ISTITUTO DI ANALISI DEI SISTEMI ED INFORMATICA
"Antonio Ruberti"
CONSIGLIO NAZIONALE DELLE RICERCHE

A. Borri, F. Carravetta, G. Mavelli, P. Palumbo

A STUDY ON THE STRUCTURAL PROPERTIES AND THE SOLUTION OF THE CHEMICAL MASTER EQUATION

R. 10, Ottobre 2012

Alessandro Borri – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: alessandro.borri@iasi.cnr.it.

Francesco Carravetta – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: francesco.carravetta@iasi.cnr.it.

Gabriella Mavelli – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: gabriella.mavelli@iasi.cnr.it.

Pasquale Palumbo – Istituto di Analisi dei Sistemi ed Informatica del CNR, viale Manzoni 30 - 00185 Roma, Italy. Email: pasquale.palumbo@iasi.cnr.it.

ISSN: 1128–3378

Collana dei Rapporti dell'Istituto di Analisi dei Sistemi ed Informatica "Antonio Ruberti", CNR

viale Manzoni 30, 00185 ROMA, Italy

tel. ++39-06-77161
fax ++39-06-7716461
email: iasi@iasi.cnr.it
URL: http://www.iasi.cnr.it

Abstract

The Chemical Master Equation (CME) is a well-known tool for studying (bio)chemical processes involving few copies of the species involved, because it is a framework able to capture random behaviors that are neglected by deterministic approaches based on the concentration dynamics. In this work, we investigate some structural properties of CMEs and their solutions, with a particular focus on the efficient computation of the stationary distribution. We introduce a generalized notion of one-step process, which results in a sparse dynamical matrix describing the collection of the scalar CMEs, also showing a recursive block-tridiagonal structure.

1. Introduction

An important research topic in the field of Systems Biology is the detection of efficient methods for modeling complex cellular mechanisms. It has been recently pointed out the importance of the noise role in the dynamics of biological processes [18]. Random fluctuations, provided by a wide set of concurring factors including, for instance, thermal noise or asynchronous occurrence of synthesis and degradation events, need to be considered when modeling most of the molecular processes involved in cellular regulation (such as phosphorylation/dephosphorylation processes or other enzymatic reactions), as well as in gene expression, see e.g. [1, 4]. These noisy behaviors are much more evident in cases when only few copies of the reacting species (DNA, RNA or regulating proteins) are involved, and standard Ordinary Differential Equation models, essentially based on the concentration dynamics, fail to capture the inherent randomness of the phenomena. Indeed, in these cases the Chemical Master Equation (CME) approach reveals to be more suitable, describing the biological process under investigation in terms of the dynamics of the probability distribution of the chemical population of the system under study [22, 13, 14].

Such an attractive, stochastic approach, which allows to simulate and keep track of the reactions occurring in a single fixed volume, has recently become more and more appealing because of the biotechnology devices available nowadays, which are able to provide single-cell experimental data: see, e.g. [9], where CME-based stochastic simulations have been used to validate a model of the Ras/cAMP/PKA signaling pathway, or the recent [17, 23, 16] where stochastic simulations have been used, in general, with the goal of *reverse engineering* from real data.

Except very basic cases, to find the exact solution of a CME is a *hard nut to crack*, even if we are interested only in the steady-state solution. Indeed, the aggregation of CMEs results to be a linear system, whose state vector at a given time t collects the probabilities for all the possible copies of all the involved species. It readily appears that, even in the case of a closed system of reacting species (i.e. the CME is finite-dimensional), the computation of the matrix exponential or of the null space (in the stationary case) would require non-trivial numerical algorithms to be implemented. Moreover, different ways of aggregation of the probabilities to define the state vector may produce different properties of the linear system to be solved. For these reasons, most efforts have been concentrated so far to implement Monte Carlo methods (such as the Stochastic Simulation Algorithm (SSA), [13, 12], or the τ-leaping algorithms, [15, 6]) with the goal of approximating the exact solutions. As a matter of fact, the performance of such algorithms is a tradeoff between the high number of Monte Carlo simulations required to approximate the exact solution and the time spent for running a single long-term simulation. It has to be stressed that, in some crucial cases, such a tradeoff could reveal to be not satisfying. This is the case when some biological traits happen rarely, thus requiring a very high number of Monte Carlo simulations in order to get a sufficiently precise statistics. Examples are usually taken from biological toggle switches, such as for instance, the ones related to the pyelonephritis-associated pili (Pap) epigenetic switch in E. coli [3] or the genetic toggle switch model of Gardner [10]. Therefore, a need exists to overcome purely numerical Monte Carlo methods and look for the solution of the original CME. Recent results on this field have been published in [19, 20], where the CME is investigated in some details and a numerical algorithm is proposed to provide the approximate solution of the CME. However, this work lacks a characterization of the dynamical properties of the CME in the exact case, which are useful to search for the solution in the presence of a large number of states.

The present work aims to further investigate the properties of the CME and its solution. Indeed, the exact stationary solution of a CME is available in closed form just for simple *one-*

dimensional cases, meaning that the system under investigation has to be 'simple enough' to be described by the number of copies of just one species. This is the case for instance of a single phosphorylation reaction. In the case of more complex reactions or even sets of chemical reactions, to find the exact equilibrium solution is not a trivial issue and, as mentioned before, it is usually computed by means of stochastic Monte Carlo simulation [13]. Here we propose a mathematical setting that can be always adopted in spite of the number or kind of reactions/reagents. The method to build up the CMEs uses an interesting recursive approach to aggregate the vector of probabilities and the dynamical matrix. The analysis of the dynamical properties is the novelty of the paper, as well as the characterization of the chemical networks of reactions in a *graph-theoretical* fashion; our results are suitably exploited to find efficiently the exact solution at the equilibrium, in a much faster way with respect to usual Monte Carlo SSA techniques; the method is validated in a well-established biochemical example [11].

The paper is organized as follows: in Section 2 we recall some preliminary definitions and introduce the general setting, supported by some examples. In Section 3 we introduce the CME in matrix form and show the recursive block-tridiagonal structure in the (generalized) one-step setting. Section 4 addresses the structural properties and the solution of the CMEs. In Section 5, we adapt the Gillespie Stochastic Simulation Algorithm (SSA) [13] to our setting for Monte Carlo simulations, and present some simulation results in a biochemical application. Section 6 offers concluding remarks.

2. Non-redundant state representation of Chemical Networks

2.1. Preliminaries

The symbols $\mathbb{N}$, $\mathbb{N}_0$, $\mathbb{R}$ and $\mathbb{R}^+$ denote the set of natural, nonnegative integer, real and positive real numbers, respectively. The cardinality of a set V is denoted by $|V|$. The transpose of a matrix A is written as A^T. The time-derivative $\frac{dp}{dt}$ is denoted by $\dot{p}$.

A *reaction* is a process leading to the transformation of a set of substances to another. Hereinafter we call *species* the kinds of players involved in a given chemical/biochemical reaction, and *elements* (or *elementary species*) the kinds of basic constituents of all the involved species, which means: any species is the aggregate of some elements, whilst no element can be expressed as the aggregate of other elements. For instance, in standard chemical reactions, the species are the kinds of molecules involved in the reactions, and the elements are simply the chemical elements. On the other hand, in the case of biochemical reactions (like, e.g., phosphorylation or protein aggregation) involving macromolecules (e.g. proteins, RNA, etc.) it will be useful to consider *elements* the basic molecules composing larger molecular complexes.

In the following, we consider a system (or network) of q (bio)chemical reactions in the general form:

$$\begin{cases} a_{11}^l Y_1 + \cdots + a_{1N}^l Y_N \longrightarrow a_{11}^r Y_1 + \cdots + a_{1N}^r Y_N \\ \vdots \\ a_{q1}^l Y_1 + \cdots + a_{qN}^l Y_N \longrightarrow a_{q1}^r Y_1 + \cdots + a_{qN}^r Y_N \end{cases} \tag{1}$$

where $Y_1, \ldots, Y_N$ are the species involved. The number $\beta_{ij} = a_{ij}^r - a_{ij}^l$ is called *stoichiometric coefficient* of species j in reaction i, for all $j = 1, \ldots, N$ and for all $i = 1, \ldots, q$. In case the right-hand-side and the left-hand-side of a given reaction are equal, respectively, to the left-hand-side and the right-hand-side of a different reaction, we will simply denote the two reactions by means of a unique formal *reversible* reaction with the symbol $\leftrightharpoons$.

Let us define $n(t)$ as the state of the system at time t, with i-th component $n_i(t) \in \mathbb{N}_0$ denoting the number of copies of the i-th species at time t. The state function $n : [0, +\infty) \to \mathbb{N}_0^N$ is a realization of a discrete-valued continuous-time stochastic Markov process with initial conditions $n_i(0) = \bar{n}_i$, $i = 1, \ldots, N$. We refer to a system (or to the underlying process) as *closed* if $n_i(t) \in \{0, 1, \ldots, N_i\}$ for some fixed $N_i \in \mathbb{N}$, for all i. A system is *open* if it is not closed, namely if the number of copies of some species is not uniformly bounded.

2.2. Detection of independent and orthogonal species

We denote by N_E the number of distinct elements that form the N species. In closed processes, the total number of copies of each element is conserved, no matter what species it may be a part of. This imposes N_E mass-balance constraints on the system, which can be written in matrix form as:

$$F \cdot n(t) = b, \tag{2}$$

where $b \in \mathbb{N}^{N_E}$ is the vector collecting the total mass, expressed in numbers of copies of each elementary species, and $F \in \mathbb{N}_0^{N_E \times N}$ is a mass-balance matrix whose (i, j)-th entry $F_{i,j}$ is the number of copies of element i present in one copy of species j. Note that Eq. (2) is independent of the particular set of reactions, instead it depends on the species involved in the reactions. Because of the N_E mass-balance equations in (2), the state vector $n(t)$ is redundant, because only $M = N - N_E$ state variables are independent, assuming the non-trivial case $N > N_E$ and $rank(F) = N_E$. This leads to the concept of independent species. In the remainder of the paper we will always assume $N > N_E$ and $rank(F) = N_E$, if not explicitly stated otherwise.

Definition 2.1 (Independent species) *Consider a system of q chemical reactions (1) involving N_E elements and $N > N_E$ species, and let $F = [F_1 \cdots F_N]$ be the matrix of mass-balance constraints in (2) such that $rank(F) = N_E$. Let $\{Y_{r_1}, \ldots, Y_{r_M}\}$, with $\{r_1, \ldots, r_M\} \subseteq \{1, \ldots, N\}$, be a subset of the species reacting in (1), with $M = N - N_E$. $\{Y_{r_1}, \ldots, Y_{r_M}\}$ is a set of* independent species *for (1) if the matrix $\bar{F}$ obtained by erasing columns $r_1, \ldots, r_M$ from F is invertible.*

Definition 1 is properly exploited in the following Proposition.

Proposition 2.2. *Consider a system of q chemical reactions involving N_E elements and $N > N_E$ species, with the mass-balance constraints given by (2) and $rank(F) = N_E$. Let $\{Y_{r_1}, \ldots, Y_{r_M}\}$, with $\{r_1, \ldots, r_M\} \subseteq \{1, \ldots, N\}$, be a set of* independent species *for the q chemical reactions. Then the reduced state vector $x(t) = [n_{r_1}(t) \cdots n_{r_M}(t)]^T$ is a non-redundant representation of vector $n(t)$ at any time t.*

Proof. The proof is a consequence of the Rouché-Capelli theorem. Denote by $\tilde{F}$ the matrix composed by the columns $r_1, \ldots, r_M$ of F, and denote by $\bar{F}$ the matrix obtained by erasing columns $r_1, \ldots, r_M$ from F, according to Definition 2.1. Let $\bar{x}(t)$ the vector obtained by erasing components $r_1, \ldots, r_M$ from the state vector $n(t)$. The proof consists in showing that it is possible to compute the state subvector $\bar{x}(t)$ as a result of the knowledge of the reduced state $x(t)$ (collecting just the components referred to the independent species) and the mass-balance constraints in (2), which can be rewritten as

$$\tilde{F} \cdot x(t) + \bar{F}\bar{x}(t) = b.$$

The definition of independent species guarantees the existence of $\bar{F}^{-1}$ such that (at any time t)

$$\bar{x}(t) = \bar{F}^{-1}(b - \tilde{F} \cdot x(t)),$$

concluding the proof. ∎

Remark 2.3. *Definition 2.1 provides a simple method to build a non-redundant state representation of a network of chemical reactions and the proof of Proposition 2.2 shows a constructive technique to infer the redundant state components from the independent ones. The procedure above can be readily generalized to the case of systems with a number of constraints N_C smaller than the number of element N_E. This is the case, for instance, when a chemical is present in the reactions in different forms (e.g. phosphorylated/non-phosphorylated, or methylated/non-methylated): it is a typical occurrence in biochemical reactions, when the chemicals involved are proteins (e.g. enzymes) that play their role depending on whether they are or not in their active state. Therefore, we may figure out a case where the same chemical is involved in a set of reactions according to two different elementary species, but the mass-balance constraint is only one. In this case, the number M of independent species is equal to $N - N_C$ and the procedure illustrated above is still valid, with the necessary modifications. We omit further details in order to keep notation simple, but we discuss an example of such a system in Example 4.*

In open systems (without mass-balance constraints), the species are always independent. A further property possibly enjoyed by independent species is given by the following Definition.

Definition 2.4 (Orthogonal species) *Consider a system of q chemical reactions as in (1). A set $\{Y_{r_1}, \ldots, Y_{r_M}\}$ of independent species is a set of* orthogonal species *for the given system of reactions if for any reaction $i \in \{1, \ldots, q\}$ there exists a unique index $j \in \{r_1, \ldots, r_M\}$ such that $\beta_{ij} \neq 0$.*

In practice, given M independent species on a given set of reactions, we say that these species are orthogonal if every chemical reaction changes just one component of the reduced state $x(t)$. We conclude this section with some examples of (bio)chemical reactions illustrating the concepts defined above. Notice that the first trivial cases have a purely didactic purpose.

Example 1. Consider the following closed system of $q = 4$ reactions:

$$\begin{cases} 2H_2 + O_2 & \rightleftharpoons & 2H_2O \\ CaCO_3 & \rightleftharpoons & CaO + CO_2, \end{cases}$$

in which we have $N = 6$ species, $Y_1 = H_2$, $Y_2 = O_2$, $Y_3 = H_2O$, $Y_4 = CaCO_3$, $Y_5 = CaO$, $Y_6 = CO_2$) and $N_E = 4$ elements (H, O, Ca, C). The matrix F is

$$F = \begin{bmatrix} 2 & 0 & 2 & 0 & 0 & 0 \\ 0 & 2 & 1 & 3 & 1 & 2 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}.$$

The number of independent species is $M = N - N_E = 2$. One possible choice is $\{H_2O, CaO\}$ ($r_1 = 3$, $r_2 = 5$), because the matrix

$$\bar{F} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 2 & 3 & 2 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix},$$

obtained from F by erasing columns 3 and 5 has a determinant $\neq 0$. Definition 2.4 is also verified, namely the chosen species are also orthogonal, because the four reactions do not change the number of H_2O and CaO simultaneously.

Example 2. Consider a closed system in which $N = 5$ species react according to:

$$\begin{cases} 2H_2 + O_2 & \rightleftharpoons & 2H_2O \\ H_2O + CO_2 & \rightleftharpoons & H_2CO_3. \end{cases}$$

One can write $N_E = 3$ constraints by the mass-balance of the elements H, O, C present in the above $q = 4$ reactions. Since $M = N - N_E = 2$, the system has a 2-dimensional reduced state vector. One possible choice of independent species is $\{H_2, H_2O\}$. These species are not orthogonal because the first two reactions change both the components of the reduced state. On the contrary, it can be verified that an orthogonal pair of independent species is $\{H_2, CO_2\}$.

Example 3 (MicroRNA Toggle Switch). CMEs can be usefully exploited also to represent (and simulate, as usual) a biological framework, already modeled by means of ordinary differential equations. In [11], for instance, the following microRNA-protein toggle switch is considered:

$$\begin{cases} \dot{p}(t) & = & \bar{\alpha} + \frac{k_1 p(t)^2}{\Gamma_1 + p(t)^2 + \Gamma_2 m(t)} - \delta p(t) \\ \dot{m}(t) & = & \beta + k_2 p(t) - \gamma m(t) \end{cases} \tag{3}$$

where $p(t)$ and $m(t)$ represent the E2F-Myc complex and the miRNA cluster concentrations, respectively. A CME approach substantially identifies the state of the system by means of the number of copies of each species under investigation, allowing them to increase or decrease of a fixed number of copies. For instance, in the case of a *one-step* orthogonal choice (i.e. each species can increase/decrease of one unit per time), we can formalize the following set of 4 reactions:

$$\begin{cases} p & \longrightarrow & p+1 \\ p & \longrightarrow & p-1 \\ m & \longrightarrow & m+1 \\ m & \longrightarrow & m-1 \end{cases}$$

which is a 2-dimensional open system ($N = 2$). The coefficients of the ODE model will play a role to set the *propensities* of the CME, as it will be clearer in the Simulation section, where such an example will be investigated in details.

Example 4 (Cyclin-dependent kinase). Let us consider the following $q = 7$ biochemical reactions:

$$\begin{cases} \mathrm{Cdk} + \mathrm{Cln} & \rightleftharpoons & \mathrm{CdkCln} & (4) \\ \mathrm{CdkCln} + \mathrm{I} & \rightleftharpoons & \mathrm{CdkClnI} & (5) \\ \mathrm{CdkClnI} & \longrightarrow & \mathrm{CdkClnI}_p & (6) \\ \mathrm{CdkClnI}_p & \rightleftharpoons & \mathrm{CdkCln} + \mathrm{I}_p & (7) \end{cases}$$

that represent the first step of the biochemical machinery leading to the G_1/S transition in yeast cell cycle. In this framework, Cdk is a *cyclin-dependent kinase*, which requires the binding of the proper cyclin Cln to be activated; at the same time a different molecule, I, plays the opposite role of inhibitor when binding to the complex $CdkCln$. The inhibited complex $CdkClnI$ is set free of the inhibitor by means of a first step of phosphorylation $CdkClnI_p$ (here considered as

a not reversible reaction, by neglecting the actions of phosphatases) and then degraded into $CdkCln$ and the phosphorylated inhibitor I_p. The chemical players previously mentioned are found in literature as $Cdk1$, $Cln3$ and $Far1$ (this last being the inhibitor), see e.g. [21] and references therein. According to our Definitions, we have $N = 7$ chemical species, $N_E = 4$ elements, but only $N_C = 3$ mass-balance constraints for the three elementary species Cdk, Cln and I, being I and I_p two different forms of the same chemical player (see Remark 2.3). The reduced state is 4-dimensional and a choice of independent species is, for example: Cdk, I, $CdkClnI$ and I_p. Notice that it is impossible to draw a subset of 4 independent orthogonal species from the set of 7 species, because one of the two species involved in reaction (6) needs to belong to the set of independent species (otherwise such reaction would not affect the reduced state). However, since they also appear in other reactions (5) and (7), an occurrence of such reactions would produce a simultaneous change in some other states.

Notice that in the given examples, although it is always possible to define a set of independent species out of a system of q reactions, it is not always possible to choose M independent and orthogonal species (see Example 4). Nevertheless, the orthogonality property can be recovered by defining state variables as appropriate linear combinations of the original species, but this is not object of the present work. A stochastic Markov process describing a network of reactions that is described by means of M independent orthogonal species is called *orthogonal process*. An important class of orthogonal processes is the class of *one-step* processes, illustrated in the next section.

3. The multi-dimensional Chemical Master Equation (CME) for generalized one-step processes

In the remainder of the paper, we will refer to systems of q chemical reactions that are already in reduced (non-redundant) form, with state $x(t) \in \mathbb{N}_0^M$, whose components are related to a choice of M independent species in the q reactions. Define $p_{n_1,n_2,\ldots,n_M}(t)$ as the joint probability of having n_i copies of the i-th species (for $i = 1, \ldots, M$) at time t:

$$p_{n_1,\ldots,n_M}(t) \doteq P\Big(x_1(t) = n_1, \ldots, x_M(t) = n_M\Big).$$

Moreover, let us define the *propensities* or transition probabilities per unit of time as:

$$g^{\alpha_1,\ldots,\alpha_M}_{n_1,\ldots,n_M} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P\Big(x_i(t+\Delta t) = n_i + \alpha_i, i = 1,\ldots,M \Big/ x_i(t) = n_i, i = 1,\ldots,M\Big),$$

in which P is the conditional probability for a step transition of the discrete amount α_i in the state-variable $x_i(t)$, $i = 1, \ldots, M$. We assume such a probability does not depend explicitly on the time t.

Consider the aggregate vectors of stoichiometric coefficients $\beta_i \doteq (\beta_{i1}, \ldots, \beta_{iM})$, for $i = 1, \ldots, q$. We will assume, as usual [22], that only one reaction per time can occur. Then, the conditional probabilities are constrained to the ones matching the variations in the number of copies of all species in some reactions, namely:

$$g^{\alpha_1,\ldots,\alpha_M}_{n_1,\ldots,n_M} = 0 \quad \text{if } (\alpha_1, \ldots, \alpha_M) \notin \{\beta_1, \ldots, \beta_q\}.$$

According to the previous definitions, one-step processes [22] have the following features:

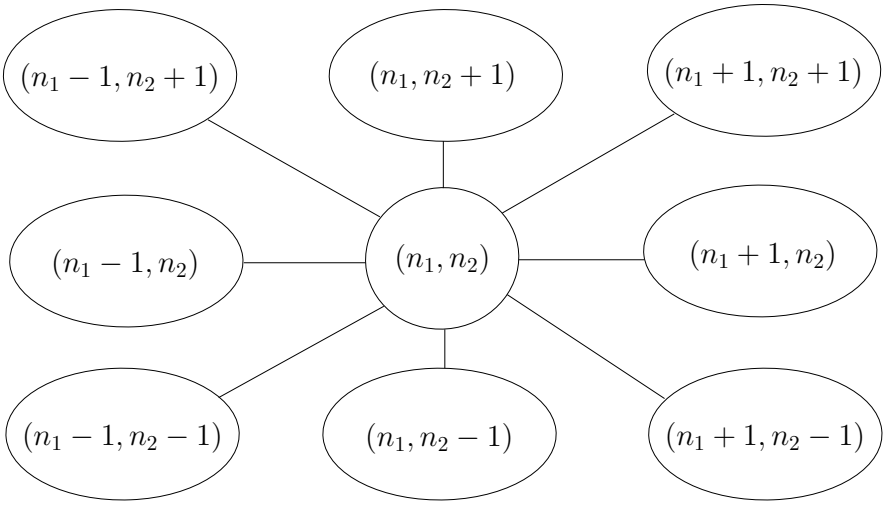

Figure 1: Example of bivariate generalized (non-orthogonal) one-step process. The links indicate non-zero transition probabilities per unit of time. Notice that the diagonal links vanish in the classical one-step case.

(a) $\beta_{ij} \in \{-1, 0, 1\}$ for all i, j (unitary steps);

(b) the M species are orthogonal.

In the following, we will generalize the one-step setting by removing the orthogonality hypothesis (b). Such a general case is referred later as *generalized* (non-orthogonal) one-step process, or *unitary* process, because reactions are allowed to cause simultaneous changes of unitary amount in the state variables. A graphical example is given in Figure 1.

For the general state $(n_1, \ldots, n_M)$ of the Markov process, one can derive from [22] the expression of the Chemical Master Equation (CME) as:

$$\begin{aligned} \dot{p}_{n_1,\ldots,n_M}(t) = & \sum_{(-\alpha_1,\ldots,-\alpha_M)\in\{\beta_1,\ldots,\beta_q\}} g^{-\alpha_1,\ldots,-\alpha_M}_{n_1+\alpha_1,\ldots,n_M+\alpha_M} p_{n_1+\alpha_1,\ldots,n_M+\alpha_M}(t) \\ & - \sum_{(\alpha_1,\ldots,\alpha_M)\in\{\beta_1,\ldots,\beta_q\}} g^{\alpha_1,\ldots,\alpha_M}_{n_1,\ldots,n_M} \cdot p_{n_1,\ldots,n_M}(t). \end{aligned} \tag{8}$$

In the case of closed systems, the equations in (8) consist of a set of $N_1 \times \cdots \times N_M$ equations providing the dynamics of the joint M-dimensional probability distribution. In the following, we propose a way to collect these equations in a compact form that ensures some interesting properties. Notice that the hypothesis of closed system is usually reasonable since, even in the case of open systems, one can truncate the system by limiting the population of the *unbounded species* at reasonable levels, in order to constrain the total number of states yet guaranteeing accurate results. This approximation is justified because the CME approach is usually used for reactions involving *few molecules* (see [22] and Section 5).

For any choice of the $N_1 \times \cdots \times N_{M-1}$ possible settings of the copies of the first $M-1$

independent species, define the N_M-dimensional vector of probabilities:

$$\mathcal{P}_{n_1,\ldots,n_{M-1}} \doteq \begin{bmatrix} p_{n_1,\ldots,n_{M-1},1} \\ p_{n_1,\ldots,n_{M-1},2} \\ \vdots \\ p_{n_1,\ldots,n_{M-1},N_M} \end{bmatrix}. \tag{9}$$

Then, the following vectors of probabilities can be recursively defined

$$\mathcal{P}_{n_1,\ldots,n_i} \doteq \begin{bmatrix} \mathcal{P}_{n_1,\ldots,n_i,1} \\ \mathcal{P}_{n_1,\ldots,n_i,2} \\ \vdots \\ \mathcal{P}_{n_1,\ldots,n_i,N_{i+1}} \end{bmatrix}, \qquad 1 \leq i \leq M-2 \tag{10}$$

up to the definition of vector $\mathcal{P}$, entailing all the probabilities involved by the CME:

$$\mathcal{P} \doteq \begin{bmatrix} \mathcal{P}_1 \\ \mathcal{P}_2 \\ \vdots \\ \mathcal{P}_{N_1} \end{bmatrix}. \tag{11}$$

Since the right-hand side of Eq. (8) is a linear combination of the joint probabilities of states of the Markov process, the equations for the joint probabilities of all states can be collected in the form of an autonomous linear system:

$$\dot{\mathcal{P}} = G\mathcal{P}, \tag{12}$$

where the dimension of G is $N_1 \times \cdots \times N_M$. In the following, we present the main result of this section showing that the assumption of (generalized) one-step process, previously described, determines a particular structure of matrix G, which is useful to compute the solution of the master equation.

Theorem 3.1. *Consider a system of q reactions involving M independent species and let β_i be the vectors of stoichiometric coefficients, for $i = 1, \ldots, q$. If $\beta_{ij} \in \{-1, 0, 1\}$ for all i, j (unitary steps), then G is block-tridiagonal, and all the non-zero blocks of G are in turn block-tridiagonal with the same structure of G.*

Theorem 3.1 claims that G is recursively block-tridiagonal for generalized (not necessarily orthogonal) one-step processes. It can be also shown that the assumptions of non-orthogonal one-step process are the mildest conditions preserving the recursive block-tridiagonal structure of G. Before proving Theorem 3.1, we specialize the result to the case of classical (orthogonal) one-step processes.

Corollary 3.2. *Consider a system of q reactions fulfilling the assumptions of Theorem 3.1. If the M species are orthogonal, then G is block-tridiagonal, with the off-diagonal blocks being diagonal matrices and with the diagonal blocks of G being in turn block-tridiagonal with the same structure of G.*

We skip the proof of Corollary 3.2 and conclude the section with a concise proof of Theorem 3.1.

Proof. [Proof of Theorem 3.1] Consider the states corresponding to the vector $\mathcal{P}_{n_1,\ldots,n_i}$, for a given set of values $n_1,\ldots,n_i$. Then, roughly speaking, the unitary-step assumption implies that a non-zero contribution to the right-hand side of Eq.(8) for the probabilities included in the vector $\mathcal{P}_{n_1,\ldots,n_{i+1}}$, for any value of n_{i+1}, is given only by the probabilities in vectors $\mathcal{P}_{n_1,\ldots,n_{i+1}-1}$, $\mathcal{P}_{n_1,\ldots,n_{i+1}}$ and $\mathcal{P}_{n_1,\ldots,n_{i+1}+1}$. Hence the set of rows in (12) corresponding to $\mathcal{P}_{n_1,\ldots,n_{i+1}}$ will just have two non-zero blocks in addition to the diagonal block.

In order to keep the notation compact and illustrate a constructive procedure for building G, we introduce the *matrix builder* Φ_ν [7, 8]. The entries of Φ_ν are sequences of equally dimensioned square matrices. The length of the first, second and third entry is set by the parameter ν at $\nu-1$, ν and $\nu-1$, respectively. By naming A_i, B_i and C_i the 3 entries of the matrix builder, it provides the following *block-tridiagonal* matrix (to avoid confusion, a further parameter is added, explicitly indicating the index of the sequences taken as inputs of the matrix builder):

$$\Phi_\nu(A_i,B_i,C_i;i)=\begin{bmatrix} B_1 & C_2 & 0 & \cdots & 0 \\ A_1 & B_2 & C_3 & 0 & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & A_{\nu-2} & B_{\nu-1} & C_\nu \\ 0 & \cdots & 0 & A_{\nu-1} & B_\nu \end{bmatrix}.$$

Therefore, define the scalar quantities:

$$g^e_{n_1,\ldots,n_M} = \sum_{(\alpha_1,\ldots,\alpha_M)\in\{\beta_1,\ldots,\beta_q\}} g^{\alpha_1,\ldots.\alpha_M}_{n_1,\ldots,n_M},$$

which allow to initialize the recursive computation of the following block-tridiagonal matrices by means of the matrix builder:

$$G^{\alpha_1,\ldots,\alpha_i}_{n_1,\ldots,n_i} = \Phi_{N_{i+1}}\Big(G^{\alpha_1,\ldots,\alpha_i,1}_{n_1,\ldots,n_i,n_{i+1}}, G^{\alpha_1,\ldots,\alpha_i,0}_{n_1,\ldots,n_i,n_{i+1}}, G^{\alpha_1,\ldots,\alpha_i,-1}_{n_1,\ldots,n_i,n_{i+1}}; n_{i+1}\Big)$$

for $i=1,\ldots,M-2$, with initial condition given by:

$$G^{\alpha_1,\ldots,\alpha_{M-1}}_{n_1,\ldots,n_{M-1}} = \Phi_{N_M}\Big(G^{\alpha_1,\ldots,\alpha_{M-1},1}_{n_1,\ldots,n_{M-1},n_M}, -g^e_{n_1,\ldots,n_M}, G^{\alpha_1,\ldots,\alpha_{M-1},-1}_{n_1,\ldots,n_{M-1},n_M}; n_M\Big).$$

The last step of the recursion provides the matrix:

$$G = \Phi_{N_1}(G^1_{n_1}, G^0_{n_1}, G^{-1}_{n_1}; n_1),$$

which concludes the proof. ∎

4. Structural properties and efficient solution of the CME

In this section, we show some interesting properties of matrix G that will reveal to be useful to compute classical tasks as the computation of the stationary solution of the CME. From a theoretical point of view, the explicit solution of the CME at time t is:

$$\mathcal{P}(t) = e^{Gt}\mathcal{P}_{init}, \tag{13}$$

where $\mathcal{P}_{init}$ is the initial distribution at time $t_0 = 0$. Note that e^{Gt} is not block-tridiagonal, in general, so the exact analytical solution of the master equation at any time t does not take any advantages of the one-step assumption. Furthermore, the large dimension of matrix G in general cases makes the computation of e^{Gt} a hard task. One way to face the numerical problem could be to approximate Eq.(13) by means of the Taylor series expansion, truncated for a sufficiently large index, so that one can take advantage of the sparsity of matrices G^k, whose submatrices are block-$(2k+1)$-diagonal. This will be addressed in further work.

An easier and much more studied task is the computation of the stationary distribution of the CME [22]. In fact, this quantity provides fundamental information for understanding the equilibrium condition of a network of chemical reactions. Starting from an initial condition $\mathcal{P}_{init}$, it can be defined by means of the classical steady-state condition:

$$\mathcal{P}_{ss} \doteq \lim_{t \to +\infty} \mathcal{P}(t) = \lim_{t \to +\infty} e^{Gt} \mathcal{P}_{init}.$$

A different way to compute $\mathcal{P}_{ss}$ avoiding the computation of the matrix exponential e^{Gt} suitably exploits the stationary distribution properties: it is a probability vector that satisfies the following steady-state conditions

$$\begin{cases} G\mathcal{P} & = 0 \\ \mathbf{1}^T \mathcal{P} & = 1 \\ \mathcal{P} & \geq 0 \end{cases} \tag{14}$$

where $\mathbf{1} = (1 \cdots 1)^T$. We point out that, instead of calculating the null space in (14), one can take advantage of the some interesting properties of matrix G, coming from the *Algebraic Graph Theory* [5].

First, we need to recall the concept of weighted directed graph (digraph).

Definition 4.1. *A weighted digraph is a triple (V, E, A), where $V = \{v_k\}$ is a set of vertices (or nodes), $E \subseteq V \times V$ is a set of ordered pairs of vertices called edges (or links), and A is a weighted adjacency matrix such that, for any pair (i, j), the entry A_{ij} is strictly positive if (v_i, v_j) is an edge, whilst $A_{ij} = 0$ otherwise.*

Note that the set of edges E can be derived from matrix A and can be therefore omitted in the previous definition. A very important matrix related to a weighted digraph is the Laplacian matrix.

Definition 4.2. *[5] The Laplacian of the digraph (V, E, A) is defined as*

$$L_{ij} = \begin{cases} \sum_k A_{ik} & i = j \\ -A_{ij} & otherwise. \end{cases}$$

We now introduce a formal graph-theoretical interpretation of a Markov process describing a network of chemical reactions.

Definition 4.3. *The digraph associated to a continuous-time discrete-state stochastic Markov process is a weighted digraph (V, E, A), where $V = \{v_k\}$ is a set of vertices, each associated to a discrete state $(n_1, \ldots, n_M)$ of the process, and A is a matrix whose generic element A_{ij} is the propensity $g^{\alpha_1,\ldots,\alpha_M}_{n_1,\ldots,n_M}$ of reaching the state $(n_1 + \alpha_1, \ldots, n_M + \alpha_M)$ from the state $(n_1, \ldots, n_M)$, with $(n_1, \ldots, n_M)$ being the state associated with vertex v_i and $(n_1 + \alpha_1, \ldots, n_M + \alpha_M)$ being the state associated with vertex v_j. The set of edges E is uniquely defined by A and includes all the links (v_i, v_j) with non-zero probability per unit of time of reaching v_j from v_i.*

We assume that the set $V = \{v_k\}$ is ordered with respect to the order of states induced by the recursive construction in Equations (9)–(11). Matrix G can be shown to share most of the properties of the graph Laplacian as a consequence of the following fundamental theorem, which is one of the main results of this work. The theorem, with its consequences, provides a novel characterization (to the best of the authors' knowledge) of the dynamical matrix G of a general CME in terms of well-known results of algebraic-graph theory (see e.g. [5]).

Theorem 4.4. *Let L be the Laplacian of the digraph associated with the Markov process describing the chemical network. Then $G = -L^T$.*

Proof. Consider any row i of the matrix G, corresponding to a state $(n_1, \ldots, n_M)$ of the Markov process. The master equation for such a state is in Eq.(8), hence

$$G_{ii} = - \sum_{(\alpha_1,\ldots,\alpha_M) \in \{\beta_1,\ldots,\beta_q\}} g^{\alpha_1,\ldots,\alpha_M}_{n_1,\ldots,n_M}$$

and

$$G_{ij} = g^{-\alpha_1,\ldots,-\alpha_M}_{n_1+\alpha_1,\ldots,n_M+\alpha_M}$$

for $i \neq j$, where the generic column j is referred to the node $(n_1 + \alpha_1, \ldots, n_M + \alpha_M)$, for some $\alpha_1, \ldots, \alpha_M$. Notice that, from Definitions 4.2 and 4.3, $G_{ii} = -L_{ii}$. Now consider the element $L_{ji} = -A_{ji}$ of the Laplacian which, from Definition 4.3, is (minus) the probability per unit of time of reaching state i, associated to $(n_1, \ldots, n_M)$, from state j, associated to $(n_1 + \alpha_1, \ldots, n_M + \alpha_M)$; hence by Definition 4.3 it is equal to $-g^{-\alpha_1,\ldots,-\alpha_M}_{n_1+\alpha_1,\ldots,n_M+\alpha_M}$, in turn equal to $-G_{ij}$. This concludes the proof. ∎

As an immediate consequence of the previous theorem, G is a Metzler matrix (namely all the off-diagonal components are nonnegative), hence the system in (12) can be regarded as a positive linear dynamical system. Other known properties of G, which we rediscover here as a consequence of Theorem 4.4, are the following:

- $\mathbf{1}^T G = \mathbf{0}^T$, hence G is singular and admits a non-trivial null space;
- all eigenvalues of G have nonpositive real part, ensuring the convergence of the dynamics to the null space;
- $\mathbf{1}^T e^{Gt} = \mathbf{1}^T$ for all t, i.e e^{Gt} is a *column-stochastic* matrix [5]. This ensures that $\mathcal{P}(t)$ is a probability vector (nonnegative entries which add up to 1) at any time t.

The following proposition is one of the main results of this section and provides a necessary and sufficient condition for the existence of a one-dimensional null space, i.e. a unique stationary distribution. The condition is that the digraph associated with the network of reactions has a globally reachable vertex, which is a milder assumption than strong connectivity; that is not a major assumption in networks of chemical reactions, where states can usually jump to adjacent states with non-zero probability per unit of time.

Proposition 4.5. *The stationary distribution of a discrete-state continuous-time Markov process is unique if and only if* $rank(G) = dim(G) - 1$*, namely if and only if the digraph associated with the Markov process has a globally reachable vertex*[1]*. Under these conditions,* 0 *is a single eigenvalue of* G*, with left eigenvector* $\mathbf{1}^T$*. The stationary distribution is unique and is given by* $\mathcal{P}_{ss} = \frac{u_0}{\mathbf{1}^T u_0}$*, where* u_0 *is the right eigenvector corresponding to the eigenvalue* 0*. The second smallest eigenvalue of* G *(also called* algebraic connectivity *of the digraph associated with the Markov process) gives the convergence speed to the stationary distribution.*

The proof of Proposition 4.5 is a consequence of Theorem 4.4 and of known properties of the Laplacian matrix [5], and is therefore omitted. We conclude this section by remarking some computational issues related to G due to its high number of elements. In particular, we recover the one-step assumption to show the complexity reduction (from *cubic* to *quadratic* complexity with respect to the number of rows) in performing the classical Gaussian elimination to solve the equilibrium problem $G\mathcal{P} = 0$. The performance of the algorithm is illustrated in the following proposition and will tested in an example in Section 5.

Proposition 4.6. *For orthogonal one-step processes, the algorithm of Gaussian elimination to solve the stationary equation* $G\mathcal{P} = 0$ *is performed in time* $\mathcal{O}(n^2)$*, where* n *is the number of rows of* G*.*

Proof. For general dense matrices the performance of Gaussian elimination is cubic; in fact, to put G in upper-triangular form, $\mathcal{O}(n)$ elementary operations need to be computed on each of $\mathcal{O}(n^2)$ elements below the main diagonal of G. In orthogonal one-step processes, matrix G is very sparse, with a maximum number of non-zero elements per row equal to $2M + 1$; hence, for any fixed number of species M, the number of elements below the main diagonal is $\mathcal{O}(n)$. Hence the total number of elementary operations is $\mathcal{O}(n^2)$, concluding the proof. ∎

5. Simulations

We consider again Example 3 in Section 2 (MicroRNA Toggle Switch) [11]. The equations describe an orthogonal one-step process described by a bivariate master equation ($M = 2$) with transition probabilities

$$\begin{cases} g^{1,0}_{n_1,n_2} & = \bar{\alpha} + \frac{k_1 n_1^2}{\Gamma_1 + n_1^2 + \Gamma_2 n_2} \\ g^{0,1}_{n_1,n_2} & = \beta + k_2 n_1 \\ g^{-1,0}_{n_1,n_2} & = \delta n_1 \\ g^{0,-1}_{n_1,n_2} & = \gamma n_2 \end{cases}$$

and zero otherwise (see [11] and references therein for more details). The chosen parameters are $\bar{\alpha} = 1.68$, $\beta = 0.202$, $\delta = 0.2$, $\gamma = 0.2$, $k_1 = 90$, $k_2 = 0.05$, $\Gamma_1 = 10300$, $\Gamma_2 = 1006$. In [11] the stationary distribution for this process was not computed exactly, but the model was reduced to one dimension, by considering a different time scale for the two reactions. In particular, n_2 was considered as a fast variable and the value of its steady state (computed by means of the deterministic concentration equation in Eq.(3) as $n_2 = \frac{\beta + k_2 n_1}{\gamma}$) was substituted into the

[1] A globally reachable vertex v is a vertex of the digraph such that there exists a directed path from any node of the graph to v.

bivariate master equation, obtaining an approximate scalar CME, whose stationary distribution could be computed theoretically. This example was presented in [11] to show the possibility of poor agreement between stochastic simulation and the $1D$ stationary distribution obtained by the steady-state approximation for n_2.

We applied the Gillespie SSA Algorithm [13] (formalized in Algorithm 1) to the described model. We repeated the stochastic simulation for the example above by means of $nMC = 2 \cdot 10^4$

Init: $g_{n_1,n_2,\ldots,n_M}^{\alpha_1,\alpha_2,\ldots,\alpha_M}$ for all n_i, α_i, $i = 1, \ldots, M$;
T (time horizon);
$count_{\max}$ (max number of reactions to observe);
$(n_1, n_2, \ldots, n_M)$ (initial state);
$t = 0$ (initial time) ;
$count = 0$ (counter) ;
Main:
while $(t \leq T) \vee (count \leq count_{\max})$ **do**
  $H := \{g_{n_1,\ldots,n_M}^{\alpha_1,\ldots,\alpha_M} \mid g_{n_1,\ldots,n_M}^{\alpha_1,\ldots,\alpha_M} \in \mathbb{R}^+\}$;
  $a_0 := \sum_{a_i \in H} a_i$;
  Draw r_1, r_2 from the continuous distribution $\mathcal{U}(0,1)$;
  $\tau := (1/a_0) \ln (1/r_1)$;
  Choose μ s.t. $\sum_{v=1}^{\mu-1} a_i < r_2 a_0 \leq \sum_{v=1}^{\mu} a_i$;
  $t := t + \tau$;
  $count := count + 1$;
  $n_i := n_i + \bar{\alpha}_i$, $i = 1, \ldots, M$, with $a_\mu = g_{n_1,\ldots,n_M}^{\bar{\alpha}_1,\ldots,\bar{\alpha}_M}$;
end

Algorithm 1: Gillespie Stochastic Simulation Algorithm (SSA).

Monte Carlo runs of Algorithm 1, with $T_{\max} = 10^3$, $count_{\max} = 10^7$, and plotted the statistics of the occurrences of the steady states for n_1. We compared them to the previously described $1D$ stationary distribution and to the exact solution of $G\mathcal{P} = 0$, obtained by a sparse-matrix implementation of Gaussian elimination for the matrix G (see also Proposition 4.6). The matrix G was built by means of an approximation of the model with a closed system with a sufficiently large number of copies ($N_1 = 300$, $N_2 = 80$), chosen so that the probabilities of generation of new molecules from those levels is negligible. While the Monte Carlo runs of the Gillespie Algorithm took some hours, the exact computation of the $2D$ stationary distribution with our method was computed in just 82 seconds on the Matlab suite through an Intel Core 2 Duo T5500 1.66GHz laptop with 4 GB RAM. The plots in Fig. 2 show the agreement of the statistical estimation with the $2D$ theoretical stationary distribution while showing the mismatch with respect to the $1D$ approximate steady-state distribution.

6. Conclusions

In this work we presented some results on the dynamical properties and the solution of the Chemical Master Equation, with a particular focus on the exact solution of the problem at the equilibrium. The application in real cases appears to be promising in that the approach is accurate and allows a cheap management of the computation resources with respect to the extensive use of stochastic simulation to approximate the stationary distribution.

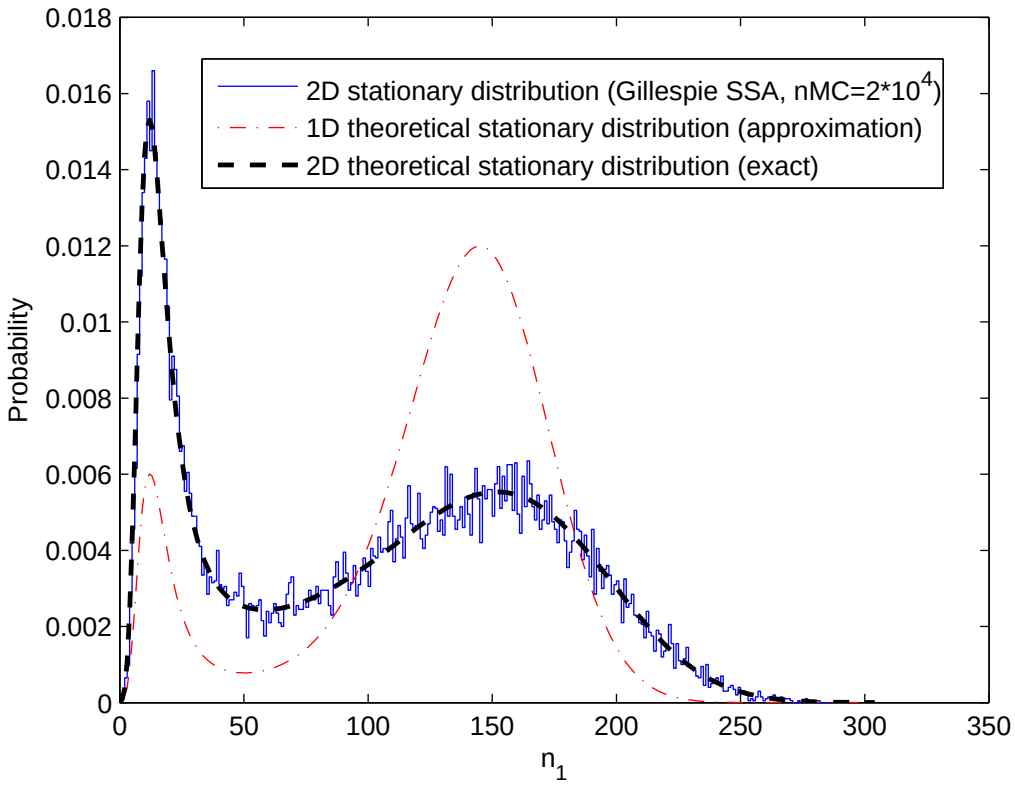

Figure 2: Comparison among the statistics of steady states provided by the Gillespie Algorithm (solid blue line), the 2D theoretical stationary distribution (dashed black line) and the 1D approximate stationary distribution (dash-dotted red line).

Acknowledgment. The authors are grateful to Gastone Castellani for fruitful discussions on the topic of the present paper.

References

[1] R. Bahar, C.H. Hartmann, K.A. Rodriguez, A.D. Denny, R.A. Busuttil, M.E. Doll, R.B. Calder, G.B. Chisholm, B.H. Pollock, C.A. Klein and J. Vijg, Increased cell-to-cell variation in gene expression in ageing mouse heart, *Nature*, 441, 1011–1014, 2006.

[2] A. Bazzani, G. Castellani, E. Giampieri, D. Remondini and L. N Cooper, Bistability in the Chemical Master Equation for Dual Phosphorylation Cycles, *J. Chem. Phys.*, 136(23), 235102, 2012.

[3] L.B. Blyn, B.A. Braaten, C.A. White-Ziegler, D.H. Rolfson and D.A. Low, Phase-variation of pyelonephritis-associated pili in Escherichia coli: evidence for transcriptional regulation, *EMBO J.*, 8(2), 613-620, 1989.

[4] F.J. Bruggeman, N. Blüthgen and H.V. Westerhoff, *Noise management by molecular networks, PLoS Computational Biology*, 5(9), 1-11. 2009.

[5] F. Bullo, J. Cortes and S. Martinez, Distributed Control of Robotic Networks, *Series in Applied Mathematics*, Princeton, 2009, ISBN 978-0-691-14195-4.

[6] Y. Cao, D.T. Gillespie and L.R. Petzold, Efficient step size selection for the tau-leaping simulation method, *J. Chem. Phys.* 124, 044109, 2006.

[7] F. Carravetta, Nearest-neighbour modelling of reciprocal chains, *Stochastics: An International Journal of Probability and Stochastic Processes.* 80 (6), 525-584, 2008.

[8] F. Carravetta, 2D-Recursive Modelling of Homogeneous Discrete Gaussian Markov Fields, *IEEE Transaction on Automatic Control.* 56 (5), 1198-1203, 2011.

[9] P. Cazzaniga, D. Pescini, D. Besozzi, G. Mauri, S. Colombo and E. Martegani, Modeling and stochastic simulation of the Ras/cAMP/PKA pathway in the yeast *Saccharomyces cerevisiae* evidences a key regulatory function for intracellular guanine nucleotides pools, *Journal of Biotechnology*, 133, 377-385, 2008.

[10] T. Gardner, C. Cantor and J. Collins, Construction of a toggle switch in Escherichia coli, *Nature* 403(6767), 339342, 2000.

[11] E. Giampieri, D. Remondini, L. de Oliveira, G. Castellani and P. Lió, Stochastic analysis of a miRNA-protein toggle switch, *Molecular Biosystems*, 7(10), 2796-2803, 2011.

[12] M.A. Gibson and J. Bruck, Efficient exact stochastic simulation of chemical systems with many species and many channels, *J. Phys. Chem.* 104, 18761889, 2000.

[13] D. T. Gillespie, Exact Stochastic Simulation of Coupled Chemical Reactions, *The Journal of Physical Chemistry* 81(25), 23402361, 1977.

[14] D.T. Gillespie, A rigorous derivation of the chemical master equation, *Physica A*, 188, 404–425, 1992.

[15] D.T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems, *J. Chem. Phys.* 115(4), 1716-1733, 2001.

[16] H. Koeppl, C. Zechner, A. Ganguly, S. Pelet and M. Peter, Accounting for extrinsic variability in the estimation of stochastic rate constants, *Int. J. Robust. Nonlinear Control*, 22, 1103-1119, 2012.

[17] G. Lillacci and M. Khammash, A distribution-matching method for parameter estimation and model selection in computational biology, *Int. J. Robust. Nonlinear Control*, 22, 1065-1081, 2012.

[18] J. Mettetal and A. van Oudenaarden, Necessary noise, *Science*, 317, 463, 2007.

[19] B. Munsky and M. Khammash, The finite state projection algorithm for the solution of the chemical master equation, *J. Chem. Phys.*, 124, 044104, 2006.

[20] B. Munsky and M. Khammash, The finite state projection approach for the analysis of stochastic noise in gene networks, *IEEE Trans. Autom. Control, Special Issue on Systems Biology*, 201-214, 2008.

[21] P. Palumbo, G. Mavelli, L. Farina and L. Alberghina, Networks and circuits in cell regulation, *Biochem. Biophys. Res. Commun.*, 881-886, 2010.

[22] N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, North Holland, third edition, 2007.

[23] C. Zechner, J. Ruess, P. Krenn, S. Pelet, M. Peter, J. Lygeros and H. Koeppl, Moment-based inference predicts bimodality in transient gene expression, *PNAS*, 109(21), 8340-8345, 2012.