

Rare Event Handling in Signalling Cascades

Benoît Barbot, Serge Haddad and Claudine Picaronny

LSV, ENS Cachan & CNRS & Inria,
Cachan, France

{barbot, haddad, picaronny}@lsv.ens-cachan.fr

Monika Heiner

Brandenburg University of Technology,
Cottbus, Germany

monika.heiner@b-tu.de

Abstract—Signalling cascades are a recurrent pattern of biological regulatory systems whose analysis has deserved a lot of attention. It has been shown that Stochastic Petri Nets (SPN) are appropriate to model such systems and evaluate the probabilities of specific properties. Such an evaluation can be done numerically when the combinatorial state space explosion is manageable or statistically otherwise. However, when the probabilities to be evaluated are too small, random simulation requires more sophisticated techniques for the handling of rare events. In this paper, we show how such involved methods can be successfully applied for signalling cascades. More precisely, we study three relevant properties of a signalling cascade with the help of the COSMOS tool. Our experiments point out interesting dependencies between quantitative parameters of the regulatory system and its transient behaviour. In addition, they demonstrate that we can go beyond the capabilities of MARCIE, which provides one of the most efficient numerical solvers.

Keywords—Rare event; Importance sampling; Signalling cascade.

I. INTRODUCTION

Signalling cascades: Signalling processes play a crucial role for the regulatory behaviour of living cells. They mediate input signals, i.e., the extracellular stimuli received at the cell membrane, to the cell nucleus, where they enter as output signals the gene regulatory system. Understanding signalling processes is still a challenge in cell biology. To approach this research area, biologists design and explore signalling networks, which are likely to be building blocks of the signalling networks of living cells. Among them are the type of signalling cascades which we investigate here.

A signalling cascade is a set of reactions which can be grouped into levels. At each level a particular enzyme is produced (e.g., by phosphorylation); the level generally also includes the inverse reactions (e.g., dephosphorylation). The system constitutes a cascade since the enzyme produced at some level is the catalyser for the reactions at the next level. The catalyser of the first level is usually considered to be the input signal, while the catalyser produced by the last level constitutes the output signal. The transient behaviour of such a system presents a characteristic shape, the quantity of every enzyme increases to some stationary value. In addition, the increases are temporally ordered w.r.t. the levels in the signalling cascade. This behaviour can be viewed as a signal travelling along the levels, and there are many interesting properties to be studied like the travelling time of the signal, the relation between the variation of the enzymes of two consecutive levels, etc.

In [1], it has been shown how such a system can be modelled by a Petri net which can either be equipped with continuous transition firing rates leading to a continuous Petri

net which determines a set of differential equations or by stochastic transition firing rates leading to a SPN. This approach emphasises the importance of Petri nets which, depending on the chosen semantics, permit to investigate particular properties of the system. In this paper, we wish to explore the influence of stochastic features on the signalling behaviour, and thus we focus on the use of SPN.

Analysis of SPNs can be performed either numerically or statistically. The former approach is much faster than the latter and provides exact results up to numerical approximations, but its application is limited by the memory requirements due to the combinatory explosion of the state space.

Statistical evaluation of rare events: Statistical analysis means to estimate the results by evaluating a sufficient number of simulations. However, standard simulation is unable to efficiently handle *rare events*, i.e., properties whose probability of satisfaction is tiny. Indeed, the number of trajectories to be generated in order to get an accurate interval confidence for rare events becomes prohibitively huge. Thus, *acceleration* techniques [2] have been designed to tackle this problem whose principles consist in (1) favouring trajectories that satisfy the property, and (2) numerically adjusting the result to take into account the bias that has been introduced. This can be done by *splitting* the most promising trajectories [3] or *importance sampling* [4], i.e., modifying the distribution during the simulation. In a previous work [5], some of us have developed an original importance sampling method based on the design and numerical analysis of a reduced model in order to get the importance coefficients. This method was first proposed for checking “unbounded until” properties over models whose semantics is a discrete time Markov chain, it has been extended to also handle “bounded until” properties and continuous time Markov chains [6].

Our contribution: In this paper, we consider three families of properties for signalling cascades that are particularly relevant for the study of their behaviour and that are (depending on a scaling parameter) potentially rare events. From an algorithmic point of view, this case study raises interesting issues since the combinatorial explosion of the model quickly forbids the use of numerical solvers and its intricate (quantitative) behaviour requires elaborated and different abstractions depending on the property to be checked.

Due to these technical difficulties, the signalling cascade analysis has led us to substantially improve our method and in particular the way we obtain the final confidence interval. From a biological point of view, experiments have pointed out interesting dependencies between the scaling parameter of the model and the probability of satisfying a property.

Organisation: In Section II, we present the biological back-

ground, the signalling cascade under study and the properties to be studied. Then, in Section III after some recalls on SPN, we model signalling cascades by SPNs. We introduce the rare event issue and the importance sampling technique to cope with in Section IV. In Section V, we develop our method for handling rare events. Then, in Section VI, we report and discuss the results of our experiments. Finally in Section VII, we conclude and give some perspectives to our work. Additional explanations related to model abstractions, algorithmic considerations, and experimental statistical analysis can be found in the following research report [7].

II. SIGNALLING CASCADES

In technical terms, signalling cascades can be understood as networks of biochemical reactions transforming input signals into output signals. In this way, signalling processes determine crucial decisions a cell has to make during its development, such as cell division, differentiation, or death. Malfunction of these networks may potentially lead to devastating consequences on the organism, such as outbreak of diseases or immunological abnormalities. Therefore, cell biology tries to increase our understanding of how signalling cascades are structured and how they operate. However, signalling networks are generally hard to observe and often highly interconnected, and thus signalling processes are not easy to follow. For this reason, typical building blocks are designed instead, which are able to reproduce observed input/output behaviours.

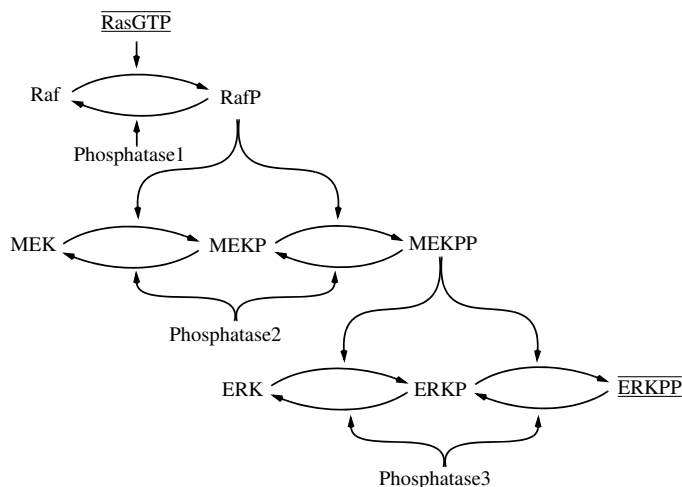


Figure 1. The general scheme of the considered three-level signalling cascade; RasGTP serves as input signal and ERKPP as output signal.

The case study we have chosen for our paper is such a signalling building block: the mitogen-activated protein kinase (MAPK) cascade [8]. This is the core of the ubiquitous ERK/MAPK network that can, among others, convey cell division and differentiation signals from the cell membrane to the nucleus. The description starts at the RasGTP complex which acts as an enzyme (kinase) to phosphorylate Raf, which phosphorylates MAPK/ERK Kinase (MEK), which in turn phosphorylates Extracellular signal Regulated Kinase (ERK). We consider RasGTP as the input signal and ERKPP (activated ERK) as the output signal. This cascade (RasGTP \rightarrow Raf \rightarrow MEK \rightarrow ERK) of protein interactions is known to control cell

differentiation, while the strength of the effect depends on the ERK activity, i.e., concentration of ERKPP.

The scheme in Figure 1 describes the typical modular structure for such a signalling cascade [9]. Each layer corresponds to a distinct protein species. The protein Raf in the first layer is only singly phosphorylated. The proteins in the two other layers, MEK and ERK, respectively, can be singly as well as doubly phosphorylated. In each layer, forward reactions are catalysed by kinases and reverse reactions by phosphatases (Phosphatase1, Phosphatase2, Phosphatase3). The kinases in the MEK and ERK layers are the phosphorylated forms of the proteins in the previous layer. Each phosphorylation/dephosphorylation step applies mass action kinetics according to the pattern $A + E \rightleftharpoons AE \rightarrow B + E$. This pattern reflects the mechanism by which enzymes act: first building a complex with the substrate, which modifies the substrate to allow for forming the product, and then disassociating the complex to release the product; for details see [10].

Having the wiring diagram of the signalling cascade, a couple of interesting questions arise whose answers would shed some additional light on the subject under investigation. Among them are an assessment of the signal strength in each level, and specifically of the output signal. We will consider these properties in Sections VI-A and VI-B. The general scheme of the signalling cascade also suggests a temporal order of the signal propagation in accordance with the level order. What cannot be derived from the structure is the extent to which the signals are simultaneously produced; we discuss this property in the technical report [7].

III. PETRI NET MODELLING

Stochastic Petri nets: Due to their graphical representation and bipartite nature, Petri nets are highly appropriate to model biochemical networks. When equipped with a stochastic semantics, yielding SPN [11], they can be used to perform quantitative analysis.

Definition 1 (SPN). A SPN \mathcal{N} is defined by a tuple $(P, T, \mathbf{Pre}, \mathbf{Post}, \{\mu_t\}_{t \in T})$ where P is a finite set of places, T is a finite set of transitions, $\mathbf{Pre}, \mathbf{Post}$ are matrices from $P \times T$ to \mathbb{N} , and $\{\mu_t\}_{t \in T}$ is a set of mappings from \mathbb{N}^P to $\mathbb{R}_{>0}$.

A marking m of SPN \mathcal{N} is an item of \mathbb{N}^P . A transition $t \in T$ is fireable in marking m if for all places $p \in P$ $m(p) \geq \mathbf{Pre}(p, t)$. Its firing leads to marking m' defined by: for all $p \in P$ $m'(p) = m(p) - \mathbf{Pre}(p, t) + \mathbf{Post}(p, t)$. This firing is denoted either $m \xrightarrow{\sigma} m'$ or as $m \xrightarrow{t}$ omitting the next marking. We extend these notations for any $\sigma = \sigma_1 \dots \sigma_n \in T^*$ of successive fireable transitions, if σ is fireable from m , that is if there exists a sequence of markings $m = m_0, m_1, \dots, m_n$ such that for all $0 \leq k < n$, $m_k \xrightarrow{\sigma_k} m_{k+1}$. Let m_0 be an initial marking, the reachability set $Reach(\mathcal{N}, m_0)$ is defined by: $Reach(\mathcal{N}, m_0) = \{m \mid \exists \sigma \in T^* m_0 \xrightarrow{\sigma} m\}$. The initialised SPNs (\mathcal{N}, m_0) that we consider do not have deadlocks: for all $m \in Reach(\mathcal{N}, m_0)$ there exists $t \in T$ such that $m \xrightarrow{t}$. each transition t is equipped with a mapping μ_t . In a marking m , each enabled transition of the Petri net randomly selects an execution time according to a Poisson process with rate $\mu_t(m)$. Then, the transition with earliest firing time is selected

TABLE I. DEVELOPMENT OF THE STATE SPACE FOR INCREASING N .

N	number of states	N	number of states
1	24,065 (4)	6	769,371,342,640 (11)
2	6,110,643 (6)	7	5,084,605,436,988 (12)
3	315,647,600 (8)	8	27,124,071,792,125 (13)
4	6,920,337,880 (9)	9	122,063,174,018,865 (14)
5	88,125,763,956 (10)	10	478,293,389,221,095 (14)

to fire yielding the new marking. Operational semantics of a SPN is a Continuous Time Markov Chain (CTMC). This can be formalized as follows.

Definition 2 (CTMC of a SPN). Let \mathcal{N} be a SPN and m_0 be an initial marking. Then the CTMC associated with (\mathcal{N}, m_0) is defined by its set of states which is $\text{Reach}(\mathcal{N}, m_0)$, its transition matrix \mathbf{P} defined by Equation 1

$$\mathbf{P}(m, m') = \frac{\sum_{m \xrightarrow{t} m'} \mu_t(m)}{\sum_{m \xrightarrow{t}} \mu_t(m)} \quad (1)$$

and, for each state m , the rate λ_m defined by:

$$\lambda_m = \sum_{m \xrightarrow{t}} \mu_t(m)$$

Running case study: We now explain how to model our running case study in the Petri net framework. The signalling cascade is made of several phosphorylation/dephosphorylation steps, which are built on mass/action kinetics. Each step follows the pattern $A + E \rightleftharpoons AE \rightarrow B + E$ and is modelled by a small Petri net component depicted in Figure 2. The mass action kinetics is expressed by the rate of the transitions. The marking-dependent rate of each transition is equal to the product of the number of tokens in all its incoming places up to a multiplicative constant given by the biological behaviour (summing up dependencies on temperature, pressure, volume, etc.).

The whole reaction network based on the general scheme of a three-level double phosphorylation cascade, as given in Figure 1, is modelled by the Petri net in Figure 3. The input signal is the number of tokens in the place RasGTP, and the output signal is the number of tokens in the place ERKPP.

This signalling cascade model represents a self-contained and closed system. It is covered with place invariants (see the research report for details), specifically each layer in the cascade forms a place invariant consisting of all states a protein can undergo; thus the model is bounded. Assuming an appropriate initial marking, the model is also live and reversible; see [1] for more details, where this Petri net has been developed and analysed in the qualitative, stochastic and continuous modelling paradigms. In our paper we extend these analysis techniques for handling properties corresponding to rare events.

We introduce a scaling factor N to parameterize how many tokens are spent to specify the initial marking. Increasing the scaling parameter can be interpreted in two different ways: either an increase of the biomass circulating in the closed system (if the biomass value of one token is kept constant), or an increase of the resolution (if the biomass value of one token inversely decreases, called level concept in [1]). The kind of interpretation does not influence the approach we pursue in this paper.

Increasing N increases the size of the state space and thus of the CTMC, as shown in Table I, which has been computed

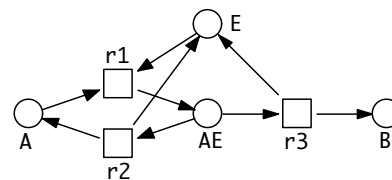


Figure 2. Petri net pattern for mass action kinetics $A + E \rightleftharpoons AE \rightarrow B + E$.

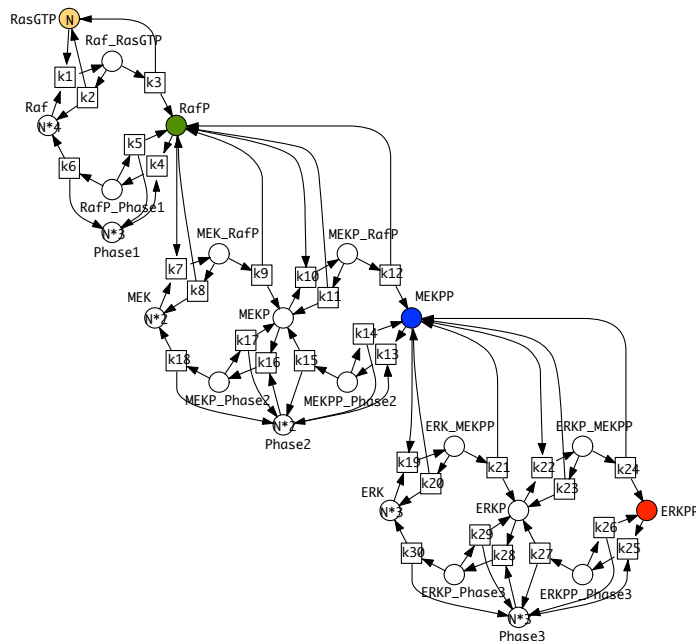


Figure 3. A Petri net modelling the three-level signalling cascade given in Figure 1; k_i are the kinetic constants for mass action kinetics, N the scaling parameter.

with the symbolic analysis tool MARCIE [12]. As expected, the explosion of the state space prevents numerical model checking for higher N and thus calls for statistical model checking.

Furthermore, increasing the number of states actually decrease the probabilities to be in a certain state, as the total probability of 1 is fixed. With the distribution of the probability mass of 1 over an increasingly huge number of states, we obtain sooner or later states with very tiny probabilities, and thus rare events. Neglecting rare events is usually appropriate when focusing on the averaged behaviour. But they become crucial when certain jump processes such as mutations under rarely occurring conditions are of interest.

IV. STATISTICAL MODEL CHECKING WITH RARE EVENTS

A. Statistical model checking and rare events

Simulation recalls: The statistical approach for evaluating the expectation $\mathbf{E}(X)$ of a random variable X related to a random path in a Markov chain is generally based on three parameters: the number of simulations K , the confidence level γ , and the width of the confidence interval lg (see [13]). Once the user provides two parameters, the procedure computes the remaining one. Then it performs K simulations of the Markov chain and outputs a confidence interval $[L, U]$ with a width of

at most lg such that $\mathbf{E}(X)$ belongs to this interval with a probability of at least γ .

Statistical evaluation of a reachability probability: Let \mathcal{C} be a Discrete Time Markov Chain (DTMC) with two absorbing states s_+ or s_- , such that the probability to reach s_+ or s_- from any state is equal to 1. Assume one wants to estimate p , the probability to reach s_+ . Then the simulation step consists in generating K paths of \mathcal{C} which end in an absorbing state. Let K_+ be the number of paths ending in state s_+ . The random variable K_+ follows a binomial distribution with parameters p and K . Unfortunately, when $p \ll 1$, the number of paths required for a small confidence interval is too large to be simulated. This issue is known as the *rare event* problem.

Importance sampling: In order to tackle the rare event problem, the importance sampling method relies on a choice of a biased distribution that will artificially increase the frequency of the observed rare event during the simulation. The generation of paths is done according to a modified DTMC \mathcal{C}' , with the same state space, but modified transition matrix \mathbf{P}' . \mathbf{P}' must satisfy Property 2:

$$\mathbf{P}(s, s') > 0 \Rightarrow \mathbf{P}'(s, s') > 0 \vee s' = s_- \quad (2)$$

which means that this modification cannot remove transitions that have not s_- as target, but can add new transitions. The method maintains a correction factor called L initialised to 1; this factor represents the *likelihood* of the path. When a path crosses a transition $s \rightarrow s'$ with $s' \neq s_-$, L is updated by $L \leftarrow L \frac{\mathbf{P}(s, s')}{\mathbf{P}'(s, s')}$. When a path reaches s_- , L is set to zero. If $\mathbf{P}' = \mathbf{P}$ (i.e., no modification of the chain), the value of L when the path reaches s_+ (resp. s_-) is 1 (resp. 0). Let V_s (resp. W_s) be the random variable associated with the final value of L for a path starting in x in the original model \mathcal{C} (resp. in \mathcal{C}'). By definition, the expectation $\mathbf{E}(V_{s_0}) = p$ and by construction of the likelihood, $\mathbf{E}(W_{s_0}) = p$. Of course, a useful importance sampling should reduce the variance of W_{s_0} w.r.t. to the one of V_{s_0} equal to $p(1-p) \approx p$ for a rare event.

V. OUR METHODOLOGY FOR IMPORTANCE SAMPLING

A. Previous work

In [5][6], we provided a method to compute a biased distribution for importance sampling: we manually design an abstract smaller model, with a behaviour close to that of the original model, that we call the *reduced model*. This is done by lumping together some states with the objective of making the rare event less rare. We perform numerical computations on this smaller model to obtain the biased distribution. We applied this method in order to tackle the estimation of time bounded property in CTMCs, that is the probability to satisfy a formula $aU^{[0, \tau]}b$, when it is a rare event. Let us outline the different steps of the method which is depicted in Figure 4.

Abstraction of the model: As discussed above, given a SPN \mathcal{N} modelling the system to be studied, we manually design an appropriate reduced one \mathcal{N}^\bullet and a correspondence function f from states of \mathcal{N} to states of \mathcal{N}^\bullet . Function f is defined at the net level (see Section VI).

Structural analysis: Importance sampling was originally proposed for DTMCs. In order to apply it for CTMC \mathcal{C} associated with net \mathcal{N} , we need to uniformize \mathcal{C} (and also \mathcal{C}^\bullet associated with \mathcal{N}^\bullet) which means finding a bound Λ for exit rate of states, i.e., markings, considering Λ as the uniform

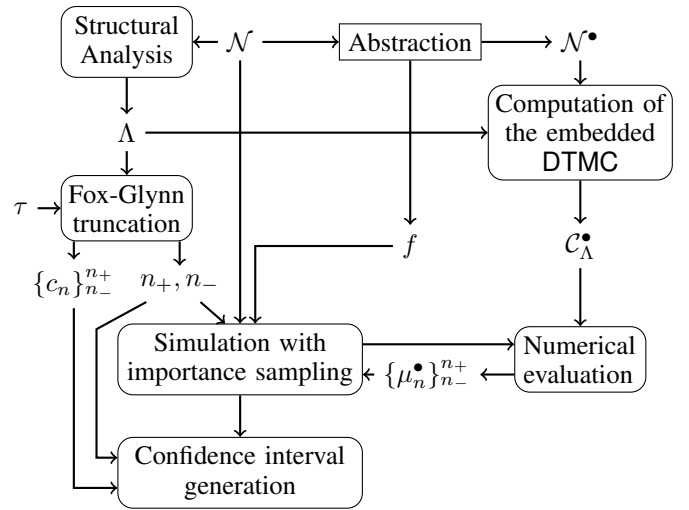


Figure 4. Principles of the methodology for estimating probability of occurrence of rare events.

exit rate of states and rescaling accordingly the transition probability matrices [14]. Since the rates of transitions depend on the current marking, determining Λ requires a structural analysis like invariant computations for bounding the number of tokens in places.

Fox-Glynn truncation: Given a uniform chain with initial state s_0 , exit rate Λ , and transition probability matrix \mathbf{P} , the state distribution π_τ at time τ is obtained by the following formula: $\pi_\tau(s) = \sum_{n>0} \frac{e^{-\Lambda\tau} (\Lambda\tau)^n}{n!} \mathbf{P}^n(s_0, s)$. This value can be estimated, with sufficient precision, by applying [15]. Given two numerical accuracy requirements α and β , truncation points n^- and n^+ and values $\{c_n\}_{n^- \leq n \leq n^+}$ are determined such that for $n^- \leq n \leq n^+$: $c_n(1 - \alpha - \beta) \leq \frac{e^{-\Lambda\tau} (\Lambda\tau)^n}{n!} \leq c_n$, $\sum_{n < n^-} \frac{e^{-\Lambda\tau} (\Lambda\tau)^n}{n!} \leq \alpha$ and $\sum_{n > n^+} \frac{e^{-\Lambda\tau} (\Lambda\tau)^n}{n!} \leq \beta$.

Computation of the embedded DTMC: We build the embedded DTMC $\mathcal{C}_\Lambda^\bullet$ of \mathcal{N}^\bullet after uniformization. Since we want to evaluate the probability to satisfy formula $aU^{[0, \tau]}b$, the states satisfying a (resp. $\neg a \wedge \neg b$) are aggregated into an absorbing accepting (resp. rejecting) state. Let $\mu_n(s^\bullet)$ be the probability to be in the accepting state at time τ starting from state s^\bullet .

Numerical evaluation: Matrix \mathbf{P}' used for importance sampling simulation in the embedded DTMC of \mathcal{N} to evaluate formulas $aU^{[0, n]}b$ for $n^- \leq n \leq n^+$, is based on the distributions $\{\mu_n^*\}_{0 < n \leq n^+}$, where $\mu_n^*(s^\bullet)$ is the probability that a random path of the embedded DTMC of \mathcal{N}^\bullet starting from s^\bullet fulfills $aU^{[0, n]}b$. Such a distribution is computed by a standard numerical evaluation. However since n^+ can be large, depending on the memory requirements, this computation can be done statically for all n or dynamically for a subset of such n during the importance sampling simulation (more details are given in the research report).

Simulation with importance sampling: Here the random distribution of the successors of a state depend on both the embedded DTMC \mathcal{C}_Λ and the values computed by the numerical evaluation. Moreover, all formulas $aU^{[0, n]}b$ for $n^- \leq n \leq n^+$ have to be evaluated increasing the time complexity of the method.

Generation of the confidence interval: After the simulations we get a family of confidence intervals indexed by $n^- \leq n \leq n^+$. Using Fox-Glynn truncation, we weight and combine the confidence intervals and return the final interval.

B. Tackling signalling cascades

Tackling signalling cascades leads us to two improvements: (1) we have performed a much more efficient importance sampling simulation and (2) we have proposed different ways of computing confidence intervals. We now detail these issues.

Importance sampling for multiple formulas: A naive implementation would require to apply statistical model checking of formulas $aU^{[0,n]}b$ for all n between n^- and n^+ , but such a number can be large. A more tricky alternative consists in producing all trajectories until time horizon n^+ and updating the simulation results at the end of a trajectory for all the intervals $[0, n]$ with $n^- \leq n \leq n^+$ as follows. If the trajectory has reached the absorbing rejecting state s^- then it is an unsuccessful trajectory for all intervals. Otherwise if it has reached the absorbing accepting state s^+ at time n_0 then for all $n \geq n_0$ it is a successful trajectory and for all $n < n_0$ it is unsuccessful. Doing this way, every trajectory contributes to all evaluations, and we significantly increase the sample size without increasing computational cost. The accuracy of the results is improved.

Confidence interval estimation: The result of each trajectory of the simulation is a realisation of the random variable $W_{s_0} = X_{s_0}L_{s_0}$ where the binary variable X_{s_0} indicates whether a trajectory starting from s_0 is successful and the positive random variable L_{s_0} is the (random) likelihood. Observe that $\mathbf{E}(W_{s_0}) = \mathbf{E}(L_{s_0}|X_{s_0} = 1)\mathbf{E}(X_{s_0})$. Since X_{s_0} follows a Bernoulli distribution, a confidence interval can be easily computed for $\mathbf{E}(X_{s_0})$. For $\mathbf{E}(L_{s_0}|X_{s_0} = 1)$ several approaches are possible among them we have selected three possible computations ranked by conservation degree. The first method assumes that the distribution is Gaussian (which is asymptotically valid if the variance is finite, thanks to the central limit theorem). Another method uses a pseudo Chernoff-Hoeffding bound. Whenever the random variable is bounded, this method is asymptotically valid. In our case we use the minimal and maximal values observed during the simulation as the bounds of L_{s_0} . The last method consists in returning the minimal and maximal observed values as the confidence interval.

VI. EXPERIMENTS

We have analysed two properties, the one of them is inspired by [1]. Recall that the initial marking of the model is parametrized by a scaling factor N . For the first property, the reduced model is the same model but with local smaller scaling factors on the different layers of phosphorylation. Every state of the initial model is mapped (by f) to a state of the abstract model which has the “closest” proportion of chemical species. For instance let $N = 4$, which corresponds to 16 species of the first layer, a state with 6 tokens in Raf and 10 tokens in RafP is mapped, for a reduced model with $N = 3$, to a state with $4 = \lfloor 6 \times 3/4 \rfloor$ tokens in Raf and $8 = \lceil 10 \times 3/4 \rceil$ tokens in RafP (see the research report for a specification of f).

All statistical experiments have been carried out with our tool COSMOS [16]. COSMOS is a statistical model checker for the HASL logic [16]. It takes as input a Petri net (or a

high-level Petri net) with general distributions for transitions. It performs an efficient statistical evaluation of the SPN by generating a code per model and formula. In the case of importance sampling, it additionally takes as inputs the reduced model and the mapping function specified by a C function and returns the different confidence intervals. All experiments have been performed on a machine with 16 cores running at 2 GHz and 32 GB of memory both for the statistical evaluation of COSMOS and the numerical evaluation of MARCIE.

We perform additional experiments with a third property, which can be found in the research report.

A. Maximal peak of the output signal

The first property is expressed as a time-bounded reachability formula assessing the strength of the output signal of the last layer: “What is the probability to reach within 10 time units a state where the total mass of ERK is doubly phosphorylated?”, associated with probability p_1 defined by:

$$p_1 = \Pr(\text{True } U^{\leq 10}(\text{ERKPP} = 3N))$$

The inner formula is parametrized by N , the scaling factor of the net (via its initial marking). The reduced model that we design for COSMOS uses different scaling factors for the three layers in the signalling cascade. The first two layers of phosphorylation which are based on Raf and MEK always use a scaling factor of 1, whereas the last layer involving ERK uses a scaling factor of N . The second column of Table II shows the ratio between the number of reachable states of the original and the reduced models.

We have performed experiments with both COSMOS and MARCIE. The time and memory consumptions for increasing values of N are reported in Table II. For each value of N we generate one million trajectories with COSMOS. We observe that the time consumption significantly increases between $N = 3$ and $N = 4$. This is due to a change of strategy in the space/time trade-off in order to not exceed the machine memory capacity. MARCIE suffers an exponential increase w.r.t. both time and space resources. When $N = 3$, it is slower than COSMOS and it is unable to handle the case $N = 4$.

Table II depicts the values returned by the two tools: MARCIE returns a single value, whereas COSMOS returns three confidence intervals (discussed above) with a confidence level set to 0.99. We observe that confidence intervals computed by the Gaussian analysis neither contain the result, the ones computed by Chernoff-Hoeffding do not contain it for $N = 3$, and the most conservative ones always contain it (when this result is available). An analysis of the likelihood L_{s_0} is detailed in the research report. It appears that the probability p_1 depends on N in an exponential way: $p_1 \approx 800(3 \cdot 10^{-15})^N$. The constants occurring in the formula could be interpreted by biologists.

B. Conditional maximal signal peak

The network structure of each layer in the signalling cascade presents a cyclic behaviour, i.e., phosphorylated proteins, serving as signal for the next layer, can also be dephosphorylated again, which corresponds to a decrease of the signal strength. Thus an interesting property of the signalling cascade is the probability of a further increase of the signal strength under the condition that a certain strength has already been reached. We estimate this quantity for the first layer in the signalling cascade, i.e., RafP, and ask specifically for the

TABLE II. COMPUTATIONAL COMPLEXITY AND NUMERICAL VALUES RELATED TO THE EVALUATION OF p_1

N	COSMOS						MARCIE		
	Reduction factor	time	memory	Gaussian CI	Chernoff CI	MinMax CI	time	memory	Output
1	-	-	-	-	-	-	4	514MB	2.07E-12
2	38	20,072	3,811MB	[3.75E-27,5.88E-26]	[3.75E-27,4.54E-25]	[3.75E-27,1.57E-23]	326	801MB	8.18E-26
3	558	15,745	15,408MB	[4.34E-42,1.72E-39]	[4.34E-42,1.82E-38]	[4.43E-42,1.87E-37]	43,440	13,776MB	2.56E-39
4	4667	40,241	3,593MB	[1.54E-57,8.54E-56]	[1.54E-57,1.98E-55]	[1.78E-57,7.05E-55]	Out of Memory: >32GB		
5	27353	51,120	19,984MB	[3.97E-73,2.33E-70]	[3.97E-73,7.30E-70]	[5.44E-73,2.24E-69]			

TABLE III. NUMERICAL VALUES ASSOCIATED WITH p_2

N	L	COSMOS		MARCIE		
		confidence interval	time	result	time	memory
2	2	[2.39E-13 , 1.07E-09]	31	5.55E-10	90	802 MB
2	3	[2.18E-10 , 6.92E-08]	110	6.64E-08	136	816 MB
2	4	[9.33E-08 , 3.54E-05]	256	3.01E-06	276	798 MB
2	5	[1.16E-05 , 6.08E-04]	1000	7.16E-05	759	801 MB
2	6	[5.42E-04 , 1.21E-03]	5612	1.27E-03	3180	804 MB
3	5	[1.82E-12 , 9.78E-09]	459	Time > 48 hours		
3	6	[3.41E-10 , 9.66E-08]	1428			
3	7	[1.81E-08 , 2.23E-06]	7067			
3	8	[8.72E-07 , 2.71E-06]	4460			
3	9	[1.42E-06 , 4.59E-05]	4301			
3	10	[2.69E-04 , 9.34E-04]	6420			
4	10	[5.12E-09 , 2.75E-08]	8423	Memory > 32GB		
4	11	[8.23E-08 , 2.97E-07]	7157			
4	12	[9.84E-07 , 1.86E-06]	18730			

probability to reach its maximal strength, $4N$: “What is the probability of the concentration of RaFP to continue its increase and reach $4N$, when starting in a state where the concentration is for the first time at least L ?”. This is a special use case of the general pattern introduced in [1].

$$p_2 = \Pr_{\pi}((\text{RaFP} \geq L) \cup (\text{RaFP} \geq 4N))$$

where π is the distribution over states when satisfying for the first time the state formula $\text{RaFP} \geq L$ (previously called a filter). This formula is parametrized by threshold L and scaling factor N . The results for increasing N and L are reported in Table III (confidence intervals are computed by Chernoff-Hoeffding method). As before, MARCIE cannot handle the case $N = 3$, the bottleneck being here the execution time.

It is clear that p_2 is an increasing function of L . More precisely, experiments point out that p_2 increases approximately exponentially by at least one magnitude order when L is incremented. However this dependency is less clear than the one of the first property.

VII. CONCLUSION AND FUTURE WORK

We have studied rare events in signalling cascades with the help of an improved importance sampling method implemented in COSMOS. Our method has been able to cope with huge models (nearly a hundred billion states) that could not be handled neither by computations nor by standard simulations. Analysis of the experiments has pointed out interesting dependencies between the scaling parameter and the quantitative behaviour of the model.

In future work, we intend to incorporate other types of quantitative properties, such as the mean time a signal needs to exceed a certain threshold, the mean travelling time from the input to the output signal, etc. We also plan to analyse other biological systems like mutation rates in growing bacterial colonies [17].

REFERENCES

- [1] M. Heiner, D. Gilbert, and R. Donaldson, “Petri nets for systems and synthetic biology,” in SFM 2008, ser. LNCS, M. Bernardo, P. Degano, and G. Zavattaro, Eds., vol. 5016. Springer, 2008, pp. 215–264.
- [2] G. Rubino and B. Tuffin, Rare Event Simulation using Monte Carlo Methods. Wiley, 2009.
- [3] P. L’ecuyer, V. Demers, and B. Tuffin, “Rare events, splitting, and quasi-monte carlo,” ACM Transactions on Modeling and Computer Simulation (TOMACS), vol. 17, no. 2, 2007, pp. 1–44.
- [4] P. W. Glynn and D. L. Iglehart, “Importance sampling for stochastic simulations,” Management Science, vol. 35, no. 11, 1989, pp. 1367–1392.
- [5] B. Barbot, S. Haddad, and C. Picaronny, “Coupling and importance sampling for statistical model checking,” in TACAS, ser. LNCS, C. Flanagan and B. König, Eds., vol. 7214. Springer, 2012, pp. 331–346.
- [6] —, “Importance sampling for model checking of continuous time Markov chains,” in Proceedings of the 4th International Conference on Advances in System Simulation (SIMUL’12), P. Dini and P. Lorenz, Eds. Lisbon, Portugal: XPS, Nov. 2012, pp. 30–35.
- [7] B. Barbot, S. Haddad, M. Heiner, and C. Picaronny, “Rare event handling in signalling cascades,” LSV, ENS Cachan, Tech. Rep. 14-10, July 2014.
- [8] A. Levchenko, J. Bruck, and P. Sternberg, “Scaffold proteins may biphasically affect the levels of mitogen-activated protein kinase signaling and reduce its threshold properties,” Proc Natl Acad Sci USA, vol. 97, no. 11, 2000, pp. 5818–5823.
- [9] V. Chickarmane, B. N. Kholodenko, and H. M. Sauro, “Oscillatory dynamics arising from competitive inhibition and multisite phosphorylation,” Journal of Theoretical Biology, vol. 244, no. 1, January 2007, pp. 68–76.
- [10] R. Breitling, D. Gilbert, M. Heiner, and R. Orton, “A structured approach for the engineering of biochemical network models, illustrated for signalling pathways,” Briefings in Bioinformatics, vol. 9, no. 5, September 2008, pp. 404–421.
- [11] M. Ajmone Marsan, G. Balbo, G. Conte, S. Donatelli, and G. Franceschinis, Modelling with generalized stochastic Petri nets. John Wiley & Sons, Inc., 1994.
- [12] M. Heiner, C. Rohr, and M. Schwarick, “MARCIE - Model checking And Reachability analysis done effiCIently,” in Proc. PETRI NETS 2013, ser. LNCS, J. Colom and J. Desel, Eds., vol. 7927. Springer, 2013, pp. 389–399.
- [13] L. J. Bain and M. Engelhardt, Introduction to Probability and Mathematical Statistics, Second Edition. Duxbury Classic Series, 1991.
- [14] A. Jensen, “Markoff chains as an aid in the study of markoff processes,” Skand. Aktuarietidskr, 1953.
- [15] B. L. Fox and P. W. Glynn, “Computing Poisson probabilities,” Commun. ACM, vol. 31, no. 4, 1988, pp. 440–445.
- [16] P. Ballarini, H. Djafri, M. Dufloy, S. Haddad, and N. Pekergin, “HASL: An expressive language for statistical verification of stochastic models,” in Proc. VALUETOOLS’11, Cachan, France, May 2011, pp. 306–315.
- [17] D. Gilbert, M. Heiner, F. Liu, and N. Saunders, “Colouring Space - A Coloured Framework for Spatial Modelling in Systems Biology,” in Proc. PETRI NETS 2013, ser. LNCS, J. Colom and J. Desel, Eds., vol. 7927. Springer, June 2013, pp. 230–249.