A generic framework to coarse-grain stochastic reaction networks by Abstract Interpretation

Jérôme Feret, Albin Salazar

A generic framework to coarse-grain stochastic reaction networks by Abstract Interpretation

Jérôme Feret[1,2] and Albin Salazar[1,2]

[1] DI ENS, École normale supérieure, Université PSL, CNRS, INRIA, 75005 Paris, France
[2] INRIA, Paris, France

Abstract. In the last decades, logical or discrete models have emerged as a successful paradigm for capturing and predicting the behaviors of systems of molecular interactions. Intuitively, they consist in sampling the abundance of each kind of biochemical entity within finite sets of intervals and deriving transitions accordingly. On one hand, formally-proven sound derivation from more precise descriptions (such as from reaction networks) may include many fictitious behaviors. On the other hand, direct modeling usually favors dominant interactions with no guarantee on the behaviors that are neglected.
In this paper, we formalize a sound coarse-graining approach for stochastic reaction networks. Its originality relies on two main ingredients. Firstly, we abstract values by intervals that overlap in order to introduce a minimal effort for the system to go back to the previous interval, hence limiting fictitious oscillations in the coarse-grained models. Secondly, we compute for pairs of transitions (in the coarse-grained model) bounds on the probabilities on which one will occur first.
We illustrate our ideas on two case studies and demonstrate how techniques from Abstract Interpretation can be used to design more precise discretization methods, while providing a framework to further investigate the underlying structure of logical and discrete models.

Keywords: Abstract interpretation · stochastic reaction networks · logical modeling · coarse-graining

1 Introduction

The field of Systems Biology is driven by the development of tools to investigate emergent behaviors in populations of biological molecules from single entity interactions. Among these tools, mathematical models of systems of interactions have been critical in identifying hidden mechanisms by perturbation studies, drivers of disease phenotypes and in generating new hypotheses. Consequently, a major interest underlies the ability for modeling tools to recapitulate biological phenomenon and its repurposing for predictive studies.

In retrospect, developing modeling tools is a continuous field of investigation as it provides means to gain understanding of biological systems through *in silico* studies. A common battle to derive an ideal representation for a system

process is the descriptive trade-off between the simplicity and accuracy. On one hand, too simple models are prone to reproduce only *a priori* knowledge. On the other hand, too descriptive models result in behaviors too difficult to parse, or even yet compute. In both cases, gaining new insights is hampered. Thus, it is arguably important in generating tools to measure the impact model selection has in capturing biological phenomena, especially those which can be verified. For example, an overview of various formal modeling frameworks for Systems Biology have been reviewed in [4]. In this paper, we assess the impact of discretization in the study of biological systems.

Logical models are a popular class of discretized models. Recent development have made it an ideal modeling tool to perform perturbation studies for a myriad of biochemical interactions. For example, some logical models have been developed to study iron metabolism in breast cancer cells [7] and a cell division process in mammalian cells [11]. A key feature of logical models is in obtaining knowledge about a process while only partial information is available about some particular interactions and their kinetics. Albeit their success in recapitulating experimental observations and predicting local system properties, their underlying modeling assumptions are often kept implicit. There is indeed a gap between hand-written logical models and the models that can be formally derived from a more concrete level of representation (such as a stochastic reaction network).

One popular approach to design logical models has been proposed by René Thomas [16,19]. With this method, each dimension is associated with a unique attractor state (or focal point), which may depend on the current state of the system. Then the transitions of the system are obtained by assuming that on each dimension, the system may get closer to its focal point. Reverse transitions may also occur (but at low probabilities), but in practice they are neglected. Such simplifications of the model is usually justified by some time-scale separation principles [11] (that are mainly asymptotic reasonings providing convergence results when scales are infinitely separated). Two critical observations emerge about this modeling process. Firstly, it is unclear to which extent these simplifications actually impact the behaviors of the systems they try to represent. Secondly, considering all reverse transitions, without any information about their potential likelihood, would lead to inaccurate models with many fictitious non-deterministic behaviors.

In order to increase confidence in the modeling process, we would like to derive formally discrete models from more precise representations (such as reaction networks). For this purpose we use the abstract interpretation framework to coarse-grain reaction networks into discrete models of abstract regions of states while preserving formal relationships between the respective behaviors of both models. Yet we have to face several issues. Firstly, several behaviors that are usually neglected in the logical models may occur with a low probability in the initial reaction network. Thus a non-deterministic abstraction would be unhelpful, because non-deterministic models provide no means to distinguish rare events from more the common ones. Instead we propose to propagate the probabilities of transitions from the reference reaction network to the coarse-grained model,

so that bounds to the probability of unlikely behaviors can be indeed computed. For this purpose, we equip each transition in the abstract model with an interval for their probabilities which is formally derived from the underlying reaction network. Secondly, even with probabilities, naive abstractions lead to very imprecise models. This means that we have to adapt our abstraction in order to highlight the main behaviors of interest. Consequently, we obtain a Discrete-time Markov chain (DTMC) whereby the exclusion of probabilities would structurally mimic logical models, yet providing an accessible tool to quantify differences between logical models and the discrete models obtained from our formalizations. One may be tempted to use Continuous-time Markov chains (CTMC) rather than DTMCs. Yet, the additional information provided by the continuous setting is not relevant in our context. On one hand, we want to obtain models comparable to logical models where the notion of time has already been abstracted away. On the other hand, the exact moment when each event occurs does not affect the computation of the probabilities of transitions between abstract regions in our coarse-grained models. What matters is only the relative order between these events, not the exact moment when they have happened. Yet, abstracting away the notion of time of a CTMC while only keeping the relative order between events induces a DTMC, which justifies our choice thoroughly. Lastly, using DTMCs instead of CTMCs deeply simplifies the underlying mathematics. For instance, in the continuous setting, probability density functions are required to define when events are likely to occur and a topology is necessary to define the probability of which set of model executions can be computed [13]. In contrast, in the discrete setting, only discrete probabilities are necessary and the probability of each execution with a finite amount of steps can be computed.

Ideally, the upper bounds computed for the probabilities of transitions that correspond to less likely behaviors should be very low. To achieve this goal, it is important to refine the abstraction process and to distinguish the abstract regions of states according to which transitions have been taken to enter them. Furthermore, it is also important that every transition between abstract regions corresponds to sequences composed of at least a few concrete transitions. This motivates the use of overlapping intervals to coarse-grain models. In the formal discretization process, an interval is composed of a pair of boundary values that enclose a concrete value. Thus, for each concrete value, an abstract interval computation is sensitive to whether a concrete value has exited a visited interval. In the abstraction, we consider that a value changes to another interval only when it actually leaves its current interval (hence leaving the overlapping region between its previous and current intervals). This way, when entering a new region of states, going in the reverse direction requires crossing through the overlap between two consecutive intervals, which is likely to have low probability when it is against the main trend of the dynamics of the system. The so-obtained abstraction ignores small fluctuations while strengthening the sequences of transitions that follow the main trend of the system dynamics.

Related work. Value sampling is widely used to simplify dynamical systems. In piece-wise linear systems [9], the dynamics are approximated by one system of

linear equations per equivalence class of states. In [15], an abstract interpretation based on support functions makes the computation of their trajectories scale while ensuring a sound over-approximation of any potential behavior. In the context of discrete modeling, the Boolean semantics of BIOCHAM [10] is also an abstract interpretation of the stochastic semantics, but it is too conservative. In [1,2], value sampling is refined by exploiting some formal properties of the initial model. But this abstraction relies on some informal time-scale separation arguments. We propose to compute conservative bounds on the probabilities on unlikely transitions in the coarse-grained model rather than ignoring them.

Our abstraction of states is history sensitive. In [3], the abstraction of each state of a reaction network also depends on the previous state in its trajectory.

As noticed in [6], refining the sampling intervals further in a discretization process does not necessarily reduce the amount of potential behaviors. This is why a new update policy has been introduced to recover this lack of monotonicity — at the cost of considering more fictitious behaviors. We do not think that non-monotonicity is an issue: when discretizing a system, refining sampling intervals introduces new check-points. Thus it provides the obtained model more opportunities to change its trajectories. We think that it is more important to relate the behaviors of the discretized models to the ones of the initial system, that is to say to ensure that every behavior of the initial system is reflected in the abstraction, would they be some additional fictitious behaviors. This is why, we prove formally that the potential behaviors of each coarse-grain model over-approximate the ones of a reference system (would it be known or not).

Lastly, our goal is also to compare different modeling paradigms in order to understand their underlying assumptions. Building a landscape of different semantics is one of the initial motivations behind the abstract interpretation framework [8].

Outline. The rest of the paper is organized as follows. In Sect. 2, we introduce a unidimensional case study to motivate and illustrate our general framework. Sect. 3 generalizes this approach to coarse-grain arbitrary stochastic reaction networks. In Sect. 4, we apply this framework on a tridimensional case study. We conclude in Sect. 5.

2 First case study: Birth and death model

In order to motivate our framework, we introduce a well-studied unidimensional system: a birth and death (BD) model.

2.1 Reaction network

The BD system is characterized by two reactions having opposite behaviors. Both reactions are given as follows:

$$r_1 \ : \ \emptyset \xrightarrow{k_A} A \qquad\qquad r_2 \ : \ A \xrightarrow{k_{A'}} \emptyset.$$

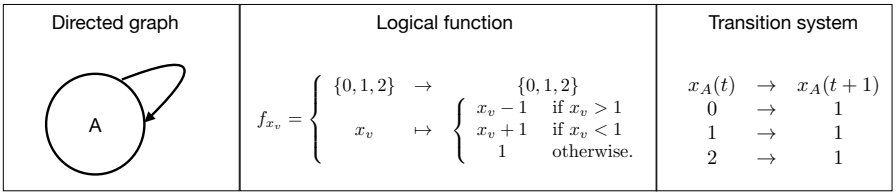

Fig. 1. A logical BD model. A directed graph (left) displays A as a self-regulator, while the logical function (center) is derived to reflect the expected BD system behaviors. The transition system (right) is the result of applying the logical function to a state $x_A \in \{0, 1, 2\}$, each representing an interval of values of molecule A: low (0), medium (1) and high (2). Note that the notion of time is discrete.

The reactions represent production and consumption events of molecules of A. The first reaction, r_1, is a birth event with kinetic constant k_A, while the second, r_2, is a death event with kinetic constant $k_{A'}$. We denote as q the state of the system. The state of the system maps the components of the system to their copy numbers. In this model, A is the unique component.

By assuming stochastic mass-action kinetics law, we can obtain the propensity k_A for the production event and the propensity $k_{A'} \cdot q(A)$ for the death event. Then, the probabilities $\lambda_{r_1}(q)$ and $\lambda_{r_2}(q)$ that the next event is an instance of the reaction r_1 or an instance of the reaction r_2 are defined as follows:

$$\lambda_{r_1}(q) = \frac{k_A}{k_A + k_{A'} \cdot q(A)} \text{ and } \lambda_{r_2}(q) = \frac{k_{A'} \cdot q(A)}{k_A + k_{A'} \cdot q(A)}.$$

A probability for an event type is the ratio between its propensity and the sum of all possible propensities for this system. We can observe that these probabilities are non-constant, as the quantity of A, $q(A)$, may vary.

2.2 Logical model

A logical model can be provided regardless of the exact structure of the reactions and the effective values of the kinetic parameters. Following René Thomas's principles, what matters is the general trend for the evolution of the copy numbers of each kind of components. In our case study, we can indeed distinguish three kinds of states: when the amount of A is such that the state of the system is likely to be stable; when (below this amount) the quantity of A is likely to increase; and when (above this amount) the quantity of A is likely to decrease.

These observations lead to the logical model that is described in Fig. 1. Firstly, a directed graph summarizes the potential regulations (or dependencies) between the components. Here a self arrow on the component A stipulates that the component A auto-regulates itself. We assume that the potential quantities of A are partitioned into three intervals, denoted as 0, 1, and 2. They stand respectively for below the steady state (low / (0)), for around the steady state (medium / (1)), and for above the steady state (high / (2)). Consequently, we derive a logical function that reflects how the system is expected to evolve: below

the steady state, the amount of A increases; around the steady state, it remains constant; above the steady state, it decreases. Then, the logical function induces a transition system that captures the integrated process. Note, however, that at this level of abstraction stochastic fluctuations cannot be observed since reversible interval transitions are not permitted. Thus, the logical model is capable of capturing only the most expected behaviors.

2.3 Formal derivation of a coarse-grained model

Now that a logical model for the BD system has been proposed, we would like to compare its behaviors to those obtained by a formal discretization. We use the same BD reaction network and formally discretize its state space. Then we show that it is possible to restore information on probabilities in this new model.

A common discretization method uses a non-overlapping interval schema. This means that the state space of chemical values are partitioned into intervals that do not share any common values. For example, in the logical BD model interval partitioning are qualitative states with implicit meaning. We can obtain a similar representation in the formally derived model by explicitly choosing a sequence of contiguous intervals (with no intersecting values). It is worth noting that we expect this new model to cope with many more behaviors than the logical model. A question then rises as to whether these behaviors are consequence of the imprecision of the formal abstraction, or whether they reveal important behaviors that are missing in the logical models.

To answer this question, we use information about the stochastic behaviors of the reaction network to recover the probabilities to navigate between intervals (see Fig. 2). More precisely, when entering an interval, we compute the probability that the process will cross this interval or go back to the previous one. It is worth noting that whether an interval is entered from below or from above matters. So we duplicate each interval accordingly. We expect to observe the convergence of the process towards the interval containing the steady state, which is a stable behavior observed by a system. This is the main qualitative behavior of the reaction network and it should be reflected in the discretize model independently of the choices of the discretization intervals.

We call the transitions between intervals macro-transitions. Macro-transition probabilities are computed as follows. Given an interval with lower bound $l \in \mathbb{N}$ and upper bound $u \in \mathbb{N}$, we consider for every state q such that $l \leq q(A) \leq u$, the probability $P(q)$ such that the quantity of A will reach the upper bound u before reaching the lower bound l, knowing that the system starts in the state q. By reasoning on the potential BD events stemming from the state q and their probabilities, we obtain the following relation:

$$P(q) = \begin{cases} 0 \text{ whenever } q(A) = l, \\ 1 \text{ whenever } q(A) = u, \\ \lambda_{r_1}(q) \cdot P([A \mapsto q(A) + 1]) + (1 - \lambda_{r_1}(q)) \cdot P([A \mapsto q(A) - 1]) \text{ otherwise.} \end{cases}$$

As boundary conditions, the probabilities $P([A \mapsto l])$ and $P([A \mapsto u])$ are set respectively to 0 and 1. The last case combines the contribution of two processes: the potential increase in the amount of A with probability $\lambda_{r_1}(q)$ and the potential decrease with probability $1 - \lambda_{r_1}(q)$.

Finite unfolding of the previous recurrence relation converges to a lower bound on the probability $P(q)$. In Sect. 3.3, we discuss how one can exploit recurrence relations to bound an exact probability by an interval of probabilities. Yet, in the BD model, a closed form equation can be derived. The probability $P(q)$ is indeed defined by the following equality:

$$P(q) = \frac{Aux(q(A))}{Aux(u)} \tag{1}$$

where $Aux(j) = \sum_{l \leq s' < j} \left(\prod_{l < s \leq s'} \left(\frac{k_{A'} \cdot s}{k_A} \right) \right)$, for each $j \in \{u, q(A)\}$. We can use Eq. 1 to compute the probability to reach first an upper (and by complement, first a lower) bound.

We show in Fig. 2 the macro-transition system that we derive this way. Underlying each rectangular region are the dynamics stemming from the BD process with kinetic constants $k_A = 20$ and $k_{A'} = 1$. Thus, the intervals displayed are chosen according to the steady state of the system, which is when the quantity of A is equal to 20, or $q(A) = 20$, and in this example it is contained in the interval $[20, 24]$. Each macro-transition is composed of a source interval, an edge labeled with a probability and a target interval. To trigger a macro-transition type, a BD event (birth or death) must push the value $q(A)$ through an interval upper or lower bound value. Thus, it is possible to enter a target interval from below (via an upper bound) or above (via a lower bound). Intervals are duplicated accordingly. The one on the left side of the transition system denotes those entered from below and the one on the right side, those entered from above. Also the exact position of the source (resp. target) of each macro-transition indicates from which border of the interval the macro-transition starts (resp. ends).

We then want to observe whether the trend of the system will proceed upwards or downwards. Since abstraction loses all information about the probabilities of individual reactions, we recover them using Eq. 1. As a result, the probability to exit from the upper bound 19 when starting from the value $q(A) = 15$ is equal to 0.27, and its complement 0.73 is the probability to exit from the lower bound 15. Putting these pieces of information together results in the macro-transition $[15, 19] \xrightarrow{0.73} [10, 14]$ in Fig. 2. Namely, after going up to the interval $[15, 19]$, the system has a higher tendency to return to a lower interval. Note that this behavior is opposite in direction to the stable interval $[20, 24]$. Actually, the general trend of a majority of macro-transitions opposes the direction of the interval containing the steady state point. Mainly this is because it requires only one transition in the initial reaction network to go back to the previous interval, whereas several ones are required to cross an interval entirely. Hence border effects give too much importance to backwards macro-transitions.

To cope with this artifact of the abstraction, we introduce a minimal effort for the system to perform fluctuations between consecutive intervals by using

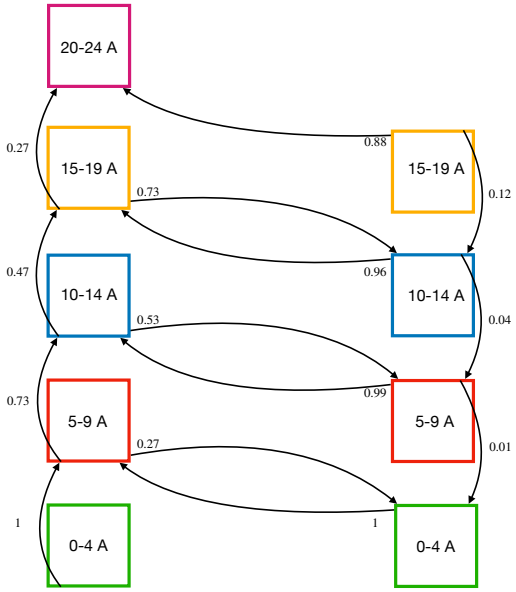

Fig. 2. A non-overlapping macro-transition system for the BD system. Each rectangle is a range of values for the molecule A and a labeled edge is a transition from a source to a target interval. Intervals are duplicated to distinguish whether they are entered from below or above.

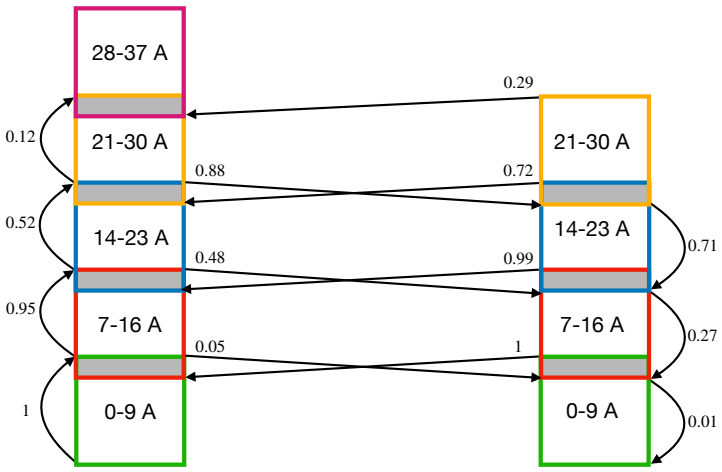

Fig. 3. An overlapping macro-transition for the BD system. The interpretation is similar to the non-overlapping case in 2 with the exception that intervals overlap (denoted by a gray region)

overlapping intervals instead. In Fig. 3, we compute a macro-transition system for the same BD system as in Fig. 2 but with overlapping intervals (overlaps are indicated in gray). The meaning of macro-transitions has to be defined carefully: we consider a macro-transition between a first interval and a second one, only when the system leaves the first interval (hence crossing the overlapping region). We adjust the position of the source and target of the macro-transitions accordingly in the drawing.

Now, the stable interval is $[14, 23]$. Starting from the value $q(A) = 10$, the probability to exit from the upper bound 16 of the interval $[7, 16]$ is equal to 0.95 while the probability to exit from its lower bound 7 is equal to 0.05. Consequently, we obtained the macro-transition $[7, 16] \xrightarrow{0.95} [14, 23]$ which shows the tendency to move towards the stable interval. Similar observations can be made about the other macro-transitions leading to the interval $[14, 23]$.

We notice two major features from our overlapping interval design. Firstly, quantities of molecule A contained in overlapping regions must surpass a buffer region to be able to go back into the previous interval, thus limiting border effects. And, secondly, the general trend of the system reflects a greater likelihood towards the stable interval which was not the case for non-overlapping intervals.

Altogether, we show how our framework is capable of coarse-graining the behaviors emerging from the BD reaction network using a more formal approach. Whereas discretization with non-overlapping intervals was not enough to keep only the likely behaviors as done in the logical models, the use of overlapping intervals introduces a minimal effort for the system to go back after entering an interval. The result is an abstract transition system when unnatural behaviors are assigned low probabilities, hence allowing to quantify the probability of the behaviors that are neglected in the hand-written logical model.

3 General case

In the previous section, we introduced as an example a logical model of the BD system and compared this hand-written model to a formally derived coarse-grained model from the same underlying reaction network. In this section, we generalize this approach to coarse-grain arbitrary reaction networks. More specifically, we build a concrete semantics to capture all the behaviors that may emerge from a system of reactions, and an abstract semantics to approximate these behaviors with intervals (overlapping or not). After which, we bridge the two semantics to restore information on probabilities to the abstract semantics.

3.1 Concrete Semantics

Firstly we define the syntax for reactions and reaction networks.

Definition 1 (Chemical Reaction). *Given a finite set $\mathbb{V}$ of chemical species, a reaction over the set of species $\mathbb{V}$ is defined as a triple $r = (M, V, k)$ such that:*

1. $M : \mathbb{V} \to \mathbb{N}$,

2. $V : \mathbb{V} \to \mathbb{Z}$,
3. $k : \mathbb{V}^{\mathbb{N}} \to \mathbb{R}_{\geq 0}$.

In Def. 1, the function M stands for the (multi-)set of reactants, V denotes the reaction vector (which cumulates the production (positively) and the consumption (negatively) of each chemical species), and k is a function mapping a vector of chemical quantities to a real number which denotes a kinetic term.

Definition 2 (Chemical Reaction Network). *A reaction network R is defined as a pair $(\mathbb{V}, (r_j)_{1 \leq j \leq n})$ such that:*

1. *$\mathbb{V}$ is a set of chemical species;*
2. *$(r_j)_{1 \leq j \leq n}$ is a set of n reactions over the set $\mathbb{V}$ indexed with an integer j between 1 and n.*

For each integer j between 1 and n, the reaction r_j is also denoted as $(M_{r_j}, V_{r_j}, k_{r_j})$.

A chemical state encodes the values of each chemical species.

Definition 3 (Chemical State). *A chemical state is defined as a function $q : \mathbb{V} \to \mathbb{N}$. The set of all the chemical states is denoted as $\mathcal{Q}$.*

Additionally, a chemical state is an input to a kinetic function to obtain a kinetic term for each reaction that involves this state.

For example, we apply our definitions to the reaction network made of both following reactions:

$$r_1 \ : \ A \xrightarrow{k_B} B \qquad\qquad r_2 \ : \ 2A \xrightarrow{k_C} C.$$

Here, the set of chemical species is $\{A, B, C\}$. This reaction network is made of two reactions r_1 and r_2, with respective multiplicity vectors $M_{r_1} = [A \mapsto 1, B \mapsto 0, C \mapsto 0]$ and $M_{r_2} = [A \mapsto 2, B \mapsto 0, C \mapsto 0]$ and with respective reaction vectors $V_{r_1} = [A \mapsto -1, B \mapsto 1, C \mapsto 0]$ and $V_{r_2} = [A \mapsto -2, B \mapsto 0, C \mapsto 1]$. Furthermore, assuming the stochastic mass-action kinetics law, the kinetic functions are defined as $k_{r_1} = [q \mapsto k_B \cdot q(A)]$ for the reaction r_1 and $k_{r_2} = \left[q \mapsto \frac{k_C \cdot q(A) \cdot (q(A)-1)}{2}\right]$ for the reaction r_2.

Until the rest of Sect. 3, we assume that we are given $(\mathbb{V}, (r_j)_{1 \leq j \leq n})$ a generic reaction network that we also denote as R. The set $\{r_1, \ldots, r_n\}$ is also written as $\mathcal{R}$. This is the set of the reactions of the network R.

Furthermore, a system is updated via a reaction application, which is called a chemical transition. This is formalized in the following definition.

Definition 4 (Chemical Transition). *A chemical transition is a triple $(q, r, q') \in \mathcal{Q} \times \mathcal{R} \times \mathcal{Q}$ relating two chemical states $q, q' \in \mathcal{Q}$ by a reaction r such that for all chemical species $v \in \mathbb{V}$:*

1. *$M_r(v) \leq q(v)$*

2. $q'(v) = q(v) + V_r(v)$.

The set of all the chemical transitions is denoted as $\mathcal{T}$.

A chemical transition captures an application of a reaction rule. Criterion 1 ensures that there are enough reactants available for a reaction to occur, while criterion 2 applies a reaction rule to update a predecessor value to obtain a successor value. Additionally, each chemical transition is given a probability. The probability that the transition will be the next one, given the current state, is defined as follows:

Definition 5 (Transition probability). *Let* $(q, r, q') \in \mathcal{T}$ *be a chemical transition. The probability* $\lambda_r(q)$ *for a chemical state* $q \in \mathcal{Q}$ *involved in a chemical transition is defined as:*

$$\frac{k_r(q)}{\sum_{r' \in \mathcal{R}} k_{r'}(q)}.$$

In Def. 5 the probability that a given chemical transition is applied next, is equal to the ratio of its kinetic term to all kinetic terms of the reaction network.

For example, an instantiation of this definition used on the previous reaction network results in the following transition probabilities:

$$\lambda_{r_1}(q) = \frac{2 \cdot k_B \cdot q(A)}{q(A) \cdot (2 \cdot k_B + k_C \cdot (q(A) - 1))} \text{ and } \lambda_{r_2}(q) = \frac{k_C \cdot q(A) \cdot (q(A) - 1)}{q(A) \cdot (2 \cdot k_B + k_C \cdot (q(A) - 1))}$$

whenever $q(A) > 0$. Note that when $q(A) = 0$, no reaction is enabled and the system is deadlocked.

Starting from an initial chemical state, it is possible to chain chemical transitions. This leads to the notion of a chemical trace.

Definition 6 (Chemical Trace). *A trace of length* $k \in \mathbb{N}$ *is a pair* $(q'_0, ((q_i, r_i, q'_i), \mu_i)_{1 \leq i \leq k}) \in \mathcal{Q} \times (\mathcal{T} \times [0, 1])^k$ *that satisfies both conditions:*

1. *for every integer* i *between* 0 *and* $k - 1$*, we have* $q'_i = q_{i+1}$*;*
2. *for every integer* i *between* 1 *and* k*, we have* $\mu_i = \lambda_{r_i}(q_i)$.

Such a trace is usually written as $q_1 \xrightarrow[\mu_1]{r_1} \ldots \xrightarrow[\mu_k]{r_k} q'_k$.

The set of all the chemical traces of a reaction network defines all the potential long-term behaviors of its underlying system. Given an initial chemical state $q'_0 = [A \mapsto 6, B \mapsto 0, C \mapsto 0]$ and the kinetic constants $k_B = 20$ and $k_C = 1$, an example of a chemical trace for the previous reaction network is given as follows:

$$(6,0,0) \xrightarrow[0.8]{r_1} (5,1,0) \xrightarrow[0.83]{r_1} (4,2,0) \xrightarrow[0.13]{r_2} (2,2,1) \xrightarrow[0.95]{r_1} (1,3,1) \xrightarrow[1]{r_1} (0,4,1),$$

where a state q is denoted as the triple $(q(A), q(B), q(C))$. At the end of this trace, no transition is available.

Finally, the following definition associates a probability to each chemical trace.

Definition 7. *Let $(q'_0, ((q_i, r_i, q'_i), \mu_i)_{1 \leq i \leq k})$ be a chemical trace that we denote as τ. The probability $P(\tau \mid q'_0)$ of the chemical trace τ, knowing that the system starts in the state q'_0 is defined as $\prod_{1 \leq i \leq k} \mu_i$.*

For example, the probability for the previous trace is: $P(\tau \mid (6, 0, 0)) = 0.08$ (since $0.80 \cdot 0.83 \cdot 0.13 \cdot 0.95 \cdot 1 = 0.08$).

3.2 Abstract Semantics

The goal of the abstract semantics is to over-approximate the behaviors emerging from chemical reaction networks. Namely, it is obtained by sampling the value domains by the means of a set of intervals.

Definition 8 (Intervals). *We consider a family $\left(\underline{q}^\sharp_p, \overline{q}^\sharp_p\right)_{1 \leq p \leq n}$ of n pairs of values in $\mathbb{N} \cup \{+\infty\}$ (where n is a natural number in $\mathbb{N}$) such that both of the following properties are satisfied:*

1. *for every natural number p between 2 and n, $\underline{q}^\sharp_{p-1} < \underline{q}^\sharp_p \leq \overline{q}^\sharp_{p-1} < \overline{q}^\sharp_p$;*
2. *$\overline{q}^\sharp_n = +\infty$.*

We denote by $D^\sharp$ the set of intervals $\{\left(\underline{q}^\sharp_p, \overline{q}^\sharp_p\right) \mid 1 \leq p \leq n\}$.

An interval $\left(\underline{q}^\sharp_p, \overline{q}^\sharp_p\right)$ denotes the set of values $\{k \in \mathbb{N} \mid \underline{q}^\sharp_p \leq k < \overline{q}^\sharp_p\}$. There are finitely many of them. Each of them is well-formed. Their lower bounds form an increasing sequence, as well as their upper bounds. Also every natural number occurs in at least one of them.

Conversely, an abstraction of a value is an interval in the domain $D^\sharp$ that contains this value. There may be several such intervals. To decide which one, the abstraction function is parameterized by a context made of a reference interval. The following definition specifies that the so-contextualized abstraction function selects the interval nearest to the reference one among the potential ones.

Definition 9 (Value Abstraction Function). *Let $(\underline{q}^\sharp_{p_\star}, \overline{q}^\sharp_{p_\star})$ be an interval in $D^\sharp$. The value abstraction function $\beta^{\mathcal{D}}_{(\underline{q}^\sharp_{p_\star}, \overline{q}^\sharp_{p_\star})} : \mathbb{N} \to D^\sharp$ maps each value $k \in \mathbb{N}$ to the unique interval $(\underline{q}^\sharp_p, \overline{q}^\sharp_p) \in D^\sharp$, such that both following properties are satisfied:*

1. *$\underline{q}^\sharp_p \leq k < \overline{q}^\sharp_p$;*
2. *for any $(\underline{q}^\sharp_{p'}, \overline{q}^\sharp_{p'}) \in D^\sharp$ such that $\underline{q}^\sharp_{p'} \leq k < \overline{q}^\sharp_{p'}$, we have: $|p_\star - p| \leq |p_\star - p'|$.*

Def. 9 is well-formed thanks to the hypotheses in Def. 8. More precisely, the existence of an interval in $D^\sharp$ containing a given value follows from the fact that the elements of $D^\sharp$ forms a covering of $\mathbb{N}$, then thanks to the monotonicity of the lower and upper bounds of the intervals, the set of intervals that contain a given value are contiguous elements in the domain. It follows the uniqueness of the interval that is the closest to the reference interval.

Now we lift the notions of values and value abstraction to all the chemical species of a reaction network.

Definition 10 (Abstract State). *An abstract state is a function* $q^\sharp : \mathbb{V} \to D^\sharp$. *The set of all abstract states is denoted* $\mathcal{Q}^\sharp$.

An abstract state contains all the interval values approximating the quantities of each chemical species.

Definition 11 (State Abstraction Function). *Let* $q^\sharp_*$ *be an abstract state in* $\mathcal{Q}^\sharp$. *The abstract state function* $\beta^{\mathcal{S}}_{q^\sharp_*} : \mathcal{Q} \to \mathcal{Q}^\sharp$ *maps each chemical state* q *to the abstract state* $\left[v \in \mathbb{V} \mapsto \beta^{\mathcal{D}}_{q^\sharp_*(v)}(q(v)) \in D^\sharp\right]$.

Similarly to the value abstraction function, the state abstraction function is parameterized by a reference abstract state. An equivalent definition can be obtained by interpreting each abstract state as the box delimited on each chemical species in $\mathbb{V}$ by its corresponding intervals, and then by abstracting each concrete state by the unique box that contains this concrete state and that is at minimal Gaussian distance from the reference abstract state.

It is worth noting that we have used the same family of intervals to abstract the quantity of every chemical species. Using different families of intervals is also possible and it would have raised no further technical difficulties. Indeed, it would have been even useful in practice since there is no reason why the quantity of each component of the system should be abstracted the same way. Yet making this simplification deeply lighten the presentation of the framework and this is why we have proceeded this way.

We can now define the abstraction of a chemical trace. Each abstract trace is obtained by lifting point-wise the state abstraction function to each chemical state along a chemical trace, taking respectively for reference the previous abstract state. For the moment, we discard probabilities. Restoring information about the probabilities of abstract transitions is the purpose of Sect. 3.3.

Definition 12 (Abstract Trace). *An abstract trace is an element of the set* $\mathcal{Q}^\sharp \times (\mathcal{Q}^\sharp \times \mathcal{R} \times \mathcal{Q}^\sharp)^\star$.

Definition 13 (Trace Abstraction Function). *The trace abstraction function* $\beta^{\mathcal{T}}$ *maps each chemical trace* $(q'_0, ((q_i, r_i, q'_i), \mu_i)_{1 \le i \le k})$ *to the abstract trace that is defined inductively as follows:*

1. $\beta^{\mathcal{T}}(q'_0, ()) = (\beta^{\mathcal{S}}_{q^\sharp_0}(q'_0), ())$;
2. *By induction, if* $\beta^{\mathcal{T}}(q'_0, ((q_i, r_i, q'_i), \mu_i)_{1 \le i < k}) = (q^{\sharp\prime}_0, (q^\sharp_i, r^\sharp_i, q'_i)_{1 \le i < k})$, *the abstract trace* $\beta^{\mathcal{T}}(q'_0, ((q_i, r_i, q'_i), \mu_i)_{1 \le i \le k})$ *is defined as* $(q^{\sharp\prime}_0, (q^\sharp_i, r^\sharp_i, q'_i)_{1 \le i \le k})$ *where* $(q^\sharp_k, r^\sharp_k, q^{\sharp\prime}_k) = (q^{\sharp\prime}_{k-1}, r_k, \beta^{\mathcal{S}}_{q^{\sharp\prime}_{k-1}}(q'_k))$.

where $q^\sharp_0$ *is the abstract state mapping every component to the interval* $(\underline{q}^\sharp_0, \overline{q}^\sharp_0)$.

The trace abstraction function starts by abstracting the initial state of a chemical trace by using the abstract state $q^\sharp_0$ as a reference; then it abstracts

each chemical transition by abstracting the successor chemical state of the corresponding chemical transition while referencing the last encountered abstract state. For example, we apply our definition on the following chemical trace:

$$3 \xrightarrow[0.87]{r_1} 4 \xrightarrow[0.83]{r_1} 5 \xrightarrow[0.20]{r_2} 4 \xrightarrow[0.83]{r_1} 5 \xrightarrow[0.20]{r_2} 4 \xrightarrow[0.83]{r_1} 5 \xrightarrow[0.80]{r_1} 6$$

for the BD model introduced in Sect. 2.3 with several interval samplings.

With the following choice of non-overlapping intervals: $((0,4),(5,9))$, we obtain the following abstract trace:

$$(0,4) \xrightarrow{r_1}{}^{\sharp} (0,4) \xrightarrow{r_1}{}^{\sharp} (5,9) \xrightarrow{r_2}{}^{\sharp} (0,4) \xrightarrow{r_1}{}^{\sharp} (5,9) \xrightarrow{r_2}{}^{\sharp} (0,4) \xrightarrow{r_1}{}^{\sharp} (5,9) \xrightarrow{r_1}{}^{\sharp} (5,9),$$

whereas with the following choice of overlapping intervals: $((0,5),(2,7))$, we obtain the following one:

$$(0,5) \xrightarrow{r_1}{}^{\sharp} (0,5) \xrightarrow{r_1}{}^{\sharp} (0,5) \xrightarrow{r_2}{}^{\sharp} (0,5) \xrightarrow{r_1}{}^{\sharp} (0,5) \xrightarrow{r_2}{}^{\sharp} (0,5) \xrightarrow{r_1}{}^{\sharp} (0,5) \xrightarrow{r_1}{}^{\sharp} (2,7).$$

We notice that with the second choice, the system remains in the first interval until finally exiting via its final chemical state of the chemical trace. By emphasizing the effort to go back and forth between consecutive intervals, the abstraction has abstracted away the fluctuations. This is not the case with the first choice of intervals. Please observe that no concrete behavior has been lost in the process. Then, we will see in Sect. 3.3, how to recover some information about the probabilities of the behaviors of the initial chemical networks. In that context, not only, as in a non-deterministic setting, any potential behavior in the concrete is reflected in the abstraction, but also the probability attached to each potential concrete behavior is over-approximated in the abstraction: transitions between abstract regions come with an upper bound on their probability to occur. In particular, in case of a rare event that may occur in the concrete only at a very small probability, by construction, this event is taken into account in the abstraction, but its likelihood may be over-estimated.

Furthermore, each choice of sampling intervals comes with a different interpretation for the abstract traces, which may highlight or hide different information accordingly. When the initial system is complex, the process of interval parameterization may proceed heuristically. Our framework provides a tool to tune granularity: too coarse interval abstractions can mask the underlying dynamics; however, upon refining the intervals one will be able to identify the behavioral trend of a system. In practice, it is enough to pick intervals that are fine-grained enough to separate the main regimes of the system.

3.3 Recovering information about transition probabilities

In Sect. 3.3, we refine the abstract semantics with some quantitative information to compare the likelihood of transitions between abstract states, the macro-transitions. This basically means that knowing that the system has just entered a new abstract state, and a given pair of potential macro-transitions, we would

like to know with which probability the first macro-transition (in the pair) will occur before the second one.

By construction of our abstract domain, macro-transitions are triggered when a given chemical species reaches a particular copy number. We introduce target regions accordingly in the following definition.

Definition 14 (Target region). *A target region is a set of concrete states of the form* $\{q \in \mathcal{Q} \mid q(v) \sqsupseteq b\}$*, where* v *is a chemical species in* $\mathbb{V}$*,* $\sqsupseteq$ *a binary relation in the set* $\{\leq, \geq\}$*, and* b *a natural number in* $\mathbb{N}$*.*

This target region is denoted as $g_{v,\sqsupseteq,b}$*.*

Until the end of Sect. 3.3, we consider $q_\bullet \in \mathcal{Q}$ a state, $\mathcal{G}$ a set of target regions, and g a specific target region in the set $\mathcal{G}$. We want to define, the probability that the system when starting from the state $q_\bullet$, will enter the specific region g before entering any other target regions of the set $\mathcal{G}$. In order not to overcount the probabilities, we cut chemical traces as soon as they enter the region g and we ignore the traces that enter another region in the set $\mathcal{G}$ before.

Definition 15 (Minimum successful traces). *We denote as* $\chi_{(q_\bullet,\mathcal{G},g)}$ *the set of the chemical traces* $(q'_0, ((q_i, r_i, q'_i), \mu_i)_{1\leq i\leq k})$ *such that the following conditions are satisfied:*

1. $q'_0 = q_\bullet$*;*
2. $q'_k \in g$*;*
3. $\forall i \in \mathbb{N}$*, such that* $0 \leq i < k$*,* $q'_i \notin \bigcup \mathcal{G}$*.*

In Def. 15, the set $\chi_{(q_\bullet,\mathcal{G},g)}$ contains all the chemical traces that start in the state $q_\bullet$ (Cond. 1), reach the target region g in their final state (Cond. 2), and have reached no other target regions before (Cond. 3).

We are now ready to integrate the probability of minimal successful traces.

Definition 16 (Probability to reach a specific goal first). *The probability* $P_g^{\mathcal{G}}(q_\bullet)$ *that the system reaches the target region* g *before any other target regions in* $\mathcal{G}$ *when starting in the state* $q_\bullet$ *is defined as:* $\sum_{\tau \in \chi_{(q_\bullet,\mathcal{G},g)}} P(\tau \mid q_\bullet)$*.*

Namely, the probability to reach a specific goal first is computed by summing the probability of all corresponding minimum successful traces.

The following proposition provides an easier way to compute this probability.

Proposition 1 (Inductive definition). *The probabilities* $P_g^{\mathcal{G}}(q)$ *for every state* $q \in \mathcal{Q}$ *are related by the following three conditions:*

1. $P_g^{\mathcal{G}}(q) = 1$ *whenever* $q \in g$*;*
2. $P_g^{\mathcal{G}}(q) = 0$ *whenever* $q \in \bigcup \mathcal{G} \setminus g$*;*
3. $P_g^{\mathcal{G}}(q) = \sum_{q \xrightarrow{r_i} q'} \lambda_{r_i}(q) \cdot P_g^{\mathcal{G}}(q')$ *whenever* $q \notin \bigcup \mathcal{G}$*.*

Prop. 1 provides an iterative scheme that computes for every state q a sequence of values that converges from below to the value of $P_g^{\mathcal{G}}(q)$. By complementing, we can also obtain an upper bound to this probability.

We can go further by the means of matrix computations. We consider one dimension for each potential state. Each function that maps states to real numbers is interpreted as a vector, whereas each function that maps pairs of states to real numbers is interpreted as a matrix. We define the vector B, and both matrices I and A as follows:

$$B(q) = \begin{cases} 1 & \text{whenever } q \in g, \\ 0 & \text{otherwise;} \end{cases} \qquad I(q,q') = \begin{cases} 1 & \text{whenever } q = q', \\ 0 & \text{otherwise;} \end{cases}$$

$$A(q,q') = \begin{cases} 0 & \text{when } q \in \bigcup \mathcal{G}, \\ \sum_{q \xrightarrow{r_i} q'} \lambda_{r_i}(q) \cdot P_g^{\mathcal{G}}(q') & \text{otherwise.} \end{cases}$$

Then, the sequence $(X_k)_{k \in \mathbb{N}}$ of vectors that is defined by:

1. $X_0 = B$;
2. $X_{k+1} = A \cdot X_k + B$ for every $k \in \mathbb{N}$;

converges component-wise to the probability $P_g^{\mathcal{G}}(q)$. It follows that $P_g^{\mathcal{G}} = (\sum_{j \in \mathbb{N}} A^j) \cdot B$. Or even, $P_g^{\mathcal{G}} = (I - A)^{-1} \cdot B$, whenever the matrix $(I - A)$ is invertible.

The computation of the probabilities $P_g^{\mathcal{G}}(q)$ can be proceeded by using any available linear algebra library. This is indeed what the model checker PRISM [17,14] is doing. Yet, having unfolded the computation offers several advantages. For instance, in our setting, the probabilities can be approximated from below by finitely approximating the formal expansion of the sums of the powers of the sparse matrix A. Secondly, when dealing with high dimensional models, expressions with scalar coefficients can be symbolically simplified into expressions over interval coefficients [18] in order to eliminate some dimensions and to tune the trade-off between accuracy and efficiency. This would not be possible with a black box approach.

In Sect. 3, we have refined our non-deterministic abstract semantics with probabilities, hence providing information about the general trend for the dynamics of models. In Sect. 4, we apply this approach on a case study taken from [2], which consists of two reactions competing for a common resource at different time-scales according to the availability of this resource. In [2] a precise coarse-grained system has been derived, but at the cost of neglecting some slow reactions. We would like to assess this assumption from a formal perspective.

4 Second case study: Competition for resources

In Sect. 2, we compared a logical model and a formally discretized one of the BD process. In this section, we will follow a similar strategy for a model of a system of reactions which compete for a common resource taken from [2].

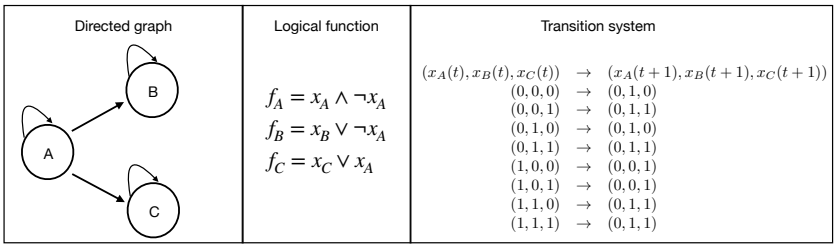

Fig. 4. A logical model for a system of resource competition. A directed graph (left) displays how molecule B and C have a common regulator, molecule A. Moreover, each kind of molecules auto-regulates it-self. Each logical function (center) is a Boolean rule which reflects the update scheme for a given chemical species. They can take values in the domain $\{0, 1\}$. The value 0 stands for low, 1 for high. The transition system (right) reflects the system dynamics from the logical functions.

4.1 Reaction network

The second case study is composed of two reactions:

$$r_1 \ : \ A \xrightarrow{k_B} B \qquad\qquad r_2 \ : \ 2A \xrightarrow{k_C} C$$

where two chemical species, B and C are produced and each consume a common resource, A. In reaction r_1, a quantity of B is produced with a kinetic constant k_B. In reaction r_2, a quantity of C is produced with kinetic constant k_C. The production of molecule B consumes a quantity of A, while C requires two.

As in the first case study, we assume stochastic mass-action kinetics law to obtain the propensities $k_B \cdot q(A)$ and $\frac{k_C \cdot q(A) \cdot (q(A)-1)}{2}$ for, respectively, the reactions r_1 and r_2. Consequently, we derive a probability function for each reaction:

$$\lambda_{r_1}(q) = \frac{2 \cdot k_B \cdot q(A)}{q(A) \cdot (2 \cdot k_B + k_C \cdot (q(A)-1))} \text{ and } \lambda_{r_2}(q) = \frac{k_C \cdot q(A) \cdot (q(A)-1)}{q(A) \cdot (2 \cdot k_B + k_C \cdot (q(A)-1))}.$$

Contrary to the first case study, this reaction system does not have reversible reactions and the quantities of B and C are strictly increasing up to the point at which all resources become depleted.

4.2 Logical model

As in the first case study, we can write by hand a logical model for this second example. The property of interest is the competition between the production of the molecules B and C. Namely, depending on the quantity of the resource A, either B or C is produced more abundantly. When the quantity of A is under a certain value, the quantity of B increases more; and when it is above this value, the quantity of C increases more.

Using this knowledge we obtain the logical model that is described in Fig. 4. The directed graph shows different kinds of regulations. Firstly, A regulates B

and C because it may be consumed to produce them. Secondly, each component auto-regulates itself since whatever it increases or decreases, its quantity at the next time step depends on the one at the current state. Lastly, A auto-regulates itself negatively (since it is consumed to produce B and C). We can abstract quantities by Boolean values. Namely, the state is a triple of Boolean variables $(x_A, x_B, x_C) \in \{0,1\}^3$, here 0 stands for low quantity and 1 for high quantity. The logical (Boolean) functions are derived to capture the main feature of the reaction network. A gets consumed, while either B or C is produced according to the qualitative abundance of A.

Several update policies exist to define the operational semantics. Our transition system is derived by assuming the synchronous one, where the value of each chemical species is updated at each time step. It induces the eight transitions that are described in Fig. 4. It is worth mentioning that similar results can be achieved in the asynchronous mode by taking into account mass preservation of invariants and using priorities [2]. The model indicates that starting with a low amount of the molecule A, the system produces some B, but no C; whereas with a high amount, the system produces firstly some C, then some B. One may wonder whether more behaviors may occur in the underlying reaction system that have been discarded by the simplification into a logical model.

4.3 Formal discretization of the reaction network

This motivates the formal discretization of the reaction network with overlapping intervals to compare the behaviors of the so-obtained model to the ones of the logical model. Specifically, we wonder whether our framework is capable of highlighting both main behaviors (production of B, or production of C followed by production of B), and provides low upper bounds for the probability of behaving differently.

Firstly, in order to simplify the computation, we would like to eliminate the variable $q(A)$ which stands for the quantity of A in the system. We denote as $q \in \mathcal{Q}$ the chemical state, which contains copy numbers of molecules A, B, and C. The system is constrained by the following mass invariant $q(A) = q_0(A) - (q(B) - q_0(B)) - 2{\cdot}(q(C) - q_0(C))$, where q_0 stands for the initial state and is fixed once for all. We can safely replace each occurrence of the variable $q(A)$ with the right hand side of this equality in any expression.

As in the first case study, the domain of values can be sampled into overlapping intervals. This way, chemical states are gathered into rectangular regions (we are left with only two dimensions since we have eliminated the quantity of A), that are called abstract states. Our goal is then to derive some quantitative information about the macro-transitions in the so-obtained discretized model. More specifically, when a chemical state enters a new abstract state, the goal is to compute the probability that the system will cross the corresponding rectangular region and exit along the same axis, or via the alternate axis. Note that when entering a new abstract state, we do not know precisely the chemical state of the system. Thus, any potential position on the entering side must be considered (e.g., the system may be arbitrary close to the corner of the rectangular

region so that exiting the rectangle via the alternate axis may require only one step in the concrete). Consequently, in order to retain a minimal effort strategy, we do not consider the next consecutive interval in the alternate axis, but the subsequent one. For instance, when an event drives molecule B into a new abstract interval, we consider as target goals the next consecutive interval for molecule B and the next two consecutive abstract values for molecule C, since we do not know precisely the concrete amount of C). The initial abstract state receives particular treatment: we can safely compute by which rectangular face a chemical state will exit, since the initial state is known perfectly.

The general framework described in Sect. 3 provides for any pair of thresholds for the quantities of molecules B and C and for any chemical state $q \in \mathcal{Q}$, the probability that the quantity of the molecule B reaches its threshold before the molecule C, and conversely. We denote by (m_B, M_B) (resp. (m_C, M_C)) an interval for the quantity of the molecule B (resp. C). We introduce $P^{\mathcal{G}}_{g_1}(q)$ (resp. $P^{\mathcal{G}}_{g_2}(q)$) as the probability for a chemical state where the quantity of B (resp. C) reaches the threshold M_B (resp. M_C) before that the quantity of C (resp. B) reaches the threshold M_C (resp. M_B) when starting from a state with $q(B)$ and $q(C)$ instances of the molecule B and C. Therefore, the probability $P^{\mathcal{G}}_{g_1}(q)$ satisfies the following relation:

$$P^{\mathcal{G}}_{g_1}(q) = \begin{cases} 1 \text{ whenever } q(B) = M_B, \\ 0 \text{ whenever } q(B) < M_B \text{ and } q(C) = M_C, \\ \lambda_{r_1}(q) \cdot P^{\mathcal{G}}_{g_1}(q') + \lambda_{r_2}(q) \cdot P^{\mathcal{G}}_{g_1}(q'') \text{ otherwise.} \end{cases} \tag{2}$$

for every $q(B), q(C) \in \mathbb{N}$, such that $m_B \leq q(B) \leq M_B$ and $m_C \leq q(C) \leq M_C$. A similar expression can be obtained for $P^{\mathcal{G}}_{g_2}(q)$ by switching the base cases for the alternate axis. First two cases stand for the boundary conditions (where thresholds are reached) whereas the third cases captures an increase in molecule B with probability $\lambda_{r_1}(q)$ or C with probability $\lambda_{r_2}(q)$.

As seen in Sect. 3.3, in general, Eq. 2 can be computed exactly by means of inverting a matrix (or equivalently solving a linear system of equations) or approximated, from below, by using a finite expansion of the sequence of the powers of a sparse matrix. Here, since the quantities of the molecules B and C never decrease, the recurrence relation can be solved exactly (up to rounding errors) in a finite amount of iterations.

In the logical version of the reaction network, the unlikely behaviors when C is produced at low abundance of A and when B is produced at high abundance of A have been discarded. We thus test our framework in capturing a low upper bound on the probability of the corresponding macro-transitions in the formal discretization of the underlying reaction network. As a result, in Fig. 5, we computed a macro-transition system for two scenarios: when the copy number of A is low (Fig. 5, left) or high (Fig. 5, right). As kinetic constant, we took $k_B = 20$ and $k_C = 1$ and parameterize intervals to reflect the system steady state, which is again when $q(A) = 20$. We tune the initial values $q_0(A)$ respectively to 15 and 100. In Fig. 5, a rectangular region represents an abstract state and is composed each of a respective range of values for molecules B and C. A labeled edge

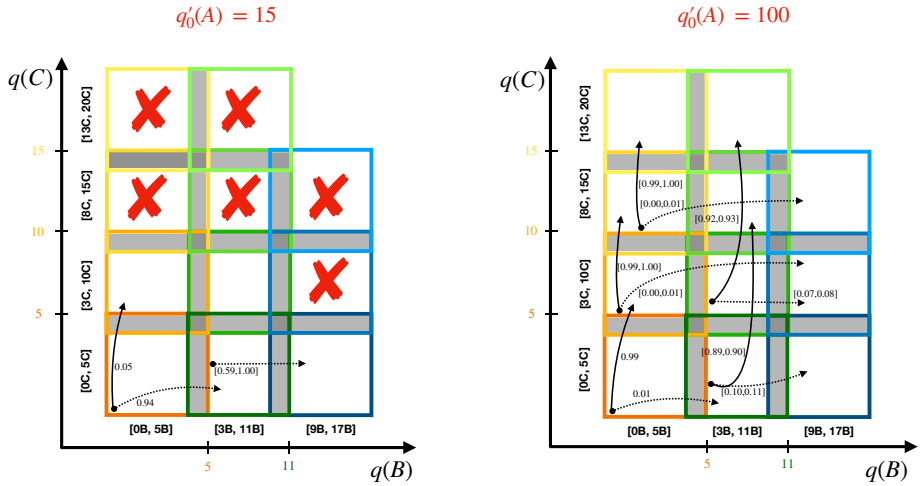

Fig. 5. The derived coarse-grained transition systems for different initial quantities of the molecule A. In the first case (left), the initial amount of A is equal to 15 whereas it is equal to 100 in the second case (right). The states of both systems are related by edges with one source and two targets, with the following meaning: upon entering a new state (the source), the range of probabilities reflects the probability to reach a target before the other one.

connects a source abstract state to a target abstract state. Each time, a dashed edge, that describes an increase in the quantity of the molecule B, competes with a solid edge, that describes an increase in the quantity of the molecule C.

Let us take an example by considering, in the first scenario, when $q_0(A) = 15$, the sequence of chemical reactions that starting from the initial state (where $q(B) = 0$ and $q(C) = 0$) drives the system out of the region $([0B, 5B], [0C, 5C]$ into the region $([3B, 11B], [0C, 5C]$. To recover the probability for this macro-transition, by Eq. 2 we can compute the probabilities to reach each next target abstract state. As such, the probability to exceed, from the initial state, the upper bound $q(B) = 5$ (resp. upper bound $q(C) = 5$) first is equal to 0.94 (resp. 0.05). The remaining 0.01 probability corresponds to the case where the system reaches the state where $q(B) = 5$ and $q(C) = 5$ (that is to say that the molecule A is completely depleted before having left the initial abstract state). This describes precisely the directional tendency of the behavior of the underlying reaction network. Combining these elements results in the macro-transition $([0B, 5B], [0C, 5C]) \xrightarrow{0.94} ([3B, 11B], [0C, 5C])$ (indicated by a dashed edge on the left in Fig. 5). Given a low resource environment, this macro-transition highlights the tendency towards a regime where the molecule B is created abundantly. Additionally, macro-transitions towards creation of the molecule C either occur with low probability or are not possible due to limited resources (indicated by red crosses).

In the second scenario (Fig. 5, right) we tune the initial resource pool to $q_0(A) = 100$ and retain the same kinetic conditions. We immediately observe that there are many more macro-transitions than in the first scenario, a consequence to the abundance in the common resource. As an example, we detail the computation for the bounds on the probability of the macro-transition $([0B, 5B], [3C, 10C]) \to ([0B, 5B], [8C, 15C])$, when the region $([0B, 5B], [3C, 10C])$ has been entered from below (i.e. by increasing the abundance of C). Before starting any computation, it is worth noting that when entering the region $([0B, 5B], [3C, 10C])$ from below, $q(B)$ ranges arbitrarily between 0 and 5, while $q(C)$ is equal to 6. Thus in the computation of the probabilities of macro-transitions we have to consider any potential value for $q(B)$ between 0 and 5. Then we compute the minimal probability that the quantity of the molecule C exceeds 10 before the quantity of the molecule B exceeds the quantity 11. By applying Eq. 2, we obtain 0.99 (it is indeed obtained when entering the region for $([0B, 5B], [3C, 10C])$ with the state $[B \mapsto 5, C \mapsto 6]$, the maximal probability that we obtain is 1.00 (when entering this region with the state $[B \mapsto 0, C \mapsto 6]$). Indeed the value is not exactly 1 but it is conservatively rounded to 1 because of floating point arithmetics. This highlights that the molecule C is abundantly created with very high probability at the begin of the system execution, and then eventually some B is synthesized.

Reflecting on the two scenarios, it becomes clear that one can bound accurately the probabilities on the likelihood for macro-transitions. Our coarse-graining approach has the following benefits. Firstly, our framework provides a means to compute lower and upper bounds on the probabilities of the transitions between abstract states by formally relating the semantics of reaction networks to its abstract counterpart. Hence providing formal confidence in the formally derived discretized models. Secondly, by enforcing a minimal effort for the system to perform any transition between abstract states, we were able to observe the expected dynamics, the same as in the logical model but without relying on arguments on concentration- and time-scale separation. This has been obtained by twisting the abstraction function instead. This way, the interpretation of the behaviors of the abstract model is less intuitive, but it is still rigorously formally specified. Finally, it is possible to assess the validity of logical models by providing upper bounds to the probabilities of the transitions that has been discarded during the modeling process without any formal justification.

5 Conclusion

We have proposed a generic framework to coarse-grain stochastic reaction networks by sampling the quantity of each kind of molecules within a set of intervals. Instead of neglecting unlikely transitions between abstract regions of states, we compute conservative bounds on their probability. Our goal is indeed to check whether we can derive as accurate models as hand-written ones while ensuring a formal relationship between the potential behaviors in the initial and the

derived models. We expect to gain new insights to understand the underlying assumptions behind logical modeling.

Getting formal — but accurate — coarse-grained models requires a specific treatment of boundary effects. It is indeed important not to amplify the importance of some unlikely behaviors. In particular we ensure that every chemical transition between abstract regions of states corresponds to a minimal number of steps. For transitions induced by reverse reactions, we use overlapping intervals. When a chemical state can be arbitrary close to the boundary of an interval, we examine the capacity to cross the next interval instead of just entering it. This induces a non-standard interpretation of the dynamics of the coarse-grained systems. Fewer trajectories are considered in the abstract, while soundness is still ensured (by construction). As in a non-deterministic setting, every concrete behavior is reflected in the abstract, but additionally an upper bound on their probability is computed. Hopefully, rare events are assigned a small upper bound on their probability to occur.

An additional advantage of our framework is the ability to perform and combine numerical abstractions, such as finite expansions of infinite increasing series and include their overall impact as a unique bound on the numerical errors made on the computation of probability values. However, scaling the framework to more complex systems would require one to formally parse intricate relations between numerous variables. For example, to deal with higher dimensional models, symbolic simplification of expressions [18] is possible. Another avenue of thought is to use exact model reduction methods based on the structure of the components of an initial reaction network [12]. Yet, in practice, exact model reduction techniques are not very efficient, especially in a stochastic setting [13]. Still, in our context, we can be more optimistic since, on the one hand not all the properties of the underlying stochastic system have to be preserved and because on the other hand, we can admit numerical approximations on probability values as any other sources of numerical imprecision and include them in the computation of sound over-approximations.

In our paper, we have dealt only with very small case studies. Our motivation was to be able to explain them thoroughly and to focus on minimal difficulties that occur pervasively in models. Ideally we would like to target bigger — but still reasonable — models such as the one for the early events of the EGFR cascade presented in [5]. These models already cope with around three hundred kinds of molecular species. To scale up to this kind of model, we will restrict our study to the competition between pairs of macro-transitions. Yet special care will have to be taken to deal with the denominator of probability functions. These denominators involve the sum over the propensities of each potential event which make them particularly tricky to abstract. Instead of using numerical approaches, we plan to use marginalization to isolate independent subnetworks and reduce the number of terms in denominators accordingly. Yet, here again perfectly independent reaction sub-networks are very unlikely to occur, thus we plan to propose a relaxed version, at the cost of including an additional component in the computation of bounds of probability values.

References

1. Abou-Jaoudé, W., Feret, J., Thieffry, D.: Derivation of qualitative dynamical models from biochemical networks. In: Roux, O.F., Bourdon, J. (eds.) Computational Methods in Systems Biology - 13th International Conference, CMSB 2015, Nantes, France, September 16-18, 2015, Proceedings. Lecture Notes in Computer Science, vol. 9308, pp. 195–207. Springer (2015). https://doi.org/10.1007/978-3-319-23401-4_17
2. Abou-Jaoudé, W., Thieffry, D., Feret, J.: Formal derivation of qualitative dynamical models from biochemical networks. Biosystems **149**, 70–112 (2016). https://doi.org/10.1016/j.biosystems.2016.09.001
3. Adélaïde, M., Sutre, G.: Parametric analysis and abstraction of genetic regulatory networks. In: Proc. 2nd Workshop on Concurrent Models in Molecular Biology (BioCONCUR'04), London, UK, Aug. 2004. Electronic Notes in Theor. Comp. Sci., Elsevier (2004), `http://www.labri.fr/~sutre/Publications/Documents/Adelaide:2004:BioCONCUR.ps.gz`
4. Bartocci, E., Lió, P.: Computational modeling, formal analysis, and tools for systems biology. PLOS Computational Biology **12**(1), 1–22 (01 2016). https://doi.org/10.1371/journal.pcbi.1004591
5. Blinov, M.L., Faeder, J.R., Goldstein, B., Hlavacek, W.S.: A network model of early events in epidermal growth factor receptor signaling that accounts for combinatorial complexity. Biosystems **83**(2), 136–151 (2006). https://doi.org/10.1016/j.biosystems.2005.06.014
6. Chatain, T., Haar, S., Paulevé, L.: Boolean networks: Beyond generalized asynchronicity. In: Baetens, J.M., Kutrib, M. (eds.) Cellular Automata and Discrete Complex Systems - 24th IFIP WG 1.5 International Workshop, AUTOMATA 2018, Ghent, Belgium, June 20-22, 2018, Proceedings. Lecture Notes in Computer Science, vol. 10875, pp. 29–42. Springer (2018). https://doi.org/10.1007/978-3-319-92675-9_3
7. Chifman, J., Arat, S., Deng, Z., Lemler, E., Pino, J.C., Harris, L.A., Kochen, M.A., Lopez, C.F., Akman, S.A., Torti, F.M., Torti, S.V., Laubenbacher, R.: Activated oncogenic pathway modifies iron network in breast epithelial cells: A dynamic modeling perspective. PLOS Computational Biology **13**(2) (2017). https://doi.org/10.1371/journal.pcbi.1005352
8. Cousot, P.: Constructive design of a hierarchy of semantics of a transition system by abstract interpretation. Theor. Comput. Sci. **277**(1-2), 47–103 (2002). https://doi.org/10.1016/S0304-3975(00)00313-3
9. de Jong, H., Gouzé, J.L., Hernandez, C., Page, M., Sari, T., Geiselmann, J.: Qualitative simulation of genetic regulatory networks using piecewise-linear models. Bulletin of Mathematical Biology **66**(2), 301–340 (2004). https://doi.org/10.1016/j.bulm.2003.08.010
10. Fages, F., Soliman, S.: Formal cell biology in biocham. In: Bernardo, M., Degano, P., Zavattaro, G. (eds.) Formal Methods for Computational Systems Biology, 8th International School on Formal Methods for the Design of Computer, Communication, and Software Systems, SFM 2008, Bertinoro, Italy, June 2-7, 2008, Advanced Lectures. Lecture Notes in Computer Science, vol. 5016, pp. 54–80. Springer (2008). https://doi.org/10.1007/978-3-540-68894-5_3
11. Faure, A., Naldi, A., Chaouiya, C., Thieffry, D.: Dynamical analysis of a generic boolean model for the control of the mammalian cell cycle. Bioinformatics **22**(14), e124–e131 (2006). https://doi.org/10.1093/bioinformatics/btl210

12. Feret, J., Danos, V., Krivine, J., Harmer, R., Fontana, W.: Internal coarse-graining of molecular systems. Proceedings of the National Academy of Sciences **106**(16), 6453–6458 (2009). https://doi.org/10.1073/pnas.0809908106
13. Feret, J., Koeppl, H., Petrov, T.: Stochastic fragments: A framework for the exact reduction of the stochastic semantics of rule-based models. Int. J. Softw. Informatics **7**(4), 527–604 (2013), `http://www.ijsi.org/ch/reader/view_abstract.aspx?file_no=i173`
14. Forejt, V., Kwiatkowska, M., Norman, G., Parker, D.: Automated verification techniques for probabilistic systems. In: Bernardo, M., Issarny, V. (eds.) Formal Methods for Eternal Networked Software Systems (SFM'11). LNCS, vol. 6659, pp. 53–113. Springer (2011). https://doi.org/0.1007/978-3-642-21455-4_3
15. Grosu, R., Batt, G., Fenton, F.H., Glimm, J., Guernic, C.L., Smolka, S.A., Bartocci, E.: From cardiac cells to genetic regulatory networks. In: Gopalakrishnan, G., Qadeer, S. (eds.) Computer Aided Verification - 23rd International Conference, CAV 2011, Snowbird, UT, USA, July 14-20, 2011. Proceedings. Lecture Notes in Computer Science, vol. 6806, pp. 396–411. Springer (2011). https://doi.org/10.1007/978-3-642-22110-1_31
16. Kauffman, S.: Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology **22**(3), 437–467 (1969). https://doi.org/10.1016/0022-5193(69)90015-0
17. Kwiatkowska, M., Norman, G., Parker, D.: PRISM 4.0: Verification of probabilistic real-time systems. In: Gopalakrishnan, G., Qadeer, S. (eds.) Proc. 23rd International Conference on Computer Aided Verification (CAV'11). LNCS, vol. 6806, pp. 585–591. Springer (2011). https://doi.org/10.1007/978-3-642-22110-1_47
18. Miné, A.: Symbolic methods to enhance the precision of numerical abstract domains. In: Emerson, E.A., Namjoshi, K.S. (eds.) Verification, Model Checking, and Abstract Interpretation, 7th International Conference, VMCAI 2006, Charleston, SC, USA, January 8-10, 2006, Proceedings. Lecture Notes in Computer Science, vol. 3855, pp. 348–363. Springer (2006). https://doi.org/10.1007/11609773_23
19. Thomas, R.: Boolean formalization of genetic control circuits. Journal of Theoretical Biology **42**(3), 563–585 (1973). https://doi.org/10.1016/0022-5193(73)90247-6