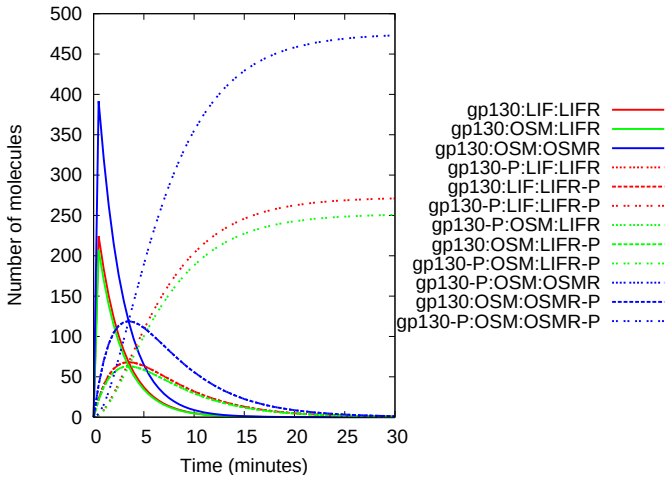

Fig. 3. Time-series by model-checking: ligands-receptors sub-model.

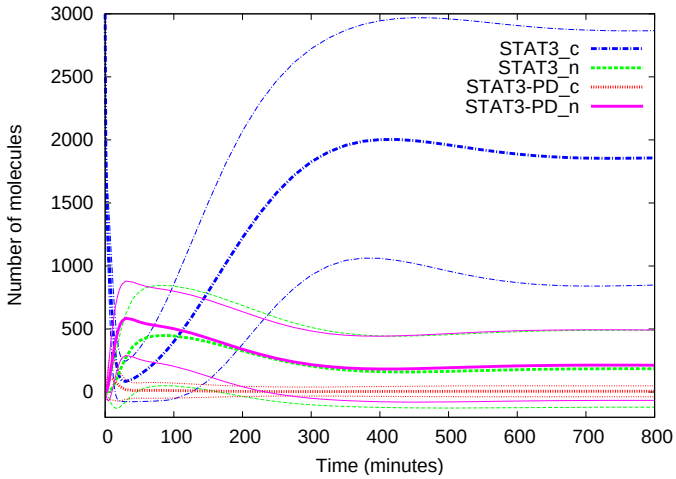

Fig. 4. Time-series by model-checking: downstream sub-model. Thick lines represent the expected numbers of molecules; thin lines represent their standard deviation.

In the next section we discuss the effect of varying the step size H on the behaviour of the system. Instead, we do not report results concerning the variation of the bounds N_L and N_U since, in this particular system, increasing the bounds does not have a significant effect: the reason for this is that no synthesis and degradation reactions are defined (with the single exception of SOCS3) and, as a consequence, the amount of most molecular species is clearly bounded by the amounts of the molecules present at system initialisation.

The choice of how to modularise the system has been carried out in order to minimise the interaction between the two modules. However, the modularisation has certainly an impact on the quantitative behaviour. In the whole system, for instance, STAT3 and SOCS3 molecules can bind to receptor dimers as soon as they start being phosphorylated; in the downstream sub-model, instead, we had to fix an initial amount of phosphorylated receptor dimers.

Despite these possible sources of approximation, comparing Fig. 4 and Fig. 2, we notice that the results obtained by analysing the downstream sub-model using PRISM instantaneous rewards do not differ significantly from the behaviour observed by averaging the results obtained by 1000 stochastic simulation runs of the whole model. Both the time-scale and the relative amounts of molecules are the same in both figures, and the only significant difference regarding the absolute amounts is the amount of cytoplasmic monomeric STAT3, which is higher in Fig. 4. We can also observe that the standard deviation reported in Fig. 4 is quite high, due to the stochastic noise which has been introduced by using a small number of levels.

Experimenting with Step Sizes. As previously stated, the choice of the step size has a great impact on both accuracy and performance of the analysis: the smaller the step size is, the larger the CTMC state space and, hence, the smaller the discretisation error introduced, but also the longer the time needed for solving the CTMC.

Before choosing the values to be used for the step size H in the analysis of the models, we have performed a number of experiments varying H in order to find values representing a good trade-off between accuracy and performance of the analysis. In Fig. 5 and Fig. 6 we report some results which show how changing the step size affects the behaviour of the system (in ligands-receptors and downstream sub-systems, respectively).

In Fig. 5, we compare the results obtained by using six different values for H $(1000, 500, 300, 250, 200, 150)$ in the analysis of the ligands-receptors sub-model, and we can observe that H in this case does not have a big impact on the results.

The first notable difference is that in Fig. 5(a) the amounts of gp130:LIF:LIFR and gp130:OSM:LIFR are equal: as expected, with $H = 1000$ (i.e. one single level for each ligand and receptor) we are not able to observe the fact that $LIFR$ has an higher binding affinity with LIF than with OSM.

The other interesting thing is that, contrary to what we expected, there is no noticeable increase of accuracy when decreasing H. Instead, after observing the similarities between Fig. 5(b), (d) and (e), and between Fig. 5(c) and (f), respectively, we drew the conclusion that the first group is the "correct" one; the reason is the rounding error introduced when computing the number of levels starting from the initial amounts (re-

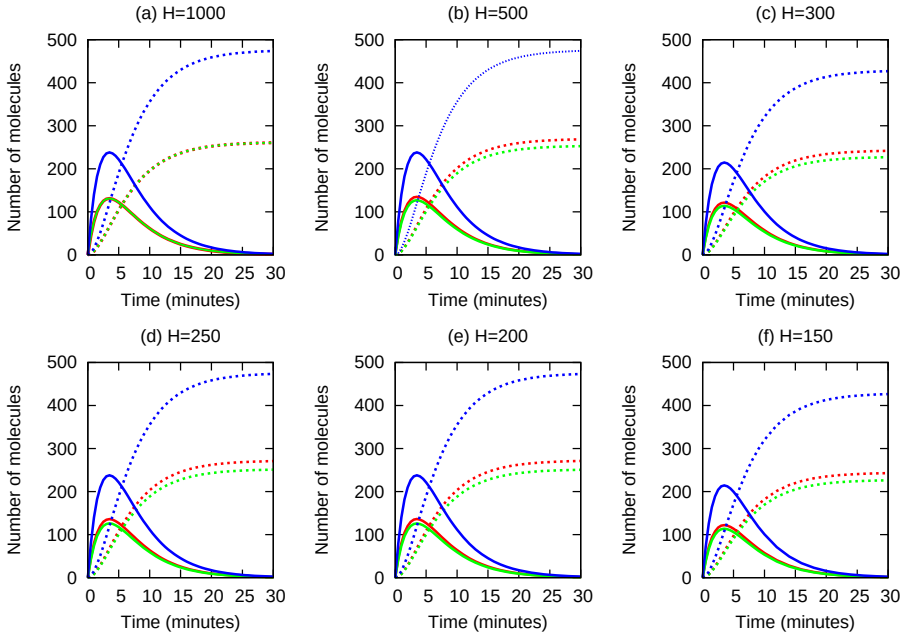

Fig. 5. Time-series by model-checking: ligands-receptors sub-model. The three types of receptor complexes are shown, gp130:LIF:LIFR (red), gp130:OSM:LIFR (green), and gp130:OSM:OSMR (blue), in the stage when one (full line) or both (dashed line) receptors are phosphorylated.

member that $N_L = \lfloor MIN/H \rfloor$ and $N_U = \lfloor MAX/H \rfloor$): when a small numbers of levels is used, this rounding error happens to be more significant than H itself.

In Fig. 6, we compare the results obtained by using five different values for H $(1000, 500, 400, 300, 270)$ in the analysis of the downstream sub-model; the value obtained by stochastic simulation is also shown.

As for the ligands-receptors sub-model, also for this sub-model we notice that when using $H = 1000$ we obtain a totally wrong behaviour, and we observe a general increase in accuracy when increasing the number of levels. For the smallest values of H, the relative values and the trends for the considered species are correctly reflected compared to the stochastic simulation results: for instance, both the peaks' amplitude and the time at which they occur are reproduced quite accurately.

Though not exact with respect to the stochastic simulation, these results are satisfactory enough for the kinds of semi-quantitative properties we are interested in analysing in the next section. Exact quantitative analysis via CTMC, instead, is infeasible for systems such as the model we consider here. Indeed, the time needed for obtaining the results shown in Fig. 6 ranges from a couple of seconds to hours, and for $H = 250$ the size of the CTMC already becomes prohibitively large to analyse.

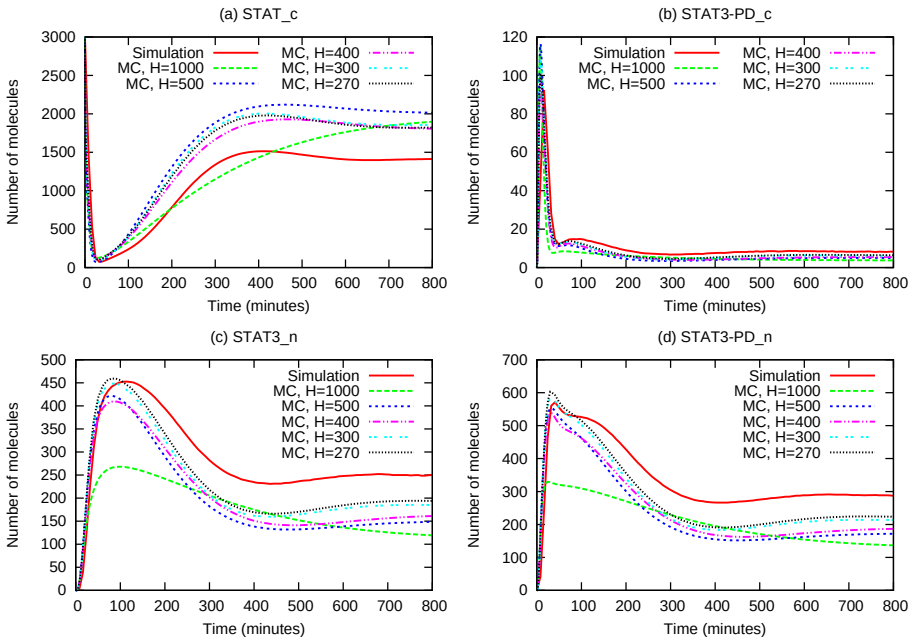

Fig. 6. Time-series by model-checking: STAT3 sub-model.

Semi-quantitative Properties. Using again the "trade-off" step sizes $H = 200$ and $H = 300$, we consider here a few more *semi-quantitative* properties of the two sub-models.

For instance, we are interested in analysing the impact that the different affinities of ligand/receptor pairs have on the consumption of the different ligands and receptors and on the relative amount of type I and type II receptors formed.

Though this kind of analysis is clearly quantitative (since it involves calculating probabilities and, hence, numbers of molecules), we consider such properties *semi-quantitative* because we are not interested in computing absolute values, but rather in knowing relative values with respect to each other.

The following properties measure the probability with which the amount of each molecular species never changes from the initial amount.

$$\mathcal{P}_{=?}[\mathbf{G}\ (LIF = N_U\_LIF)] \rightarrow\ 7.53 \cdot 10^{-2}$$

$$\mathcal{P}_{=?}[\mathbf{G}\ (OSM = N_U\_OSM)] \rightarrow\ 1.45 \cdot 10^{-6}$$

$$\mathcal{P}_{=?}[\mathbf{G}\ (gp130 = N_U\_gp130)] \rightarrow\ 0$$

$$\mathcal{P}_{=?}[\mathbf{G}\ (LIFR = N_U\_LIFR)] \rightarrow\ 1.24 \cdot 10^{-4}$$

$$\mathcal{P}_{=?}[\mathbf{G}\ (OSMR = N_U\_OSMR)] \rightarrow\ 4.56 \cdot 10^{-4}$$

From the obtained results we notice, for instance, that *gp130* is always used (indeed, it is necessary to form all receptor dimers), and that it is more likely for *OSM* to be consumed than *LIF* (indeed, *LIF* is only used in the formation of one type of receptor dimers).

We measure also the probability with which the amount of each molecular species reaches its lower bound.

This group of properties shows that *gp130* is totally consumed in any possible evolution of the system, that *LIF* and *OSM* are never totally consumed, and that the probability of *LIFR* being totally consumed is equal to the probability of *OSMR* not being used at all. These results mean that *gp130* is the bottleneck of the system, while *LIF* and *OSM* are present in abundance.

$$\mathcal{P}_{=?}[\mathbf{F}\ (LIF = N_L\_LIF)] \rightarrow\ 0$$

$$\mathcal{P}_{=?}[\mathbf{F}\ (OSM = N_L\_OSM)] \rightarrow\ 0$$

$$\mathcal{P}_{=?}[\mathbf{F}\ (gp130 = N_L\_gp130)] \rightarrow\ 1$$

$$\mathcal{P}_{=?}[\mathbf{F}\ (LIFR = N_L\_LIFR)] \rightarrow\ 4.56 \cdot 10^{-4}$$

$$\mathcal{P}_{=?}[\mathbf{F}\ (OSMR = N_L\_OSMR)] \rightarrow\ 1.24 \cdot 10^{-4}$$

Finally, we consider the reward-based property

$$\mathcal{R}^{i}_{=?}[C \leq T]$$

and we verify it on the downstream sub-model for time points $T \leq 800$ minutes, where i is an integer variable referring to a transition reward structure.

In addition to state rewards, in fact, PRISM allows for the definition of reward structures which associate with each transition a cumulative reward equal to its expected number of occurrences up to the considered time.

In Fig. 7 and Fig. 8 the expected number of occurrences for some of the reactions of the downstream sub-model is shown.

In particular, in Fig. 7 we compare the number of occurrences of receptors/STAT3 and receptors/SOCS3 binding reactions, which shows intuitively the different binding affinities of STAT3 and SOCS3 to the receptor dimers.

In Fig. 8, instead, we consider the number of occurrences of transport reactions of STAT3 molecules from cytoplasm to nucleus and back. In Fig. 8(a) we compare the number of occurrences of transport in the two directions: we count each transport from cytoplasm to nucleus twice since STAT3 molecules are translocated in the nucleus in dimeric form and, hence, a pair of STAT3 molecules is moved at each reaction occurrence. We observe that, since at system initialisation no STAT3 molecule is present in the nucleus, the difference between the two curves in Fig. 8(a) (multiplied by the step size H) must be the number of STAT3 molecules present in the nucleus. This consideration is confirmed by the perfect agreement of the two curves in Fig. 8(b).

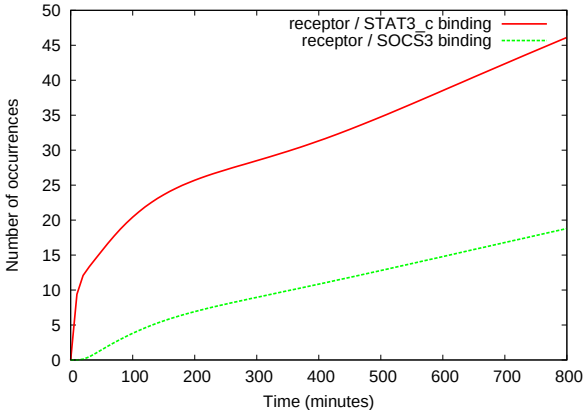

Fig. 7. Expected number of occurrences of receptor binding reactions in the downstream sub-model. The full red line plots the number of occurrences of reactions $bind\_rcpt\_DP\_stat_{27}$ and $bind\_rcpt\_DP\_stat_{28}$, while the dashed green line plots the number of occurrences of reactions $bind\_rcpt\_DP\_socs_{62}$ and $bind\_rcpt\_DP\_socs_{63}$.

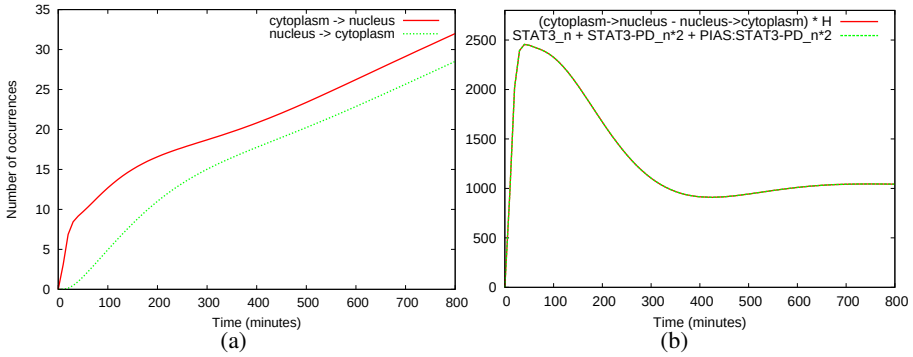

Fig. 8. Expected number of occurrences of transport reactions in the downstream sub-model. In (a) the full red line plots twice the number of occurrences of reaction $reloc\_stat\_cn_{58}$, while the dashed green line plots the number of occurrences of reaction $reloc\_stat\_nc_{60}$. In (b) the full red line is the difference between the lines in (a) multiplied by H, while the dashed green line is the total current amount of STAT3 molecules in the nucleus ($STAT3\_n + STAT3\text{-}PD\_n \cdot 2 + PIAS\text{:}STAT3\text{-}PD\_n \cdot 2$).

8 Related Work

Given its significant impact on various cellular processes, the gp130/JAK/STAT pathway has been subject of numerous studies, both experimental and computational. Consequently, a few variants of the pathway model have been developed in order to analyse different aspects of it.

In [18] the focus is on the shuttling of STATs from nucleus to cytoplasm and back. A more complete model is developed in [19], which also reports the results of a global sensitivity analysis of parameter interaction. The role of inhibitory mechanisms is instead studied in [20]. These three works are based on mathematical modelling and the analysis is performed by ODEs solvers.

In [21], a process algebra based computational model of the gp130/JAK/STAT pathway is presented and analysed using the BetaWB tool [31], a stochastic simulator for the BlenX language [32]. The Bio-PEPA model we present here is strongly based on the BlenX model described in [21], and the simulation results of the two models match well. This agreement is particularly interesting in view of the conceptual differences existing in the two process algebras. One of these differences concerns the treatment of complexes, which in BlenX are considered as molecular species consisting of the individual molecules composing them, while in Bio-PEPA they are considered as different species not explicitly related to the sub-components. Secondly, immediate reactions can be defined in BlenX, while they are not admitted in Bio-PEPA because of Bio-PEPA's underlying CTMC semantics. Finally, stoichiometric information can be specified in Bio-PEPA, while they cannot be explicitly coded in BlenX (requiring reactions involving stoichiometry greater than one to be decomposed into multiple steps). In addition to these theoretical differences between the languages, we mention that the focus in the two works is quite different. In [21] the effects of a number of experiments involving quantitative parameters are analysed and compared with experimental data. The aim of the present work, instead, is to exploit model-checking, in addition to stochastic simulation, to analyse both qualitative and quantitative properties of the model behaviour.

A few works have recently been published regarding the application of model-checking techniques to the analysis of biochemical systems. In [33] the authors demonstrate how the PRISM model-checker can be adopted to model and analyse biochemical pathways, using the FGF pathway as a case study. The approach proposed in this work differs from ours in the level of abstraction considered. Instead of taking a variable number of levels into account, the authors of [33] consider an abstraction in which one single copy of each involved molecular species is present and such that module variables represent changes in state of the molecules. This approach has the evident advantage of reducing the CTMC state space, though it might not be quantitatively correct in general: it can be seen as a level of abstraction equivalent to ours when one single level is used for each species. In the same work, the authors also consider a number of state space reduction techniques, some of which (based on lumpability and symmetry reduction) are exact, meaning that the behaviour of the reduced CTMC is preserved.

The notion of CTMC with levels of concentrations has been introduced in [34], in which the ERK signalling pathway was used as a case study, and in [35] the PRISM model-checker is used to analyse it. Following these works, the notion of discrete levels of concentrations has been adopted also in IDD-CSL [36], an Interval Decision Diagram

based model-checker for stochastic Petri nets, which allows for the verification of CSL properties.

In [37] the authors propose a framework, based on Petri nets, in which qualitative and quantitative (stochastic and continuous) analysis of biochemical pathways are integrated. Qualitative properties such as boundedness, liveness and reversibility are considered, in addition to the possibility to check for P- and T-invariants, and behavioural properties are verified by probabilistic model-checking.

Finally, BIOCHAM [38, 39] is a framework for modelling, simulating and analysing biochemical systems, in which different semantics (differential, stochastic, discrete, and boolean) are considered. BIOCHAM allows for the verification of temporal properties expressed in the Computation Tree Logic (CTL) by using the NuSMV model-checker [40].

9 Conclusions and Future Work

In this work we have used the gp130/JAK/STAT signalling pathway as a case study for modelling and analysis using the Bio-PEPA process algebra. Among the possible analysis methods made available by the Bio-PEPA Workbench, we have considered stochastic simulation and model-checking.

The results obtained by simulation agree well with existing mathematical and computational models. The application of the model-checking approach to the analysis of the pathway model, though limited by the state space explosion problem, provided us with some useful insight. First, it can be used for consistency checking, in order to guarantee the satisfaction of essential properties and, therefore, the absence of modelling errors. Second, it allows us to check for the satisfaction of semi-quantitative behavioural properties over the whole model, without the need for computing average values over a number of stochastic simulation runs.

In order to deal with the computational complexity of model-checking, we have subdivided the pathway model into two distinct sub-models. The time-series analysis obtained by analysing the sub-models individually via model-checking shows a reasonably good agreement with the behaviour obtained via stochastic simulation. The issue of modularisation of models of biochemical systems is a complex one. In this work we have adopted a simple approach which is adequate for this particular case study. A general approach for modularisation of models deserves additional study, in particular in view of the possible performance improvement which this technique could bring in model-checking.

Finally, in order to fully exploit the framework provided by Bio-PEPA further analysis could be performed on the MATLAB model generated by the Bio-PEPA Workbench using ODEs based methods to perform, for instance, bifurcation, stability, and continuation analysis.

Acknowledgments. The author wishes to thank Jane Hillston for her helpful comments. This research is supported by the EPSRC grant EP/E031439/1 “Stochastic Process Algebra for Biochemical Signalling Pathway Analysis”.

References

1. Regev, A., Silverman, W., Shapiro, E.: Representation and simulation of biochemical processes using the π-calculus process algebra. In: Proceedings of Pacific Symposium on Biocomputing (PSB'01). Volume 6. (2001) 459–470
2. Regev, A., Panina, E.M., Silverman, W., Cardelli, L., Shapiro, E.Y.: BioAmbients: an Abstraction for Biological Compartments. Theoretical Computer Science **325**(1) (2004) 141–167
3. Cardelli, L.: Brane Calculi - Interactions of Biological Membranes. In: Proceedings of Computational Methods in Systems Biology (CMSB'04). Volume 3082 of LNCS. (2005) 257–278
4. Priami, C., Quaglia, P.: Operational patterns in Beta-binders. Transactions on Computational Systems Biology **1** (2005) 50–65
5. Danos, V., Laneve, C.: Formal molecular biology. TCS **325**(1) (2004)
6. Regev, A., Shapiro, E.: Cells as Computation. Nature **419**(6905) (2002) 343
7. Ciocchetta, F., Hillston, J.: Bio-PEPA: an extension of the process algebra PEPA for biochemical networks. In: Proc. of FBTC'07. Volume 194 of ENTCS. (2008) 103–117
8. Ciocchetta, F., Hillston, J.: Bio-PEPA: a Framework for the Modelling and Analysis of Biochemical Networks. Theoretical Computer Science (2009, in press)
9. Hillston, J.: A Compositional Approach to Performance Modelling. Cambridge University Press (1996)
10. Ciocchetta, F., Hillston, J.: Calculi for Biological Systems. In: Formal Methods for Computational Systems Biology (SFM'08). Volume 5016 of LNCS. Springer-Verlag (2008) 265–312
11. Bio-PEPA Workbench Home Page: `http://www.dcs.ed.ac.uk/home/stg/software/biopepa/`
12. Ramsey, S., Orrell, D., Bolouri, H.: Dizzy: stochastic simulation of large-scale genetic regulatory networks. J. Bioinf. Comp. Biol. **3**(2) (2005) 415–436
13. PRISM Home Page: http://www.prismmodelchecker.org
14. Aziz, A., Sanwal, K., Singhal, V., Brayton, R.: Model-checking continuous-time Markov chains. ACM Trans. Comput. Logic **1**(1) (2000) 162–170
15. Underhill-Day, N., Heath, J.: Oncostatin M (OSM) Cytostasis of Breast Tumor Cells: Characterization of an OSM Receptor β-Specific Kernel. Cancer Research **66**(22) (2006) 10891–10901
16. Heinrich, P., Behrmann, I., Haan, S., Hermanns, H., Müller-Newen, G., Schaper, F.: Principles od interleukin (IL)-6-type cytokine signalling and its regulation. Biochem. J. **374** (2003) 1–20
17. Kisseleva, T., Bhattacharya, S., Braunstein, J., Schindler, C.: Signaling through the JAK/STAT pathway, recent advances and future challenges. Gene **285** (2002) 1–24
18. Swameye, I., Müller, T., Timmer, J., Sandra, O., Klingmüller, U.: Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. PNAS **100** (2003) 1028–1033
19. Mahdavi, A., Davey, R.E., Bhola, P., Yin, T., Zandstra, P.W.: Sensitivity Analysis of Intracellular Signaling Pathway Kinetics Predicts Targets for Stem Cell Fate Control. PLoS Computational Biology **3**(7) (2007) 1257–1267
20. Singh, A., Jayaraman, A., Hahn, J.: Modeling Regulatory Mechanisms in IL-6 Transduction in Hepatocytes. Biotechnology and Bioengineering **95**(5) (2006) 850–862
21. Guerriero, M.L., Dudka, A., Underhill-Day, N., Heath, J.K., Priami, C.: Narrative-based computational modelling of the Gp130/JAK/STAT signalling pathway. BMC Systems Biology **3**(1) (2009) 40
22. Bio-PEPA Home Page: `http://www.biopepa.org/`

23. Hinton, A., Kwiatkowska, M., Norman, G., Parker, D.: PRISM: A tool for automatic verification of probabilistic systems. In: Proc. 12th International Conference on Tools and Algorithms for the Construction and Analysis of Systems (TACAS'06). Volume 3920 of LNCS., Springer (2006) 441–444
24. Dizzy Home Page: `http://magnet.systemsbiology.net/software/Dizzy`
25. Aziz, A., Kanwal, K., Singhal, V., Brayton, V.: Verifying continuous time Markov chains. In: Proc. 8th International Conference on Computer Aided Verification (CAV'96). Volume 1102 of LNCS., Springer (1996) 269–276
26. Baier, C., Katoen, J.P., Hermanns, H.: Approximate Symbolic Model Checking of Continuous-Time Markov Chains. In: Proceedings of CONCUR'99. Volume 1664 of LNCS. (1999) 146–161
27. Saez-Rodriguez, J. Kremling, A., Gilles, E.: Dissecting the puzzle of life: modularization of signal transduction networks. Computers and Chemical Engineering **29** (2005) 619–629
28. Conzelmann, H., Saez-Rodriguez, J., Sauter, T., Bullinger, E., Allgöwer, F., Gilles, E.: Reduction of mathematical models of signal transduction networks: simulation-based approach applied to EGF receptor signalling. Systems Biology **1**(1) (2004) 159–169
29. Monteiro, P., Ropers, D., Mateescu, R., Freitas, A., de Jong, H.: Temporal logic patterns for querying dynamic models of cellular interaction networks. ECCB **24** (2008) 227–233
30. Gibson, M., Bruck, J.: Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. The Journal of Chemical Physics **104** (2000) 1876–1889
31. Dematté, L., Priami, C., Romanel, A.: The Beta Workbench: a computational tool to study the dynamics of biological systems. Briefings in Bioinformatics **9**(5) (2008) 437–449 (Tool available at `http://www.cosbi.eu/Rpty_Soft_BetaWB.php`).
32. Dematté, L., Priami, C., Romanel, A.: The BlenX Language: A Tutorial. In: Formal Methods for Computational Systems Biology (SFM'08). Volume 5016 of LNCS. Springer-Verlag (2008) 313–365
33. Heath, J., Kwiatkowska, M., Norman, G., Parker, D., Tymchyshyn, O.: Probabilistic Model Checking of Complex Biological Pathways. Theoretical Computer Science **319** (2008) 239–257
34. Calder, M., Gilmore, S., Hillston, J.: Modelling the Influence of RKIP on the ERK Signalling Pathway Using the Stochastic Process Algebra PEPA. Transactions on Computational Systems Biology (Proc of BioConcur'04) **7** (2006) 1–23
35. Calder, M., Vyshemirsky, V., Gilbert, D., Orton, R.: Analysis of signalling pathways using continuous time Markov chains. Transactions on Computational Systems Biology **6** (2006) 44–67
36. The Idd-CSL Home Page: `http://www-dssz.informatik.tu-cottbus.de/software/software.html`
37. Heiner, M., Gilbert, D., Donaldson, R.: Petri Nets for Systems and Synthetic Biology. In: Formal Methods for Computational Systems Biology (SFM'08). Volume 5016 of LNCS. Springer-Verlag (2008) 215–264
38. The BIOCHAM Home Page: http://contraintes.inria.fr/BIOCHAM/
39. Fages, F., Soliman, S., Chabrier-Rivier, N.: Modelling and querying interaction networks in the biochemical abstract machine BIOCHAM. Journal of Biological Physics and Chemistry **4**(2) (2004) 64–73
40. NuSMV Home Page: http://nusmv.irst.itc.it/