Available online at www.sciencedirect.com
ScienceDirect
Electronic Notes in Theoretical Computer Science
ELSEVIER
Electronic Notes in Theoretical Computer Science 171 (2007) 197–208
www.elsevier.com/locate/entcs

Modelling of Biochemical Reactions by Stochastic Automata Networks

Verena Wolf[1]

University of Mannheim, A5 B 119, D-68131 Mannheim, Germany

Abstract

This paper presents a stochastic modelling framework based on stochastic automata networks (SANs) for the analysis of complex biochemical reaction networks. Our approach takes into account the discrete character of quantities of components (i.e. the individual populations of the involved chemical species) and the inherent probabilistic nature of microscopic molecular collisions. Moreover, as for process calculi that have recently been applied to systems in biology, the SAN approach has the advantage of a modular design process being adequate for abstraction purposes. The associated composition operator leads to an elegant and compact representation of the underlying continuous-time Markov chain in form of a Kronecker product. SANs have been extensively used in performance analysis of computer systems and a large variety of numerical and simulative analysis algorithms exist. We illustrate that describing a biochemical reaction network by means of a SAN offers promising opportunities to get insight into the quantitative behaviour of systems in biology while taking advantage of the benefits of a compositional modelling approach.

Keywords: Biochemical Reactions, Stochastic Automata Networks, Markov Chain

1 Introduction

In recent years, computational modelling of large networks of biochemical reactions has become increasingly important and is a main challenge in systems biology. Stochastic approaches have emerged as a significant alternative to the classical deterministic approaches for quantitative analysis of intracellular dynamics. In this area Gillespie's simulation algorithm [13] is very popular and it is based on a framework that accounts for populations of molecules and reflects stochastic phenomena caused by the randomness of molecule collisions. The underlying model is that of a continuous-time Markov chain (CTMC) [4,15] originally used to study the performance behaviour of parallel and distributed computer systems.

As opposed to the stochastic methodology the extremely successful deterministic approach for modelling and analysis of complex biochemical reactions is based on the law of mass action, an empirical law leading to chemical kinetics rate equation

[1] Email: wolf@informatik.uni-mannheim.de

doi:10.1016/j.entcs.2007.05.017

models. This macroscopic approach provides a complete picture of concentrations of involved species over time but ignores the discrete character of quantities of components and the inherent probabilistic nature of microscopic molecular collisions. Especially for the regulation of gene expression where transcription factors interact with DNA binding sites, random fluctuations are inevitable and the macroscopic view using rate equations turns out to be less adequate than the stochastic approach. On the contrary, in large scale systems, i.e. systems with large populations of interacting species, the random behaviour averages out. For a more detailed discussion see [26,25,3] and the references therein.

Recently, formal system description techniques, such as process algebras and petri nets, originating in computer science, have been applied to the modelling of complex biological systems [22,23,7,17,2,11,18,16]. These approaches offer facilities to reason about molecular networks in a compositional way such that models remain open and allow an incremental description. Hence, it is possible to add data to an existing model without the need of building a completely new one. Various techniques are provided in this area to consider the model on different abstraction levels reducing its size (possibly on cost of loosing information). According to the given semantics the corresponding low-level descriptions are generated automatically (e.g. [6]) mostly in form of transition systems or stochastic simulation techniques are applied directly to the high-level language representations (e.g. [20]). In addition, high-level languages offer the possibility to exploit the regular structure of biological systems during analysis.

All stochastic frameworks have in common that the underlying model is a CTMC and the attraction for numerical analysis of CTMCs lies in that exact results are provided compared to simulative techniques that come along with the difficulty of statistical errors. Unfortunately, numerical analysis requires the generation of a transition matrix being liable to encounter state space explosion problems.

In process calculi the usual way of decomposing a network of biochemical reactions into components is such that each molecule corresponds to a process that is basically a finite state machine describing the possible behaviours and conformations of the molecule. For example, an enzyme molecule can be free, i.e. able to bind to a certain substrate molecule, or bound to a complex capable of dissociation. The molecule can also degenerate meaning that its state machine moves to a deadlock state. Interactions between molecules are modelled via *parallel synchronisation* of processes (with respect to certain behaviours) or *channels* are used for communication between processes. Information about the current population of a species are encoded only indirectly leading to difficulties in the quantitative analysis especially if *reaction rates*, determining the speed of the reactions and depending on substrate populations, come into play.

In this work, we propose a stochastic modelling approach for biochemical reaction networks based on *stochastic automata networks* (SANs) [21,12] which are used to efficiently model very large CTMCs whose state space is on the order of millions. The basis of the SAN formalism is a generalised tensor algebra with a Kronecker product operation that correctly reflects the calculation of reaction rates

in biochemical reaction networks. SANs have been extensively used in performance analysis of computer systems and a large variety of numerical or simulative analysis algorithms exist and are implemented in tools like PEPS [1], APNN [5] and SMART [9]. The advantages of the SAN approach are the modular design process being adequate for abstraction purposes, the elegant and compact representation of the Markov chain using a well-known formalism and the direct encoding of the discrete quantities of interest (i.e. the respective populations of chemical species). The underlying matrix representation keeps track of the network structure facilitating numerical analysis algorithms that overcome the state space explosion problem with efficient storage mechanisms.

The number of factors used in the Kronecker representation of a biochemical reaction network grows only linearly in the number of involved species and reactions and is independent of the population size whereas the sizes of the individual matrices are depending on the maximum numbers of molecules of the participating chemical species. For example, the enzyme-catalysed substrate conversion, used as running example throughout the paper, can be described by four automata and three different interactions.

To the best of our knowledge, the SAN formalism has not been used to construct stochastic models for systems in biology. An exception is [19] where T SAN descriptors are constructed to represent homogeneous clusters of intracellular Ca^{2+} channels.

The paper is organised as follows. In Section 2 we give some preliminary definitions related to the tensor product and Section 3 formally describes the underlying model. The SAN representation is derived in Section 4. Finally, Section 5 concludes the paper and gives directions of further research.

2 Preliminaries

We recall some useful definitions related to the tensor product. The Kronecker (tensor) product of two matrices $A \in \mathbb{R}^{n_1 \times m_1}$ and $B \in \mathbb{R}^{n_2 \times m_2}$ is defined as $C = A \otimes B$, $C \in \mathbb{R}^{n_1 n_2 \times m_1 m_2}$ where

$$C(k_1 \cdot n_2 + k_2, l_1 \cdot m_2 + l_2) = A(k_1, l_1)B(k_2, l_2)$$

$(1 \leq k_h \leq n_h, 1 \leq l_h \leq m_h, h \in \{1, 2\})$. We consider a simple example with

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \text{ and } B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \\ b_{31} & b_{32} & b_{33} & b_{34} \end{pmatrix}.$$

The tensor product $C = A \otimes B$ is given by

$$C = \begin{pmatrix} a_{11}B & a_{12}B \\ a_{21}B & a_{22}B \end{pmatrix}$$

$$= \left(\begin{array}{cccc|cccc} a_{11}b_{11} & a_{11}b_{12} & a_{11}b_{13} & a_{11}b_{14} & a_{12}b_{11} & a_{12}b_{12} & a_{12}b_{13} & a_{12}b_{14} \\ a_{11}b_{21} & a_{11}b_{22} & a_{11}b_{23} & a_{11}b_{24} & a_{12}b_{21} & a_{12}b_{22} & a_{12}b_{23} & a_{12}b_{24} \\ a_{11}b_{31} & a_{11}b_{32} & a_{11}b_{33} & a_{11}b_{34} & a_{12}b_{31} & a_{12}b_{32} & a_{12}b_{33} & a_{12}b_{34} \\ \hline a_{21}b_{11} & a_{21}b_{12} & a_{21}b_{13} & a_{21}b_{14} & a_{22}b_{11} & a_{22}b_{12} & a_{22}b_{13} & a_{22}b_{14} \\ a_{21}b_{21} & a_{21}b_{22} & a_{21}b_{23} & a_{21}b_{24} & a_{22}b_{21} & a_{22}b_{22} & a_{22}b_{23} & a_{22}b_{24} \\ a_{21}b_{31} & a_{21}b_{32} & a_{21}b_{33} & a_{21}b_{34} & a_{22}b_{31} & a_{22}b_{32} & a_{22}b_{33} & a_{22}b_{34} \end{array}\right).$$

Some important properties of tensor products and additions are

- Associativity: $A \otimes (B \otimes C) = (A \otimes B) \otimes C$
- Distributivity over (ordinary matrix) addition:
  $(A + B) \otimes (C + D) = (A \otimes C) + (B \otimes C) + (A \otimes D) + (B \otimes D)$
- Compatibility with (ordinary matrix) multiplication:
  $(A \times B) \otimes (C \times D) = (A \otimes C) \times (B \otimes D)$
- Compatibility with (ordinary matrix) inversion:
  $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

3 Biochemical Reactions and Markov Chains

In the following, we describe how networks of biochemical reactions such as signalling or metabolic pathways can be mapped onto a stochastic discrete-event model, more precisely a continuous-time Markov chain (CTMC for short). Each reaction between different molecular species in the network corresponds to an event and the state space of the model is characterised by the corresponding populations.

Formally, if J is the number of participating substrates (i.e. the different molecular species $S_1, S_2, \ldots, S_J$), we define $X(t) = \big(X_1(t), X_2(t), \ldots, X_J(t)\big)$ as a vector such that $X_j(t), j \in \{1, 2, \ldots, J\}$ is a discrete random variable describing the number of molecules of type S_j at time instant $t \geq 0$. If $X(t) = \bar{x} := (x_1, x_2, \ldots, x_J) \in \mathbb{N}^J$, the system is in state $\bar{x}$ at time t meaning that for each S_j the current number of molecules is x_j. The number of molecules of each species is bounded since either the starting substance is exhausted or an equilibrium is reached. We define $n_j \in \mathbb{N}$ as the maximum number of molecules of S_j, which implies that the finite state space of the model is given by

$$\mathcal{X} := \{(x_1, x_2, \ldots, x_J) \in \mathbb{N}^J \mid 0 \leq x_j \leq n_j, j \in \{1, 2, \ldots, J\}\}.$$

The state space size $|\mathcal{X}| = (n_1 + 1) \cdot (n_2 + 1) \cdots (n_J + 1)$ grows exponentially in the

number of species (also known as the problem of state space explosion). Moreover, the maximum number of molecules of a species can be very large.

The system evolves from one state to another by a set of transitions that are related to the events, i.e. to the chemical reactions. We are interested in the temporal interaction amongst large numbers of molecules to understand the functional activity of the network. According to the common chemical kinetics, we associate a rate with units of reciprocal time with each transition leading to a representation in terms of a CTMC, that is a stochastic process with discrete state space where the future evolution of the process depends only on the current state (and not on the process history or the current time instant).

The starting point for the model construction is a set $\{R_i \mid 1 \leq i \leq I\}$ of biochemical reactions. Each reaction R_i is given by

$$s_{i_1}S_{i_1} + s_{i_2}S_{i_2} + \ldots + s_{i_m}S_{i_m} \xrightarrow{c_i} s_{i_{m+1}}S_{i_{m+1}} + s_{i_{m+2}}S_{i_{m+2}} + \ldots + s_{i_l}S_{i_l}, \qquad (1)$$

where $0 \leq m \leq l \leq 2J$ and $s_{i_1}, \ldots, s_{i_l} \in \mathbb{N}$ are stoichiometric coefficients. We call the left-hand substrates *reactants* and the substrates on the right-hand are called *products* if they do not appear on the left and *catalysts* otherwise. Equation (1) describes how the reaction affects the population vector, i.e. for each $h \in \{1, \ldots, m\}$ the number of molecules of chemical species S_{i_h} that are used up is s_{i_h} and for $h \in \{m+1, \ldots, l\}$ the number of molecules of chemical species S_{i_h} that are produced by the reaction is s_{i_h}. The *(stochastic reaction) rate constant* $c_i \in \mathbb{R}_{>0}$ determines the speed of the reaction in a way explained below.

In most cases, the number of reactants and the number of products and catalysts is small, i.e. $m \leq 2$ and $l - m \leq 2$ and also the stoichiometric coefficients are mostly equal to one. All other reaction types are extremely rare because the probability that three or more independent molecules collide at the same time or within a small time interval is very small.

We consider two running examples in this paper:

Example 3.1 The enzyme-catalysed substrate conversion [2]

$$S_1 + S_2 \underset{c_2}{\overset{c_1}{\rightleftharpoons}} S_3 \xrightarrow{c_3} S_1 + S_4$$

of a substrate S_2 into a product S_4 via an *enzyme-substrate complex* S_3, catalysed (accelerated) by an enzyme S_1. Here, the number of participating species is $J = 4$ and the number of reactions is $I = 3$. All stoichiometric coefficients are equal to one.

Example 3.2 We consider a two-gene positive feedback loop with $I = 10$ reactions

[2] Reaction sets are often written as chains where bidirectional arrows are used for reactions that can happen in both directions.

involving $J = 4$ chemical species [3]:

$$
\begin{aligned}
R_l^{stcr} &: \emptyset \xrightarrow{c_l^{\mathsf{stcr}}} \mathsf{mRNA}_l && \text{(slow transcription of } \mathsf{mRNA}_l\text{)},\\
R_l^{ftcr} &: \mathsf{Protein}_j \xrightarrow{c_l^{\mathsf{ftcr}}} \mathsf{mRNA}_l + \mathsf{Protein}_j && \text{(fast transcription of } \mathsf{mRNA}_l\text{)},\\
R_j^{tsl} &: \mathsf{mRNA}_j \xrightarrow{c_j^{\mathsf{tsl}}} \mathsf{mRNA}_j + \mathsf{Protein}_j && \text{(translation of } \mathsf{Protein}_j\text{)},\\
R_j^{mdeg} &: \mathsf{mRNA}_j \xrightarrow{c_j^{\mathsf{mdeg}}} \emptyset && \text{(degradation of } \mathsf{mRNA}_j\text{)},\\
R_j^{pdeg} &: \mathsf{Protein}_j \xrightarrow{c_j^{\mathsf{pdeg}}} \emptyset && \text{(degradation of } \mathsf{Protein}_j\text{)},
\end{aligned}
$$

$(j, l \in \{1, 2\}, j \neq l)$. This set of reactions describes a regulatory network controlling the transcription of two genes into mRNA and the translation of the two corresponding types of mRNA into proteins. The transcription of gene 1 is accelerated by the existence of $\mathsf{Protein}_2$ molecules (reaction R_1^{ftcr}), which are translation products of mRNA_2 (R_2^{tsl}), and vice versa, $\mathsf{Protein}_1$, resulting from the translation of mRNA_1 (R_1^{tsl}), acts as regulatory protein for the transcription of gene 2 (R_2^{ftcr}). The transcription is slow if no activating protein molecules are available (R_l^{stcr}) and molecules degrade according to the reactions R_l^{mdeg} and R_l^{pdeg}.

We now consider the general case, i.e. we assume reaction R_i is described by (1). Let $\mathsf{REA}(i)$ ($\mathsf{PRO}(i)$) be the set of reactants (products, resp.), i.e. $\mathsf{REA}(i) := \{S_{i_1}, \ldots, S_{i_m}\}$ and $\mathsf{PRO}(i) := \{S_{i_{m+1}}, \ldots, S_{i_l}\}$. Furthermore, let $\mathsf{CAT}(i) := \mathsf{REA}(i) \cap \mathsf{PRO}(i)$ be the subset of reactants that act as catalysts, i.e. each species $S_h \in \mathsf{CAT}(i)$ occurs also on both hands of the reaction [4].

A transition describes a rule how the system evolves from the current state $\bar{x}$ to another state depending on $\bar{x}$ and the reaction type. Direct successor states are given by the function $\mathsf{next}_i : \mathcal{X} \to \mathcal{X}$ that returns the next state of $\bar{x}$ if reaction R_i happens. We define $\mathsf{next}_i(\bar{x}) = \mathsf{next}_i(x_1, x_2, \ldots, x_J) := \bar{x}$ if there exists some $S_h \in \mathsf{REA}(i)$ with $x_h < s_{i_h}$ or $S_h \in \mathsf{PRO}(i) \setminus \mathsf{CAT}(i)$ with $x_h > n_h - s_{i_h}$, i.e. reaction R_i does not take place if not enough reactant molecules are left or if one of the produced substrates exceeds its maximum number of molecules via R_i. Otherwise, i.e. if R_i can take place, we put $\mathsf{next}_i(x_1, x_2, \ldots, x_J) = (x'_1, x'_2, \ldots, x'_J)$ where

$$
x'_j = \begin{cases} x_j & \text{if } S_j \notin \big(\mathsf{REA}(i) \cup \mathsf{PRO}(i)\big) \text{ or } S_j \in \mathsf{CAT}(i),\\ x_j - s_{i_j} & \text{if } S_j \in \mathsf{REA}(i) \setminus \mathsf{CAT}(i),\\ x_j + s_{i_j} & \text{if } S_j \in \mathsf{PRO}(i) \setminus \mathsf{CAT}(i). \end{cases}
$$

The population of S_j remains unchanged if either S_j does not take part in R_i or acts as a catalyst, as opposed to the second case in the definition where S_j is a reactant that is consumed in R_i. In the last case, the population of S_j increases by

[3] The $\emptyset$ symbol on the left-hand (right-hand) indicates that no reactant (no product) is needed (is produced, resp.).

[4] We assume w.l.o.g. that the two stoichiometric coefficients of a catalyst have the same value.

s_{i_j}. In a similar way, we define the uniquely determined predecessor $\mathsf{pred}_i(\bar{x})$ that corresponds to reaction R_i such that $\mathsf{next}_i(\mathsf{pred}_i(\bar{x})) = \bar{x}$.

The *propensity function* $\mathsf{rate}_i : \mathcal{X} \to \mathbb{R}_{\geq 0}$ that returns the transition rate of reaction R_i and state $\bar{x} = (x_1, x_2, \ldots, x_J)$ is defined by

$$\mathsf{rate}_i(\bar{x}) = \begin{cases} c_i \cdot \prod\limits_{S_h \in \mathsf{REA}(i)} \binom{x_h}{s_{i_h}} & \text{if } \mathsf{next}_i(\bar{x}) \neq \bar{x}, \\ 0 & \text{otherwise,} \end{cases} \tag{2}$$

i.e. the transition rates of the underlying CTMC are proportional to the distinct combinations of reactant molecules. For details and a rigorous formal justification see [13,15].

The probability of leaving $\bar{x}$ within a small time interval of length Δt via a reaction of type R_i is given by $\mathsf{rate}_i(\bar{x})\Delta t$. Correspondingly, the probability of staying in $\bar{x}$ within this interval is given by $1 - \mathsf{rate}(\bar{x})\Delta t$ where the *exit rate* $\mathsf{rate}(\bar{x}) := \mathsf{rate}_1(\bar{x}) + \mathsf{rate}_2(\bar{x}) + \ldots + \mathsf{rate}_I(\bar{x})$ equals the sum of all outgoing rates of $\bar{x}$. The value $1/\mathsf{rate}(\bar{x})$ is the mean sojourn time in $\bar{x}$. Let $p_t(\bar{x})$ be the probability that $X(t) = \bar{x}$. Then [5]

$$p_{t+\Delta t}(\bar{x}) = (1 - \mathsf{rate}(\bar{x})\Delta t) \cdot p_t(\bar{x}) + \textstyle\sum_{i=1}^{I} \mathsf{rate}_i(\mathsf{pred}_i(\bar{x}))\Delta t \cdot p_t(\mathsf{pred}_i(\bar{x})).$$

This leads to the derivation of p_t given by differential equations

$$\dot{p}_t = \frac{d}{dt}\, p_t = \lim_{\Delta t \to 0} \frac{p_{t+\Delta t} - p_t}{\Delta t} = Q p_t \tag{3}$$

where $p_t \in \mathbb{R}^n_{\geq 0}$ is the vector with entries $p_t(\bar{x})$ and $Q \in \mathbb{R}^{|\mathcal{X}| \times |\mathcal{X}|}$ is defined by [6]

$$Q(\bar{x}, \bar{x}') := \begin{cases} -\mathsf{rate}(\bar{x}), & \text{if } \bar{x} = \bar{x}', \\ \sum\limits_{i:\mathsf{next}_i(\bar{x}) = \bar{x}'} \mathsf{rate}_i(\bar{x}), & \text{if } \bar{x}' \neq \bar{x}, \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

The CTMC $X(t)$ is uniquely described by the *(infinitesimal) generator matrix* Q and an initial distribution (cf. [4,10]). In general, the stochastic interpretation of chemical equations in the style of (1) always yields a CTMC as indicated in [13,14].

Example 3.3 In the enzyme-catalysed substrate conversion of Example 3.1 we assume that initially the system is in state $X(0) = (200, 3000, 0, 0) =: \bar{x}^{(0)}$ which means that we start with 200 enzyme molecules and 3000 molecules of the substrate S_2. Then state $\bar{x}^{(0)}$ can only be left via reaction

$$R_1 : \; S_1 + S_2 \xrightarrow{c_1} S_3$$

[5] The function $p_t(\cdot)$ is depending on the initial state $X(0) = x^{(0)}$ of the system.

[6] We assume that the states space is mapped to $\mathbb{N}$.

with $\mathsf{rate}_1(\bar{x}^{(0)}) = 200 \cdot 3000 \cdot c_1$ by entering state $\mathsf{next}_1(\bar{x}^{(0)}) = (199, 2999, 1, 0)$ which means that one enzyme-substrate complex has been formed. Now, state $(199, 2999, 1, 0)$ has three outgoing transitions, one via R_1 with rate $199 \cdot 2999 \cdot c_1$ to state $(198, 2998, 2, 0)$, one via

$$R_2 : \; S_3 \xrightarrow{c_2} S_1 + S_2$$

with $\mathsf{rate}_2(199, 2999, 1, 0) = 1 \cdot c_2$ back to the state $\mathsf{next}_2(199, 2999, 1, 0) = \bar{x}^{(0)}$ and one transition to state $\mathsf{next}_3(199, 2999, 1, 0) = (200, 2999, 0, 1)$ with $\mathsf{rate}_3(199, 2999, 1, 0) = 1 \cdot c_3$ which means that a dissociation of the complex molecule into one enzyme molecule and one product molecule via reaction

$$R_3 : \; S_3 \xrightarrow{c_3} S_1 + S_4$$

happened. The exit rate of state $(198, 2998, 2, 0)$, for instance, is given by $\mathsf{rate}(198, 2998, 2, 0) = 198 \cdot 2998 \cdot c_2 + 2 \cdot (c_1 + c_3)$.

The generator matrix Q for a network of biochemical reactions is always sparse since each state has at most only I transitions. All row sums are zero and the negative exit rates appear on the main diagonal. However, this representation yields to a mathematical treatment in terms of matrix operations which is advantageous for the realization of analysis algorithms and the definition of composition operations that match operators of a specific matrix algebra (cf. Section 4).

Stochastic systems in general, and in particular Markov chains, are analysed with respect to their temporal evolution where one distinguishes *transient* and *steady-state* analysis. The latter refers to systems in equilibrium whereas the former refers to the phase where an equilibrium has not yet been reached. A large amount of work exists on the numerical solution of Markov chains [24], where numerical solution means to compute probability distributions, either time-dependent transient distributions or steady-state distributions. Different quantitative measures can be derived from the transient and the stationary distribution of models of biochemical reaction network. For example, one might be interested in the expected number of molecules of a certain substrate S_j at time instant t, in the number of molecules of each substrate in the limit or in the expected time until the population of a substrate reaches a certain threshold.

4 Kronecker Representations for Markov Models

A SAN consists of a number of individual stochastic automata that operate more or less independently of each other. Our idea is to construct an automaton for each chemical species which counts the number of corresponding molecules. The rates at which each automaton increments its counter are local in the sense that they constitute the multiplicative factor the corresponding species contributes to the overall reaction rate.

Several types of matrices are needed to give Kronecker representations for biochemical reaction networks. They all have in common that they are sparse, more precisely, all entries are zero except the entries of one of the diagonals. Hence, each

matrix is fully described by its size, a vector and a variable $d \in \mathbb{Z}$ that determines at which diagonal the vector appears. For a reactant S_j we define the vector

$$\mathsf{dep}_j := \left(\binom{0}{s_{i_j}}, \binom{1}{s_{i_j}}, \ldots, \binom{n_j}{s_{i_j}}\right) \in \mathbb{N}^{(n_j+1)}$$

that appears at the d-th upper diagonal of the matrix for S_j where d equals the stoichiometric coefficient s_{i_j}. The vector dep_j contains the factors S_j contributes to the calculation of the transition rate of reaction R_i (compare Equation (2)). If S_j is not a reactant the associated vector is $\mathsf{ind}_j = (1, 1, \ldots, 1) \in \mathbb{N}^{(n_j+1)}$. This ensures that the transition rate of the reaction is independent of the current population of S_j, i.e. the contributed factor is one. If the reaction increases (decreases) the population of S_j by s_{i_j}, the corresponding vector appears at the $|d|$-th upper (lower) diagonal, i.e. $d = s_{i_j}$ ($d = -s_{i_j}$, respectively). If the population of S_j remains unchanged by the reaction we set $d = 0$ and the main diagonal of the matrix contains the non-zero entries.

The matrices $\mathsf{Dep}_j^{(d)}$ and $\mathsf{Ind}_j^{(d)}$ of size $(n_j + 1) \times (n_j + 1)$ are defined by $\mathsf{Dep}_j^{(d)}(k, l) = \mathsf{dep}_j(k)$ and $\mathsf{Ind}_j^{(d)}(k, l) = 1$ if $(k + d) = l$ $(1 \leq k, l \leq (n_j + 1),\ -n_j \leq d \leq n_j)$ and all remaining entries are zero. We have, for instance,

$$\mathsf{Ind}_j^{(1)} := \begin{pmatrix} 0 & 1 & & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ 0 & 0 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 1 \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix}, \quad \mathsf{Dep}_j^{(-1)} := \begin{pmatrix} 0 & 0 & & \ldots & 0 \\ 1 & 0 & 0 & \ldots & 0 \\ 0 & 2 & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \ldots & n_j & 0 \end{pmatrix}.$$

Then the effect of reaction R_i on substrate S_j is given by the matrix $E_j^{(i)}$ where

$$E_j^{(i)} = \begin{cases} \mathsf{Ind}_j^{(0)} & \text{if } j \notin \mathsf{REA}(i) \cup \mathsf{PRO}(i), \\ \mathsf{Dep}_j^{(0)} & \text{if } j \in \mathsf{CAT}(i), \\ \mathsf{Dep}_j^{(-s_{i_j})} & \text{if } j \in \mathsf{REA}(i) \setminus \mathsf{CAT}(i), \\ \mathsf{Ind}_j^{(s_{i_j})} & \text{if } j \in \mathsf{PRO}(i) \setminus \mathsf{CAT}(i). \end{cases}$$

We put $D_j^{(i)} = diag(E_j^{(i)}\mathbf{e}^T)$ where $\mathbf{e}$ is a unit row vector of appropriate size and the operator $diag(v)$ constructs a diagonal matrix from the vector v, i.e. v appears on the main diagonal. Now, let $\mathcal{R} = \{R_1, \ldots, R_I\}$ be a set of reactions such that R_i has the form (1) and $\{S_1, \ldots, S_J\}$ are the different chemical species that are involved in $\mathcal{R}$. The generator matrix Q of the underlying CTMC of $\mathcal{R}$ is given by

$$Q = \sum_{i=1}^{I} c_i \left(\bigotimes_{j=1}^{J} E_j^{(i)} - \bigotimes_{j=1}^{J} D_j^{(i)} \right). \tag{5}$$

	j	1	2	3	4
i	Reaction	mRNA_1	mRNA_2	$\mathsf{Protein}_1$	$\mathsf{Protein}_2$
1	R_1^{stcr}	$E_1^{(1)} = \mathsf{Ind}_1^{(1)}$	$E_2^{(1)} = \mathsf{Ind}_2^{(0)}$	$E_3^{(1)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(1)} = \mathsf{Ind}_4^{(0)}$
2	R_2^{stcr}	$E_1^{(2)} = \mathsf{Ind}_1^{(0)}$	$E_2^{(2)} = \mathsf{Ind}_2^{(1)}$	$E_3^{(2)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(2)} = \mathsf{Ind}_4^{(0)}$
3	R_1^{ftcr}	$E_1^{(3)} = \mathsf{Ind}_1^{(1)}$	$E_2^{(3)} = \mathsf{Ind}_2^{(0)}$	$E_3^{(3)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(3)} = \mathsf{Dep}_4^{(0)}$
4	R_2^{ftcr}	$E_1^{(4)} = \mathsf{Ind}_1^{(0)}$	$E_2^{(4)} = \mathsf{Ind}_2^{(1)}$	$E_3^{(4)} = \mathsf{Dep}_3^{(0)}$	$E_4^{(4)} = \mathsf{Ind}_4^{(0)}$
5	R_1^{tsl}	$E_1^{(5)} = \mathsf{Dep}_1^{(0)}$	$E_2^{(5)} = \mathsf{Ind}_2^{(0)}$	$E_3^{(5)} = \mathsf{Ind}_3^{(1)}$	$E_4^{(5)} = \mathsf{Ind}_4^{(0)}$
6	R_2^{tsl}	$E_1^{(6)} = \mathsf{Ind}_1^{(0)}$	$E_2^{(6)} = \mathsf{Dep}_2^{(0)}$	$E_3^{(6)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(6)} = \mathsf{Ind}_4^{(1)}$
7	R_1^{mdeg}	$E_1^{(7)} = \mathsf{Dep}_1^{(-1)}$	$E_2^{(7)} = \mathsf{Ind}_2^{(0)}$	$E_3^{(7)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(7)} = \mathsf{Ind}_4^{(0)}$
8	R_2^{mdeg}	$E_1^{(8)} = \mathsf{Ind}_1^{(0)}$	$E_2^{(8)} = \mathsf{Dep}_2^{(-1)}$	$E_3^{(8)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(8)} = \mathsf{Ind}_4^{(0)}$
9	R_1^{pdeg}	$E_1^{(9)} = \mathsf{Ind}_1^{(0)}$	$E_2^{(9)} = \mathsf{Ind}_2^{(0)}$	$E_3^{(9)} = \mathsf{Dep}_3^{(-1)}$	$E_4^{(9)} = \mathsf{Ind}_4^{(0)}$
10	R_2^{pdeg}	$E_1^{(10)} = \mathsf{Ind}_1^{(0)}$	$E_2^{(10)} = \mathsf{Ind}_2^{(0)}$	$E_3^{(10)} = \mathsf{Ind}_3^{(0)}$	$E_4^{(10)} = \mathsf{Dep}_4^{(-1)}$

Table 1
The matrices used for the construction of the SAN representing the two-gene positive feedback loop.

Note that subtracting the $D_j^{(i)}$ ensures that Q contains the negative exit rates on the main diagonal and that the row sums are zero. The matrix Q agrees with the generator defined by Equation (4) up to the ordering of states.

Example 4.1 For the enzyme-catalysed substrate conversion of Example 3.1 the Kronecker product representation of the underlying CTMC is defined as follows. The automaton that describes how the population of enzyme molecules (species S_1) is affected by the different reactions is given by the matrices $E_1^{(1)} = \mathsf{Dep}_1^{(-1)}$ (the population decreases by one via R_1) and $E_1^{(2)} = E_1^{(3)} = \mathsf{Ind}_1^{(1)}$ (the population increases by one via R_2 or R_3) where the reaction rates that correspond to R_2 and R_3 are independent of the S_1 population. The reactant S_2 is described by $E_2^{(1)} = \mathsf{Dep}_2^{(-1)}$, $E_2^{(2)} = \mathsf{Ind}_2^{(1)}$ and $E_2^{(3)} = \mathsf{Ind}_2^{(0)}$ (reaction R_3 is independent of S_2 and has no impact on its population). For the enzyme-substrate complex (species S_3) we have $E_3^{(1)} = \mathsf{Ind}_3^{(1)}$ and $E_3^{(2)} = E_3^{(3)} = \mathsf{Dep}_3^{(-1)}$. Finally, for the product S_4 the corresponding matrices are given by $E_4^{(1)} = E_4^{(2)} = \mathsf{Ind}_4^{(0)}$ and $E_4^{(3)} = \mathsf{Ind}_4^{(1)}$, i.e. the population of product molecules is only affected by reaction R_3, the dissociation of the complex. The generator matrix is then given by Equation (5).

Example 4.2 In case of the two-gene positive feedback loop, Table 1 defines the matrices that are needed to construct the generator of the underlying CTMC according to Equation (5).

5 Conclusion

We have presented the construction of stochastic automata network (SAN) descriptors for biochemical reaction networks. Each chemical species is modelled as

a stochastic automaton that counts the number of corresponding molecules. The rates at which each automaton increments its counter are local in the sense that they constitute the multiplicative factor the corresponding species contributes to the overall reaction rate. This ensures a modular design process similar as provided by processes calculi. The discrete quantities of interest (i.e. the respective populations of chemical species) can be directly retrieved from the local states of the automata. The transition matrix of the underlying stochastic model, that is a continuous-time Markov chain, is not generated but implicitly represented as a Kronecker product of (smaller) component matrices. The attraction of our framework lies in that the representation remains compact, even as the number of states of the underlying model begins to explode, and the structure of the network of biochemical reactions is reflected in the SAN description.

A large variety of numerical or simulative analysis algorithms for the solution of SANs exists. Of particular interest are techniques using decision diagrams (e.g. [8]) exploiting the fact that for systems, as considered in this paper, the component matrices are all sparse. As future research we plan case studies with tools supporting Kronecker based representations to analyse gene expression and cell signalling.

Acknowledgement

The author wishes to express her gratitude to Prof. Dr. Udo R. Krieger, University of Bamberg, Germany, who provided the initial stimulus for this paper.

References

[1] A. Benoit, P. Fernandes, B. Plateau, and W. Stewart. The PEPS Software Tool. In *13th International Conference on Modelling Techniques and Tools for Computer Performance Evaluation TOOLS 2003*, pages 98–115, Urbana, Illinois, USA, 2003.

[2] R. Blossey, L. Cardelli, and A. Phillips. A compositional approach to the stochastic dynamics of gene networks. In *Transactions on Computational Systems Biology*, volume 3939 of *LNCS*, pages 99–122. Springer, 2006.

[3] J. M. Bower and H. Bolouri. *Computational Modeling of Genetic and Biochemical Networks*. The MIT Press, 2001.

[4] P. Bremaud. *Markov Chains*. Springer, 1998.

[5] P. Buchholz and P. Kemper. A toolbox for the analysis of discrete event dynamic systems. In *Proceedings of the 11th International Conference on Computer Aided Verification*, pages 483–486, London, UK, 1999. Springer-Verlag.

[6] M. Calder, S. Gilmore, and J. Hillston. Automatically deriving odes from process algebra models of signalling pathways. In *Proceedings of Third International Workshop on Computational Methods in Systems Biology*, pages 204–215. University of Edinburgh, 2005.

[7] M. Calder, S. Gilmore, and J. Hillston. Modelling the influence of rkip on the erk signalling pathway using the stochastic process algebra pepa. *Transactions on Computational Systems Biology*, 2006. to appear.

[8] G. Ciardo and A. Miner. A data structure for the efficient kronecker solution of gspns. In *Proceedings of the The 8th International Workshop on Petri Nets and Performance Models*, page 22, Washington, DC, USA, 1999. IEEE Computer Society.

[9] G. Ciardo and A. Miner. SMART: The stochastic model checking analyzer for reliability and timing. In *Proceedings of the 1st International Conference on Quantitative Evaluation of Systems*, pages 338–339, 2004.

[10] D. Cox and H. Miller. *The Theory of Stochastic Processes.* Chapman and Hall, 1965.

[11] V. Danos and C. Laneve. Formal molecular biology. *Theoretical Computer Science*, 325(1):69–110, 2004.

[12] P. Fernandes, B. Plateau, and W. J. Stewart. Numerical evaluation of stochastic automata networks. In *Proceedings of the Third International Workshop on Modeling, Analysis, and Simulation On Computer and Telecommunication Systems*, pages 179–183, 1995.

[13] D. T. Gillespie. A general method for numerically simulating the time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22:403–434, 1976.

[14] D. T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[15] D. T. Gillespie. *Markov Processes.* Academic Press., 1992.

[16] P. Goss and J. Peccoud. Quantitative modeling of stochastic systems in molecular biology by using stochastic petri nets. In *Proceedings of the National Academy of Science USA*, pages 6750–6755, 1998.

[17] M. Heiner and I. Koch. Petri net based model validation in systems biology. In *Proceedings of the 25th International Conference on Applications and Theory of Petri Nets*, pages 216–237, 2004.

[18] R. Hofestädt and S. Thelen. Quantitative modeling of biochemical networks. *In Silico Biology*, 1:6, 1998.

[19] V. Nguyen, R. Mathias, and G. Smith. A stochastic automata network descriptor for markov chain models of instantaneously coupled intracellular Ca^{2+} channels. *Bulletin of Mathematical Biology*, 67(3):393–432, 2005.

[20] A. Phillips and L. Cardelli. A correct abstract machine for the stochastic pi-calculus. In *Bioconcur.* ENTCS, August 2004.

[21] B. Plateau. On the stochastic structure of parallelism and synchronization models for distributed algorithms. In *In Proceedings of the Sigmetrics Conference on Measurement and Modeling of Computer Systems*, pages 147–154, 1985.

[22] C. Priami, A. Regev, E. Shapiro, and W. Silverman. Application of a stochastic name-passing calculus to representation and simulation of molecular processes. *Inf. Process. Lett.*, 80(1):25–31, 2001.

[23] A. Regev, W. Silverman, and E. Shapiro. Representation and simulation of biochemical processes using the pi-calculus process algebra. In *Pacific Symposium on Biocomputing*, pages 459–470, 2001.

[24] W. Stewart. *Introduction to the Numerical Solution of Markov Chains.* Princeton University Press, 1995.

[25] T. Turner, S. Schnell, and K. Burrage. Stochastic approaches for modelling in vivo reactions. *Computational biology and chemistry*, 28(3):165–178, 2004.

[26] O. Wolkenhauer, M. Ullah, W. Kolch, and K. Cho. Modeling and simulation of intracellular dynamics: Choosing an appropriate framework. *IEEE Transactions on NanoBioscience*, 3(3):200–207, 2004.