# Stochastic modelling of non-Markovian dynamics in biochemical reactions

Davide Chiarugi[4], Moreno Falaschi[1], Diana Hermith[1], Carlos Olarte[2], and Roberto Marangoni[3]

[1] Dip. Ingegneria dell'Informazione e Scienze Matematiche, Università degli Studi di Siena. Italy.
[2] Departamento de Electrónica y Ciencias de la Computación, Univ. Javeriana Cali. Colombia.
[3] Dip. di Informatica, Università degli Studi di Pisa. Italy.
[4] Istituto di Scienza e Tecnologie dell'Informazione, CNR, Pisa. Italy

**Abstract.** Biochemical reactions are often modelled as discrete-state continuous-time stochastic processes evolving as memoryless Markov processes. However, in some cases, biochemical systems exhibit non-Markovian dynamics. We propose a methodology for building stochastic simulation algorithms which model accurately non-Markovian processes in some specific situations. Our methodology is based on and implemented in Concurrent Constraint Programming (CCP). Our technique allows us to randomly sample waiting times from probability density functions (PDFs) not necessarily distributed according to a negative exponential function. In this context, we discuss an important case-study in which the PDF for waiting times is inferred from single-molecule experiments. We show that, by relying on our methodology, it is possible to obtain accurate models of enzymatic reactions that, in specific cases, fit experimental data better than the corresponding markovian models.

## 1 Introduction

Recently there has been a significant interest in discrete stochastic (DS) models of biological systems since experimental data have shown that stochasticity arising at the molecular level plays an important role in determining the overall behaviour of living organisms [1, 13]. Gillespie's Stochastic Simulation Algorithm (SSA) [6], based on [2] and [11], is the most widespread algorithm used for implementing DS simulations of biological systems. Gillespie's SSA requires that some hypotheses are satisfied, namely solutions are well stirred and in termal equilibrium and, more importantly, it holds only for elementary chemical reactions i.e. those reactions occurring in one reactive event. Even though it has ben shown that the SSA can work besides the prescribed scope of applicability as proved by the success of various stochastic models against experimental data (see e.g.[14, 7, 15]), it is difficult to describe biochemical systems in terms of elementary reactions: often there is an incomplete knowledge of the full set of elementary reactions and mesoscopic or macroscopic transformations are the only observable ones. Most commonly this problem is circumvented abstracting away the not observable elementary steps lumping them in a single reaction event, modelled as a single

"Markov jump" with the waiting time $\tau$ sampled from a negative exponential distribution depending on an overall rate constant. However abstractions usually introduce approximations in the behaviour of the models whose impact is not easy to evaluate or estimate, as noticed by Gillespie in [15] for enzymatically catalysed reactions. One crucial point in this abstraction approach concerns the modelling of the time needed for a reaction to occur. Indeed, even though the elementary reactions underlying a given biochemical process can be modelled as a DC Markov process (and, thus, with waiting times distributed according to a negative exponential PDF) on a mesoscopic or macroscopic scale the process may exhibit non-Markovian behaviours. This fact is highlighted also in [12] and [5] and shown by various experimental evidences, e.g. [18, 17]. These arguments suggest the need for modelling approaches embedding a more general notion of transition times, possibly allowing to manage non-memoryless (non-Markovian) system's evolution. In terms of waiting times, this corresponds to consider frameworks in which transitions times can be non necessarily exponential. Various approaches have been proposed for addressing this issue. BioPEPAd [5] allows to add deterministic delays to the duration of a reaction. The work in [9] proposes an extension for process calculi allowing to express activity durations through general probability distributions. A similar approach is proposed in [4] for extending Petri Nets. Furthermore, in [12] the authors improved the Beta Workbench (BWB) framework with the possibility of computing waiting times according to PDFs different from the negative exponential, such as the Erlang, or Hyperexponential trying to obtain better matches with observed non-markovian biological behaviours.

In this work we present a constraint programming (CP) approach to embed MC algorithms for discrete-state continuous-time stochastic simulation of biochemical reactions. In our approach, waiting times can be sampled according to PDFs that non necessarily follow a negative exponential PDF thus allowing to model non-markovian processes. More precisely, our approach allows us to consider (1) general continuous PDFs for waiting times such as Erlang, Hypoexponential or Hyperexponential as in [12]; and (2) case-specific PDFs such as those inferred from experiments. To highlight this last feature, we discuss a case study represented by the simulation of enzymatically catalysed reactions. We use our method for sampling waiting times from a PDF derived from *wet-lab* experiments and we provide a stochastic simulation method that results to be more accurate than the corresponding Markovian approach. The contribution of this paper is hence twofold: on the one hand we propose a general method for building discrete-state continuous time non-markovian simulations; on the other hand we present a case-specific application of our method discussing the differences with respect to the simulations obtained using Gillespie's SSA.

## 2 The CP Approach for simulating sampling waiting times

In this section we present our simulation method, based on a CP approach. We first review the computational framework, then we show how to model and simulate network of reactions. Finally, we discuss a relevant case study.

### 2.1 The computational framework

The crucial step of the Inverse Transform Sampling (ITS) technique is the solution of the equation $F(\tau) = r$, where $F(\tau)$ is the cumulative function calculated from the PDF $f(\tau)$ and $r$ is a random number generated by the computer. In order to compute a solution for the equation $F(\tau) = r$, we use the XRI library (`http://home.gna.org/xrilpoz/`), an implementation of a real interval constraint system. Constraint systems are at the heart of Constraint Logic Programming [16] where problems are solved declaratively: one states the problem and a Search Engine searches *automatically* for a solution. The XRI library implements efficient techniques such as Hull, Box and kB-Consistency ([3],[10]) for solving sets of numerical constraints. The advantage of this technique with respect to classical numerical methods is that: 1) no initial parameters for iteration are required; 2) interval arithmetic allows us to bound numeric errors since real numbers are represented by means of intervals instead of single floating-point numbers; and 3) constraints represent *relations* on the variables rather than *assignments* to values. Hence, by using the XRI on top of the Mozart system (`http://www.mozart-oz.org/`), we can impose the constraint $F(\tau) = r$ and we are able to both determining $r$ given a $\tau$ and computing $\tau$ given some random number $r$. This is a very general method that can be applied to all continuous PDFs. Indeed we can easily obtain correct random samples from general PDFs such as Erlang, Gamma and Hyperexponential distributions and we are able to correctly reproduce all the results presented in [12].

### 2.2 Simulating network of reactions: a toy example

Now we show how our method can be used to simulate networks of biochemical reactions. Noticeably our technique allows to build stochastic models of systems composed by set of reactions not necessarily characterised by a unique PDF for waiting times. This feature can be illustrated through a simple example where two different PDFs are used to sample the duration of the reactions. The procedure we follow is:

1. We sample a $\tau_i$ for each reaction.
2. We choose the smallest $\tau_i$ (i.e., the fastest reaction).
3. We change the concentration of the species according to the chosen reaction.

*Example 1 (A simple Network).* Consider the set of reactions:

$$(1)\ A \to B \qquad (2)\ A \to C$$

Reaction (1) follows the PDF in Equation 1 and $\tau$ in Reaction (2) is sampled from the exponential PDF:

$$F(\tau) = 10[S]e^{-[S]x}$$

The output of the simulation is depicted in Figure 1.

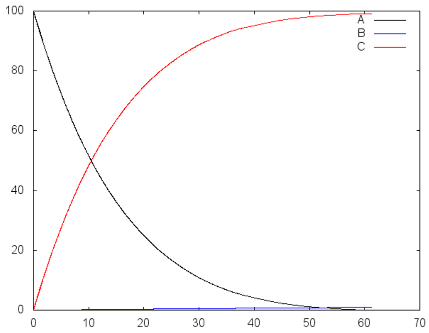


**Fig. 1.** A simple network of reactions: (1) $A \rightarrow B$, (2) $A \rightarrow C$. PDFs and parameters are described in Example 1.

### 2.3 Single molecules enzymatic reactions: a case study

Noticeably our technique works also for PDFs inferred from experiments. To show this feature we present a specific case-study regarding wet-lab experiments on single molecules of enzymes. In these experiments, it was measured the *waiting time* between a reaction and the following one [17]. In the studied cases, $\tau$ resulted to be distributed according to the following probability density function:

$$\begin{gathered} f(\tau) = \frac{k_1 k_2 [S]}{2A} \left[ e^{(A+B)\tau} - e^{(B-A)\tau} \right] \\ A = \sqrt{\frac{(k_1[S]+k_{-1}+k_2)^2}{4} - k_1 k_2 [S]} \; B = \frac{-(k_1[S]+k_{-1}+k_2)}{2} \end{gathered} \tag{1}$$

where $[S]$ is the substrate concentration and $k_1, k_{-1}, k_2$ are kinetic constants. It is worth noticing that this equation is very different from the negative exponential PDF for waiting times proposed by Gillespie. This suggests that a markovian approach may be not adequate to simulate the biochemical reactions catalysed by enzymes.

The solution of Equation (1), in the Constraint Programming approach, amounts to declare XRI variables for $r, \tau, A$ and $B$ and then, to impose the necessary constraints restricting the values the variables can take as depicted in the code of Figure 2. Hence the search engines looks for the values (intervals) for the variables that satisfy the above constraints (i.e., a solution for the set of equations). The results are depicted in Figure 3 and they show that the sampled data are distributed accordingly to the target PDF (Equation 1) as demonstrated also by a goodness of fit tests. The sampled waiting times can be used in the context of the algorithm presented in subsection 2.2 allowing accurate simulations of network of enzymatically catalysed reactions.

### 2.4 Comparing Markovian and non Markovian simulations

The problem of the approximation introduced in using Markovian models for describing non-Markovian phenomena is well known in physics as discussed, e.g., in [8]. In this

```
{XRI.hc4 eq(A sqrt(sub(divide(
          pow(plus(times(K1 S) K2 K3) 2.0) 4.0)
          times(K1 K2 S) )) )}

{XRI.hc4 eq(B divide(
              times(-1.0 plus(times(K1 S) K3 K2)) 2.0) )}

{XRI.hc4 eq(ALPHA divide(times(K1 K2 S) times(2.0 A)))}

{XRI.hc4 eq(BETA plus(A B))}

{XRI.hc4 eq(GAMMA sub(B A))}

{XRI.hc4
 eq(R
    sub(
       times(ALPHA sub(
           divide(pow(E times(BETA TAU)) BETA)
           divide(pow(E times(GAMMA TAU)) GAMMA) ))
       times(ALPHA sub(
           divide(1.0 BETA)
           divide(1.0 GAMMA) )) ) )}
```

**Fig. 2.** XRI Constraints to solve $r = F(\tau)$ in the Equation (1). We use the subexpressions `ALPHA`,`BETA` and `GAMMA` to ease readability in the expression `R` $=$ `F(TAU)`

subsection we will address this issue for a biological scenario. To this aim we consider the well known Michaelis-Menten reaction scheme for enzymatic reactions:

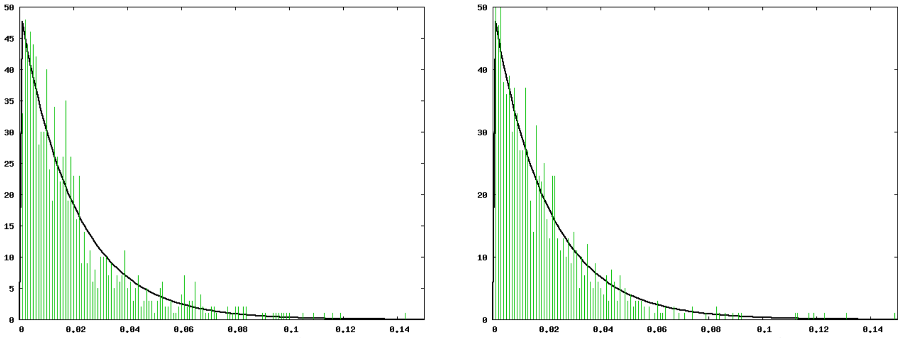
$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \tag{2}$$

where $S$ and $P$ indicates the substrate and product molecules respectively, while $E$ represent the enzyme molecule and $ES$ the enzyme-substrate complex.

In particular we analyse the differences emerging in describing reaction set (2) through two different discrete stochastic models. In one case, we use an approach *à la* Gillespie, lumping the reaction scheme (2) into one $S \to P$ reaction. This reaction is seen as a single Markov jump having the waiting time distributed accordingly to a negative exponential PDF with propensity $a = v$ calculated through Equation (3). Here $v$ is the rate of product formation, $[S]$ the substrate concentration and $K_M$ the Michaelis-Menten constant (see e.g., [15]).

$$v = \frac{k_2[S]}{[S] + K_M} \tag{3}$$

In the other case, we reduce reaction scheme (2) to a single $S \to P$ reaction but the waiting time of the process is calculated (using our CCP based method) according

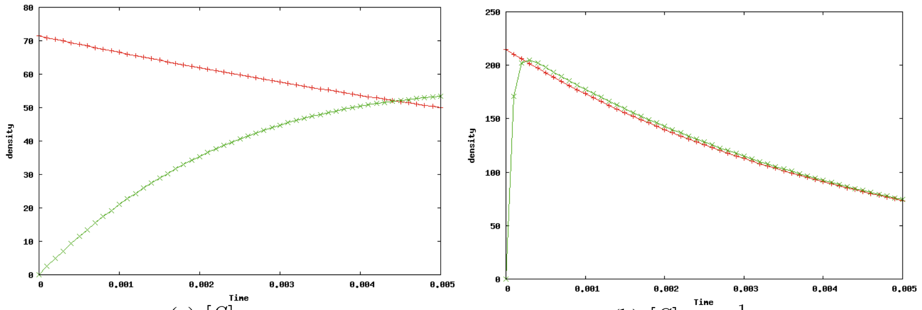


(a) A1 ($k_{-1} = 0s^{-1}$)

(b) A2 ($k_{-1} = 50s^{-1}$)

**Fig. 3.** Histograms represent the occurrences of $\tau$ obtained using our method. The continuous line represents the plot of (1) calculated using a specific set of parameters.

to Equation (1) , thus obtaining a non-markovian model. As previously noticed, experimental evidences provided by studies on single molecules suggest that the occurrence of a reaction catalysed by enzymes when observed at the mesoscopic level (i.e as a single $S \to P$ reaction) exhibit non-markovian dynamics. Our aim here is to evaluate the impact of the approximation introduced by simulating single-molecule enzymatic reactions as a Markov process through the Gillespie's SSA. To do this, we first compare the trend of the two PDFs for waiting times (the negative exponential with propensity $a$ and Equation (1)) when the initial amount of $S$ ($S_0$), $k_1$, $k_{-1}$ and $k_2$ are varied. As highlighted by figures 4(a) and 4(b) it can be noticed that the differences between the two PDF regard mainly the low values of waiting times: accordingly to the negative exponential PDF waiting times close to zero have higher values of $P(\tau)$, while accordingly to the PDF in equation (1) the higher values for $P(\tau)$ are not close to zero. Noticeably these differences are greater when $S_0$ is low and when $k_{-1}$ is significantly minor than $k_2$. This implies that, with respect to the experimental findings on single molecule enzymatic reactions, the Markovian approach introduces more approximation when the modelled system is composed by a low number of reactants and when the kinetic characteristics of the enzymatic reaction (i.e. various combinations of the values of $k_1$, $k_{-1}$ and $k_2$) prevent the system to reach quickly the steady state. To study the impact of the differences amongst PDFs on the dynamics of the system, we simulated the occurrence of single-enzyme catalysed reactions as single $S \to P$ Markov or non-Markov jumps. We compared the obtained results with the outputs of simulations performed describing reaction set (2) through the Gillespie's SSA specifying each single step (call it *full model*). As proposed in [15], given the correctness of the Gillespie's SSA, the full model can be used as a convenient benchmark for evaluating the precision of simulations against the Michaelis-Menten model. As it can be noticed observing Figure 5 (and confirmed by Pearson's test) the results of our simulations fit the full model significantly better than the corresponding markovian approach. Thus, our technique allows to describe the considered enzymatic reaction as a single $S \to P$ step with an accuracy comparable to that of the full model providing a safe technique for model reduction.



(a) $[S]$ vs. $v$

(b) $[S]$ vs. $\frac{1}{<\tau>}$

**Fig. 4.** Plots of the negative-exponential PDF (red) and the PDF according to equation (1) (green). The plots on the left is obtained with low $S_0$ and with $k_{-1} << k_2$ while the plot on the right with higher $S_0$ and $k_2 > k_{-1}$.

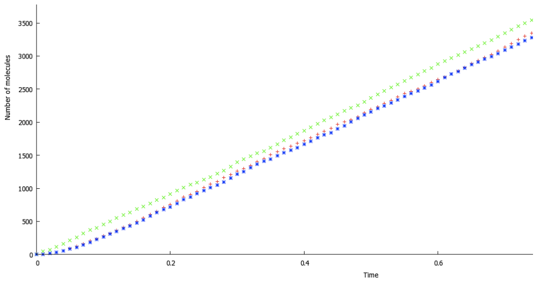


**Fig. 5.** Plots of number of product molecules vs. time obtained simulating a single molecule reaction through a Markovian (green) or non-Markovian (red) single jump. The blue plot is obtained decribing the corresponding full model.

## 3 Concluding remarks

In this paper we presented a novel technique for the stochastic simulation of biochemical reactions. The core of our proposal is a CCP based algorithm that allow to rely on a Monte Carlo method for sampling waiting times from continuous PDF in general. In this way it is possible to simulate also stochastic processes with memory, thus enabling accurate descriptions of non-Markovian phenomena that can be observed in nature. Noticeably the possibility of dealing with all continuous PDF, allow to build models using straightforwardly experimentally measured waiting times with no need of reproducing observed dynamics by tuning model parameters.

We applied our method to a case-study regarding reactions catalysed by single enzyme molecules. We showed that this method takes some advantages on the corresponding markovian approach. In particular it turned to be more accurate expecially in those cases characterised by a low number of molecules and by dynamics that makes the steady-state not quickly reachable.

Finally it is important to point out that at least for the cases that we tested, the time requested for computing simulations is similar to that needed by Gillespie's SSA.

## Acknowledgements

This work is partially supported by the FLAGSHIP ”InterOmics” project (PB.P05) which is funded and supported by the italian MIUR and CNR organizations

## References

1. G. Balzsi, A. van Oudenaarden, and J.J Collins. Cellular decision making and biological noise: from microbes to mammals. *Cell*, (6):910–925, 2011.
2. A. F. Bartholomay. Molecular set theory: A mathematical representation for chemical reaction mechanisms. *Bulletin of Mathematical Biology*, (3):285–307, 1960.
3. F. Benhamou, Fréderic Goualard, and Laurent Granvilliers. Revising hull and box consistency. In *Proceedings of ICLP'99 - MIT Press*, pages 230–244, 1999.
4. A. Bobbio and M. Telek. Computational restrictions for spn with generally distributed transition times. In D. Hammer K. Echtle and D. Powell (Eds.), editors, *First European Dependable Computing Conference (EDCC-1)*, volume 852 of *LNCS*, pages 131–148. Springer Berlin Heidelberg, 1994.
5. G. Caravagna and J. Hillston. Bio-pepad: A non-markovian extension of bio-pepa. *Theoret. Comp. Sci.*, (419):26–49, 2012.
6. DT. Gillespie. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem.*, 58:35–55, 2007.
7. D. Gonze, J. Halloy, and J. Goldbetter. A deterministic versus stochastic model for circadian rhythms. *J. Biol. Phys.*, (28):637–653, 2002.
8. P. Hanngi and P. Talkner. Memory index of first-passage time: a simple measure of non-markovian character. *Phys. Rev. Letters*, pages 2242–2248, 1983.
9. M. R. Lakin, L. Paulev, and A. Phillips. Stochastic simulation of multiple process calculi for biology. *Theoretical Computer Science*, page in press, 2012.
10. O. Lhomme. Consistency techniques for numeric csps. In *Proceedings of the 13th IJCAI, IEEE Computer Society Press*, pages 232–238, 1993.
11. D. A. McQuarrie. Stochastic approach to chemical kinetics. *Journal of Applied Probability*, (3):413–478, 1967.
12. I. Mura, D. Prandi, C. Priami, and A. Romanel. Exploiting non-markovian bio-processes. *IET Systems Biology.*, (253):83–98, 2009.
13. A. Raj, S. A. Rifkin, E. Andersen, and A. van Oudenaarden. Variability in gene expression underlies incomplete penetrance. *Nature*, (463):913, 2010.
14. C.V. Rao and A. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: application to the gillespie algorithm. *J. Chem. Phys.*, (118):4999–5010, 2003.
15. K.R. Sanft, D.T. Gillespie, and L.R. Petzold. Legitimacy of the stochastic michaelis-menten approximation. *IET Systems Biology.*, (5):58–69, 2011.
16. Vijay A. Saraswat. *Concurrent Constraint Programming*. Logic Programming. MIT Press, 1993.
17. S.C. Kou *et al.* Single-molecule michaelis-menten equations. *Journal of Physical Chemistry*, (109):19068–19081, 2005.
18. T. Franosch *et al.* Resonance arising from hydrodynamic memory in brownian motion. *Nature*, (478):85–88, 2011.