Bioinformatics, 33, 2017, i293–i300
doi: 10.1093/bioinformatics/btx249
ISMB/ECCB 2017

A scalable moment-closure approximation for large-scale biochemical reaction networks

Atefeh Kazeroonian,[1,2,3,*] Fabian J. Theis[1,2] and Jan Hasenauer[1,2,*]

[1]Institute of Computational Biology, Helmholtz Zentrum München - German Research Center for Environmental Health, 85764 Neuherberg, Germany, [2]Department of Mathematics, Technische Universität München, 85748 Garching, Germany and [3]Institut für Medizinische Mikrobiologie, Immunologie und Hygiene, Fakultät für Medizin, Technische Universität München, 81675 München, Germany

*To whom correspondence should be addressed.

Abstract

Motivation: Stochastic molecular processes are a leading cause of cell-to-cell variability. Their dynamics are often described by continuous-time discrete-state Markov chains and simulated using stochastic simulation algorithms. As these stochastic simulations are computationally demanding, ordinary differential equation models for the dynamics of the statistical moments have been developed. The number of state variables of these approximating models, however, grows at least quadratically with the number of biochemical species. This limits their application to small- and medium-sized processes.
Results: In this article, we present a scalable moment-closure approximation (sMA) for the simulation of statistical moments of large-scale stochastic processes. The sMA exploits the structure of the biochemical reaction network to reduce the covariance matrix. We prove that sMA yields approximating models whose number of state variables depends predominantly on local properties, i.e. the average node degree of the reaction network, instead of the overall network size. The resulting complexity reduction is assessed by studying a range of medium- and large-scale biochemical reaction networks. To evaluate the approximation accuracy and the improvement in computational efficiency, we study models for JAK2/STAT5 signalling and NFκB signalling. Our method is applicable to generic biochemical reaction networks and we provide an implementation, including an SBML interface, which renders the sMA easily accessible.
Availability and implementation: The sMA is implemented in the open-source MATLAB toolbox CERENA and is available from https://github.com/CERENADevelopers/CERENA.
Contact: jan.hasenauer@helmholtz-muenchen.de or atefeh.kazeroonian@tum.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Cellular mechanisms are subject to inherent biological noise that stems from stochastic events such as bursty gene expression. Due to such stochasticity, isogenic cells can behave differently under identical conditions (Elowitz *et al.*, 2002), giving rise to heterogeneous cell populations. Rather than being a nuisance, biological noise has been proven to be crucial in the functioning of biological systems such as microbial populations and biological tissue (Raj and van Oudenaarden, 2008), e.g. increasing their robustness. Studying the stochasticity of biological processes, therefore, can shed light on their underlying mechanisms and is crucial for a better understanding of their behaviour.

Many biological processes, e.g. gene expression and signal transduction, are modelled as networks of chemical species that undergo chemical reactions. The dynamics of chemical reaction networks, i.e. the temporal evolution of the counts of individual species, is usually described by continuous-time discrete-state Markov chains (CTMCs). The statistics of CTMCs are described by the Chemical Master Equation (CME). As the simulation of the CME is computationally intractable for most processes due to their high- or even infinite-dimensional state space, several methods have been proposed to approximate the statistical moments, e.g. moment-closure approximations (MAs) (Engblom, 2006; Lee *et al.*, 2009) and system-size expansions (Grima, 2010; van Kampen, 2007). These methods yield ordinary differential equations (ODEs) that approximate the temporal evolution of the statistical moments. These ODEs are usually lower-dimensional than the CME, rendering their numerical simulation more tractable. However, already for the analysis of the mean and covariance of the stochastic process, the size of the state space of the approximating models grows quadratically with the number of biochemical species. This limits the application of these methods to small- and medium-scale biochemical reaction networks if the calculation of all statistical moments is required. However interestingly, in a range of applications, including parameter estimation (Fröhlich *et al.*, 2016; Munsky *et al.*, 2009), information about a subset of statistical moments can be sufficient.

In this study, we introduce a scalable second-order moment-closure approximation (s2MA) which is feasible for large-scale biochemical reaction networks. The s2MA is designed for the accurate description of selected statistical moments, including means and variances. We introduce an algorithm that exploits the structure of the reaction network to select the subset of moments which are most relevant for the reliable approximation of means and variances. Using analytical results for toy networks and published biological models, we show the superior scaling of s2MA over other methods for moment approximation, which renders the s2MA tractable for large reaction networks. To assess the accuracy and computational efficiency of s2MA, we simulated several network motifs and models for JAK2/STAT5 and TNF signalling.

2 Approach

We consider a biochemical reaction network of n species, $S_1, \ldots, S_n$, and n_r reactions, $R_1, \ldots, R_{n_r}$. The state of this network is denoted by $\mathbf{X} = (X_1, X_2, \ldots, X_n)^T$ where X_i is the number of molecules of species S_i. Upon the firing of reaction R_r, the state $\mathbf{X}$ undergoes the transition $\mathbf{X} \xrightarrow{a_r} \mathbf{X} + \mathbf{v}_r$, in which ν_r and $a_r(\mathbf{X})$ denote the stoichiometry and the propensity of reaction R_r, respectively. Due to the stochastic nature of chemical reactions, the state vector $\mathbf{X}$ evolves stochastically over time. The probability distribution of $\mathbf{X}$ at time t is denoted by $p(\mathbf{x}|t)$ over all possible states $\mathbf{x}$.

The temporal evolution of the statistical moments of $p(\mathbf{x}|t)$ can be approximated using MAs of different orders. The order of an MA is the highest order of the statistical moments which are modelled. The second-order MA (2MA) is an ODE with $n(n+3)/2$ state variables which describes the dynamics of the mean $\mathbf{m} = \sum_\mathbf{x} \mathbf{x} p(\mathbf{x}|t)$ and covariance $\mathbf{C} = \sum_\mathbf{x} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T p(\mathbf{x}|t)$:

$$\begin{aligned}
\frac{\partial m_i}{\partial t} &= \textstyle\sum_r \nu_{ri} \left(a_r(\mathbf{m}) + \frac{1}{2} \displaystyle\sum_{k,l} \frac{\partial^2 a_r}{\partial x_k \partial x_l}\bigg|_\mathbf{m} C_{kl} \right), \\
\frac{\partial C_{ij}}{\partial t} &= \textstyle\sum_r \left[\nu_{ri}\nu_{rj} a_r(\mathbf{m}) + \displaystyle\sum_k \frac{\partial a_r}{\partial x_k}\bigg|_\mathbf{m} (\nu_{ri} C_{jk} + \nu_{rj} C_{ik}) \right. \qquad (1) \\
&\quad \left. + \frac{1}{2} \sum_{k,l} \frac{\partial^2 a_r}{\partial x_k \partial x_l}\bigg|_\mathbf{m} (\nu_{ri}\nu_{rj} C_{kl} + \nu_{ri} C_{jkl} + \nu_{rj} C_{ikl}) \right],
\end{aligned}$$

where C_{ikl} denotes the third-order moment of X_i, X_k and X_l. Due to the symmetry $C_{ij} = C_{ji}$ only C_{ij} with $i \leq j$ is considered. As in (1), the evolution equations for second-order moments usually depend on third-order moments. To close the 2MA equations, moment-closure techniques are applied which approximate the third-order moments as functions of first- and second-order moments (Hespanha, 2008). The moment closure introduces an approximation error to the otherwise exact moment equations, as it relies on assumptions about $p(\mathbf{x}|t)$ (e.g. normality or log-normality; Singh and Hespanha, 2006).

The 2MA (1) describes the covariances of all pairs of species and thus possesses $O(n^2)$ state variables. This quadratic scaling with respect to the number of species, n, poses a challenge for the applicability of 2MA to large biological networks that may contain several hundreds up to thousands of species. However, it is usually observed that in large biochemical networks, many pairwise correlations between species are small. This implies a comparably low covariance and a small contribution to the right-hand side of (1). Consequently, for an approximation of the dynamics of the biochemical network, it may not be necessary to model all covariances.

Studying a series of networks, including the JAK2/STAT5 signalling pathway described by Bachmann *et al.* (2011), we observed that species that directly influence each other via a reaction have a stronger pairwise correlation. For the JAK2/STAT5 signalling pathway, depicted in Figure 1A, we found that >50% of the correlation coefficients do not exceed an absolute value of 0.1 (Fig. 1B). Furthermore, the correlation coefficients decrease as the distance between species in the network increases (Fig. 1C). Since in many cases biological networks are sparsely connected and distances between species are relatively large (Fig. 1D), a significant portion of the covariances may be negligible.

Motivated by this observation, we develop a scalable s2MA that models a subset of covariances. The s2MA is designed to provide a good approximation for means and variances of species, as those moments are essential in a range of applications including parameter estimation (Munsky *et al.*, 2009; Fröhlich *et al.*, 2016). Accordingly, the s2MA captures the subset of covariances that are expected to influence the temporal evolution of the means and variances most strongly. In the simplest case, we only consider the covariances $\mathbf{C}^*$ that have a direct influence on the means and variances, i.e. those that appear in their evolution equations for m_i and C_{ii}:

- Covariances C_{ik} for which a reaction R_r exists with $\nu_{ri} \neq 0$ and $\frac{\partial a_r}{\partial x_k} \neq 0$. This is the case if S_k is a modifier or reactant in a reaction producing or consuming S_i.
- Covariances C_{kl} for which a reaction R_r exists with $\nu_{ri} \neq 0$ and $\frac{\partial^2 a_r}{\partial x_k \partial x_l} \neq 0$. This is the case if both, S_k and S_l, are modifiers or reactants in a reaction producing or consuming S_i.

The remaining covariances are set to zero. The resulting MA exploits the network structure and is similar to a recently proposed MA for spatially distributed systems exploiting the neighbourhood structure (Feng *et al.*, 2016). In the following, we present a mathematical formulation of the s2MA as well as extensions to control its size and approximation accuracy.

3 Materials and Methods

To simulate the statistical moments of the trajectories of large-scale stochastic biochemical reaction networks, we introduce scalable moment-closure approximations (sMAs). These sMA are based on the afore-mentioned findings and exploit the structure of the biochemical reaction network. In the following, we present the required graph characteristics and the derivation of the s2MA.

3.1 Graph representation of biochemical reaction networks

The s2MA uses the structure of the reaction network to identify the covariances that are most relevant to accurately approximate the means and variances of species. To establish a simple structure-based procedure, we exploit the graph structure of the biochemical reaction networks. This graph structure is best represented using the Systems Biology Graphical Notation (SBGN) process diagram

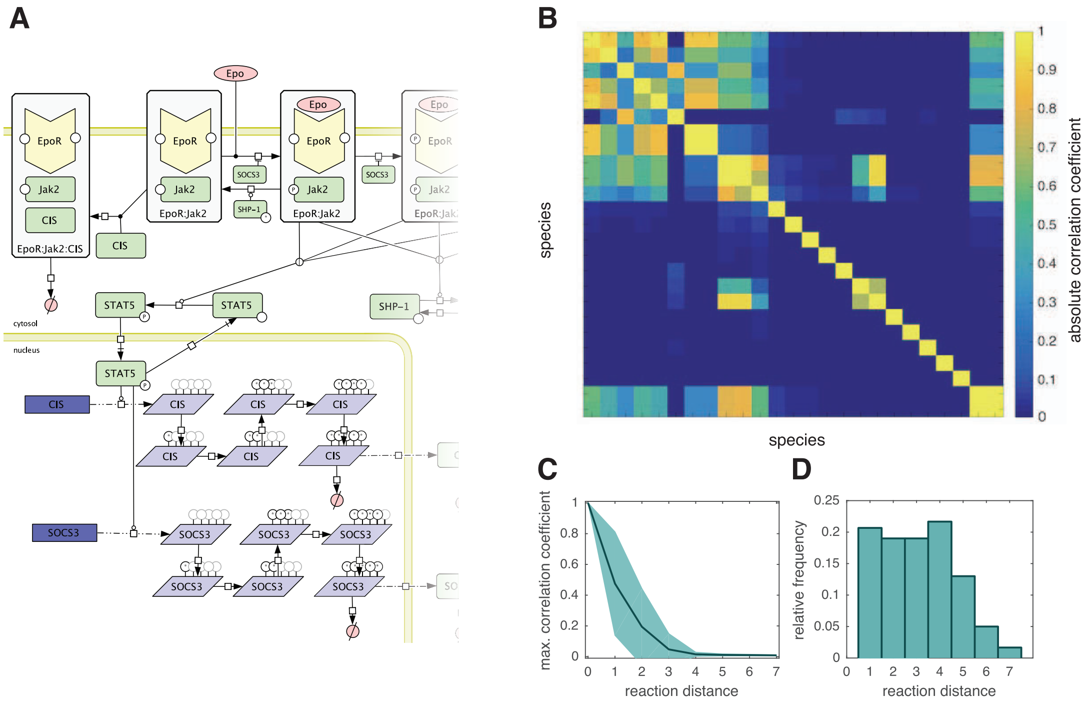

Fig. 1. Correlation coefficients in the simulated JAK2/STAT5 signalling pathway. (**A**) A partial schematic of the JAK2/STAT5 signalling pathway. (**B**) Maximum absolute pairwise correlation coefficients found in the simulation of the JAK2/STAT5 signalling pathway. (**C**) Maximum absolute pairwise correlation coefficients as function of the distance between species. (**D**) Frequency distribution of distance between species pairs

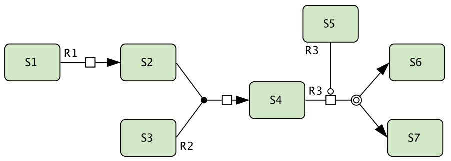

Fig. 2. Illustration of SBGN process diagram of a simple biochemical reaction network. Biochemical species (boxes), biochemical processes (squares) and interactions/dependencies (arcs) are visualised. Label S_i indicates species S_i

(Le Novère *et al.*, 2009). In essence, SBGN process diagram is a graph which consists of *entity nodes* representing biochemical species, *process nodes* representing biochemical reactions and arcs indicating the interactions/dependences. The incoming edges to a process node indicate all the reactants, as well as the modifiers, of the corresponding reaction, while the outgoing edges from a process node mark the products. For instance, reaction R_2 in Figure 2 is a bimolecular reaction where species S_2 and S_3 react to form species S_4. In reaction R_3, species S_5 acts as a modifier that activates the conversion of S_4 into S_6 and S_7. The graph structure is encoded in the propensities and the stoichiometric coefficients and can be easily visualized for Systems Biology Markup Language (SBML) models using software toolboxes such as CellDesigner (Funahashi *et al.*, 2008).

We use the graph representation to define a *dependency matrix* D which summarizes direct dependencies between species in the network. Following the arguments in Section 2, we say that a species S_j directly depends a species S_i, if the evolution equations for the mean or the variance of S_j, i.e., m_j and C_{jj}, depend on moments of S_i. Accordingly, it can be shown that:

- The products of a reaction depend on the reactants and the modifiers.
- The reactants of a reaction depend on the other reactants and the modifier.

This yields the dependency matrix D,

$$D_{ij} = \begin{cases} 1 & \text{if } S_i \text{ directly influences } S_j \\ 0 & \text{otherwise} \end{cases}$$

Note that D is not necessarily symmetric as the defined dependency is a directed property. In the model depicted in Figure 2, S_4 depends on S_2 ($D_{24} = 1$) but not vice versa ($D_{42} = 0$). The dependency matrix D encodes the necessary information for the construction of the s2MA.

3.2 The scalable s2MA

The exact evolution equations for means $\mathbf{m}$ and covariances $\mathbf{C}$ (1) can be written as

$$\begin{aligned} \frac{\partial m_i}{\partial t} &= F_{m,i}(\mathbf{m}, \mathbf{C}, \mathbf{H}), \quad i \in \{1, \ldots, n\} \\ \frac{\partial C_{ij}}{\partial t} &= F_{C,ij}(\mathbf{m}, \mathbf{C}, \mathbf{H}), \quad (i,j) \in I \\ &\text{with } I = \{(i,j) \in \{1, \ldots, n\}^2 | i \le j\}. \end{aligned} \tag{2}$$

where $\mathbf{H}$ denotes all moments with orders greater than two. To avoid redundancies caused by the symmetry of the covariances, $C_{ij} = C_{ji}$, we consider only the subset I of covariances. The higher-order moments $\mathbf{H}$ result from reactions with non-linear propensities and their temporal evolution is not to described by (2). To obtain a closed formulation, the higher-order moments $\mathbf{H}$ are approximated by functions of lower-order moments, $\mathbf{H} \approx \bar{H}(\mathbf{m}, \mathbf{C})$, using moment closure techniques. Common techniques include zero-cumulant closure (Matis and Kiffe, 1999), low-dispersion closure (Hespanha, 2008), and derivative-matching (Singh and Hespanha, 2007). This yields the 2MA,

$$\begin{aligned} \frac{\partial m_i}{\partial t} &= F_{m,i}\big(\mathbf{m}, \mathbf{C}, \bar{H}(\mathbf{m},\mathbf{C})\big) =: \bar{F}_{m,i}(\mathbf{m},\mathbf{C}), \quad i \in \{1,\dots,n\} \\ \frac{\partial C_{ij}}{\partial t} &= F_{C,ij}\big(\mathbf{m}, \mathbf{C}, \bar{H}(\mathbf{m},\mathbf{C})\big) =: \bar{F}_{C,ij}(\mathbf{m},\mathbf{C}), \quad (i,j) \in I. \end{aligned}$$

The solution of the 2MA yields an approximation to the moments of the state of the biochemical reaction network. The quality of this approximation depends on the accuracy of the moment closure (Kazeroonian *et al.*, 2016; Schnoerr *et al.*, 2015).

The 2MA possesses $n(n+3)/2$ state variables, thus, it grows quadratically with n. The simplest s2MA, the first-degree s2MA, reduces the growth rate by considering only the covariances on which the temporal evolution of the means and variances depends directly. This reduced set of covariances, $\mathbf{C}_{ij}$ with $(i,j) \in I^{(1)}$, can be determined using the dependency matrix D,

$$I^{(1)} = \left\{ (i,j) \in \{1,\dots,n\}^2 \middle| i \le j \wedge \left(D + D^T\right)_{ij} \neq 0 \right\}.$$

The covariances C_{ij} with $(i,j) \in I \backslash I^{(1)}$ are not modelled by the first-degree s2MA but can be approximated using the means, the variances and the reduced set of covariances. In this study, we use the low-dispersion closure, $C_{ij} = 0$ for $(i,j) \in I \backslash I^{(1)}$.

The approximation quality of the s2MA can be controlled using the cut-off degree. The second-degree s2MA describes the covariances that influence the temporal evolution of the means and variances either directly or via an intermediate step. More precisely, the second-degree s2MA considers the covariances $\mathbf{C}_{ij}$, $(i,j) \in I^{(1)}$ and the covariances which appear in their evolution equations. The set of these covariances, $\mathbf{C}_{ij}$, $(i,j) \in I^{(2)}$, is defined by the second power of the dependency matrix D^2. More generally, we define the δth-degree s2MA (s2MA-δ) which describes the reduced set of covariances $\mathbf{C}_{ij}$ with $(i,j) \in I^{(\delta)}$,

$$I^{(\delta)} = \left\{ (i,j) \in \{1,\dots,n\}^2 \middle| i \le j \wedge \left(D^\delta + \left(D^\delta\right)^T\right)_{ij} \neq 0 \right\}.$$

The degree $\delta \ge 1$ denotes the maximal intermediate dependency steps between species pairs (S_i, S_j) for which covariances are included in the s2MA. For a given δ, we obtain the s2MA-δ,

$$\begin{aligned} \frac{\partial m_i}{\partial t} &= \bar{F}_{m,i}(\mathbf{m},\mathbf{C}), \quad i \in \{1,\dots,n\} \\ \frac{\partial C_{ij}}{\partial t} &= \bar{F}_{C,ij}(\mathbf{m},\mathbf{C}), \quad (i,j) \in I^{(\delta)} \\ C_{ij}(t) &= 0, \quad (i,j) \in I \backslash I^{(\delta)}. \end{aligned} \qquad (3)$$

We focus on the case $\delta = 1$, in which merely covariances of interacting species are considered. To capture long-range interactions, we considered $\delta \ge 2$, which can improve the approximation accuracy of the s2MA in biological systems with complex or highly non-linear kinetics. The potentially enhanced approximation accuracy comes at the cost of higher computational complexity as the number of state variables increases with δ. In Section 4, we demonstrate that one can usually find a satisfactory tradeoff between the computational cost and approximation quality for complex biological networks.

3.3 Implementation

We implemented methods for the construction and simulation of the s2MA in the ChEmical REaction Network Analyzer (CERENA), an open source MATLAB toolbox (Kazeroonian *et al.*, 2016). The advanced version of CERENA supports automatic construction of the 2MA and the s2MA using symbolic calculus and allows for a range of moment closure schemes. The proposed construction algorithm circumvents the formulation of the full 2MA to ensure feasibility for large-scale networks. Biochemical reaction networks can be defined in the SBML or in a simple m-file format. For efficient numerical simulation, C-code simulation files are compiled using the Advanced MATLAB Interface for CVODES and IDAS (Fröhlich *et al.*, 2016). This C-code employs sophisticated numerical methods implemented in CVODES (Serban and Hindmarsh, 2005), facilitating the study of a wide range of models. In addition, simulation using MATLAB internal ODE solvers is supported. CERENA is freely available from GitHub (http://cerenadevelopers.github.io/CERENA/) and its functionality is described in a detailed documentation.

4 Results

In the following, we study the properties of the s2MA and illustrate its importance for the study of large-scale biochemical reaction networks. For this purpose, we analyse various network motifs as well as published pathway models for which available methods are computationally demanding or even infeasible.

4.1 Scaling properties

The size of the s2MA for a given network as well as its scaling properties depends on network characteristics. To highlight the scaling properties, we considered reoccurring network motifs and performed a general theoretical assessment. As verification, we inspected published signalling and metabolic pathways with different numbers of biochemical species.

4.1.1 Theoretical scaling for network motifs and generic networks

To study the scaling properties of s2MA, we considered three different network motifs illustrated in Figure 3A–C:

- A *chain of monomolecular reactions* as observed in metabolic processes (Krumsiek *et al.*, 2011) and delay representations (Bachmann *et al.*, 2011).
- A *2D grid of monomolecular reactions* as observed in histone methylation (Zheng *et al.*, 2012).
- A *sequence of bimolecular reactions with a hub* as observed in polymerisation related processes, e.g. prion aggregation (Rubenstein *et al.*, 2007).

For these network motifs, we derived the size of the s2MA-1 and -2 (see Table 1). For all three motifs, we found a linear scaling of the size of the s2MA-1 with respect to the number of species n. The same holds for the s2MA-2 of the chain of monomolecular reactions and the 2D grid of monomolecular reactions. The s2MA-2 of the sequence of bimolecular reactions with a hub is identical to the 2MA as all species are connected via at most one intermediate species (the hub). Accordingly, the analysis of selected motifs suggests that the s2MA allows for a substantial size reduction in the absence of central hubs.

For generic network structures, the scaling of the s2MA depends on the degree distribution $P(d)$ of nodes in the graph representation of the biochemical reaction network (see Section 3.1). By construction, the number of covariances in the s2MA-1 is the sum of node degrees over two,

$$\text{number of covariances in s2MA} - 1 = \frac{1}{2}\sum\nolimits_{i=1}^{n} d_i = \frac{n\bar{d}}{2},$$

in which d_i denotes the degree of node i and the division by two is required as covariances are associated to two nodes. Introducing the average node degree, $\bar{d} = \frac{1}{n}\sum\nolimits_{i=1}^{n} d_i$, the s2MA-1 describes the temporal evolution of n means, n variances and $\frac{n\bar{d}}{2}$ covariances, and thus possesses $\frac{n}{2}\left(4+\bar{d}\right)$ state variables. If we assume that there are no long-ranged connections in the network and every node is only connected to a subset of neighbouring nodes, then we can assume that $\bar{d}$ is independent of the size of the network n, and s2MA-1 will scale linearly with the number of species.

The degree distribution in biological systems have been reported to follow a power-law (Albert, 2005), $P(d) \propto d^{-\gamma}$, with an exponent of $2 < \gamma < 3$. Networks with this property are usually referred to as scale-free networks. The expected value of the average node degree in scale-free networks is

$$\mathbb{E}\left[\bar{d}\right] = \sum\nolimits_{i=1}^{n} d_i = \sum\nolimits_{d=1}^{n-1} d \cdot P(d) = \sum\nolimits_{d=1}^{n-1} d^{1-\gamma}.$$

Using the lower bound of γ and the upper bound on the partial sums of the harmonic series, we obtain

$$\text{if} \quad \gamma > 2 \quad \Rightarrow \quad \mathbb{E}\left[\bar{d}\right] < (\ln(n-1)+1).$$

Evaluating this upper bound, we notice that even for networks with up to $n = 10^4$ species, $\bar{d}$ hardly exceeds 10, making it behave like a constant compared to n. Accordingly, we conclude that the size of the s2MA-1 should scale (only slightly worse than) linearly with the network size.

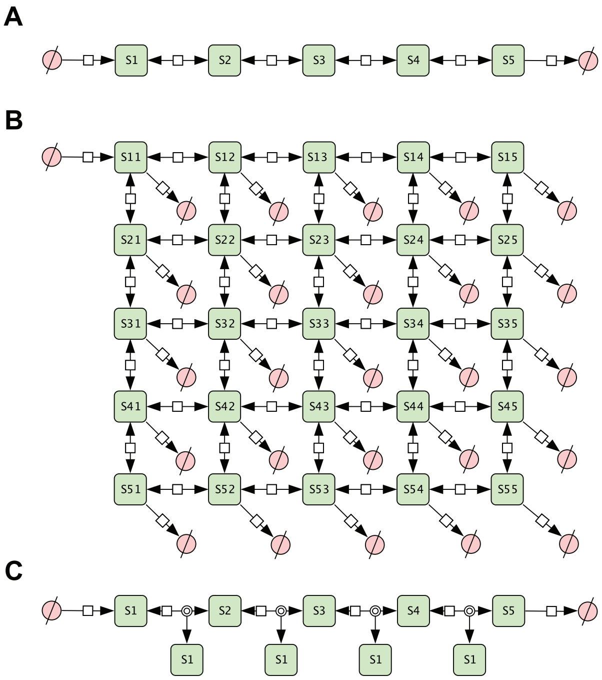

Fig. 3. Illustration of considered network motifs. (**A**) Chain of monomolecular reactions ($n = 5$). (**B**) 2D grid of monomolecular reactions ($n = 25$). (**C**) Chain of bimolecular reactions with a hub ($n = 5$)

Table 1. Comparison of the sizes of the 2MA and the s2MA for different network motifs

Network motif	Number of state variables		
	2MA	s2MA-1	s2MA-2
Chain of monomolecular reactions	$\frac{n(n+3)}{2}$	$3n-1$	$4n-3$
2D grid of monomolecular reactions	$\frac{n(n+3)}{2}$	$4n-\sqrt{n}$	$7n-7\sqrt{n}+1$
Chain of bimolecular reactions	$\frac{n(n+3)}{2}$	$4n-3$	$\frac{n(n+3)}{2}$

4.1.2 Scaling for published biochemical reaction networks

To corroborate the theoretical predictions derived under the assumption of scale-free networks, we studied a collection of 50 published biochemical reaction networks. These networks were extracted from the BioModels, NetPath and Reactome database. They include between 17 and 1277 biochemical species and a range of rate laws. A comprehensive list of the networks is provided in Supplementary Table S1.

We used an extension of the MATLAB toolbox CERENA to generate the s2MAs for the networks and recorded the sizes (Fig. 4). The analysis verified our prediction of a roughly linear relation between the size of the s2MA-1 and the number of species. The s2MA-1, on average, possessed only five times more state variables than the reaction rate equations, ensuring the applicability of the s2MA-1 to large-scale networks. For the largest network, a size reduction by a factor of >120 was achieved compared to the 2MA.

As the consideration of pair-wise correlations between reaction partners might not be sufficient for a particular application, we also assessed the scaling of the s2MA-2 and -3. In agreement with the results for the network motifs, we found that the size of the s2MA of degree ≥ 2 grew stronger than linear, namely with order 1.25 and 1.49. This implies that for realistic pathway structures, also the size of the s2MA of degree 2 and 3 grows substantially slower than the size of the 2MA, facilitating the analysis of stochasticity in large-scale networks.

4.2 Approximation accuracy

The improved scalability of the s2MA is achieved by merely modelling a subset of covariances. In the following section, we will assess the resulting approximation error and its dependence on the degree

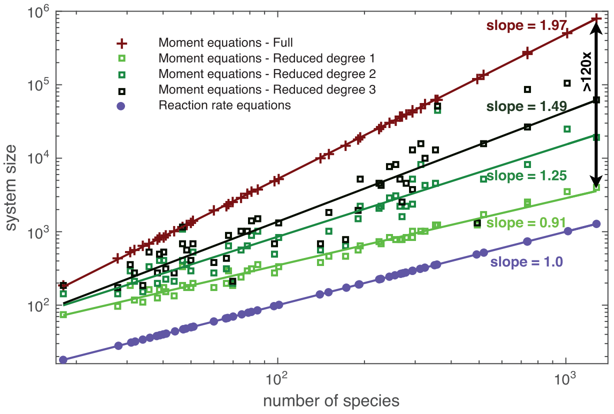

Fig. 4. Scaling of different moment-closure approximations for published networks. Moment-closure approximations for individual networks (markers) and fitted regression curves (lines) are shown

of the s2MA. For this analysis, we consider two network motifs and two published signalling pathways.

4.2.1 Comparison of approximation methods for network motifs

For an initial assessment of the approximation accuracy, we considered the *chain of monomolecular reactions* ($n = 10$) and the *sequence of bimolecular reactions with a hub* ($n = 20$) with mass action kinetics (Fig. 3A and C). The initial conditions and parameter values are reported in Supplementary Tables S2 and S3. As a measure for the approximation accuracy the relative errors in the means and variances were used, e.g.

$$100\% \times \frac{|C_{ii}^{\mathrm{s2MA}}(t) - C_{ii}^{\mathrm{2MA}}(t)|}{\max_t C_{ii}^{\mathrm{2MA}}(t)},$$

in which $C_{ii}^{\mathrm{s2MA}}(t)$ and $C_{ii}^{\mathrm{2MA}}(t)$ denote the time-dependent variance of species i calculated by s2MA and 2MA, respectively.

The numerical simulation revealed a good agreement of means and variances of 2MA and s2MA-1 (Fig. 5). Neglecting the covariances that are not modelled by the s2MA; however, resulted in a relative error $< 1\%$ for the means and $< 20\%$ for the variances. Given a size reductions of 55.4 and 66.5%, the low relative error supported the validity of the approach.

4.2.2 Comparison of approximation accuracy for s2MA of different degrees on published biochemical reaction networks

To assess the approximation accuracy of s2MAs of different degrees for realistic pathway topologies, we considered the published models of JAK2/STAT5 signalling and TNF signalling. These models were also considered in the scalability analysis (Section 4.1.2).

The model of JAK2/STAT5 signalling describes the activity of the transcription factor STAT5 in response to Epo treatment (Bachmann *et al.*, 2011). STAT5 regulates cell proliferation, differentiation and inflammation. The considered model accounts for 25 biochemical species and includes biochemical reactions with non-mass action kinetics. Its 2MA possesses 350 state variables while the s2MA-1 has less than one-third of the state variables, namely 112. Nonetheless, the simulation revealed a good agreement of 2MA and s2MA-1 for the means and variances (Fig. 6A). The means and variances computed using s2MA-2 and s2MA-3 were essentially indistinguishable from those computed using 2MA. For all s2MAs, we observed a reduction in the computation time comparable to the size reduction.

The model of TNF signalling describes the activation of pro- and antiapoptotic factors, i.e. caspases and NFκB, in response to TNF treatment (Schliemann *et al.*, 2011). Apoptosis is a form of programmed cell death which is relevant, among others, in immune

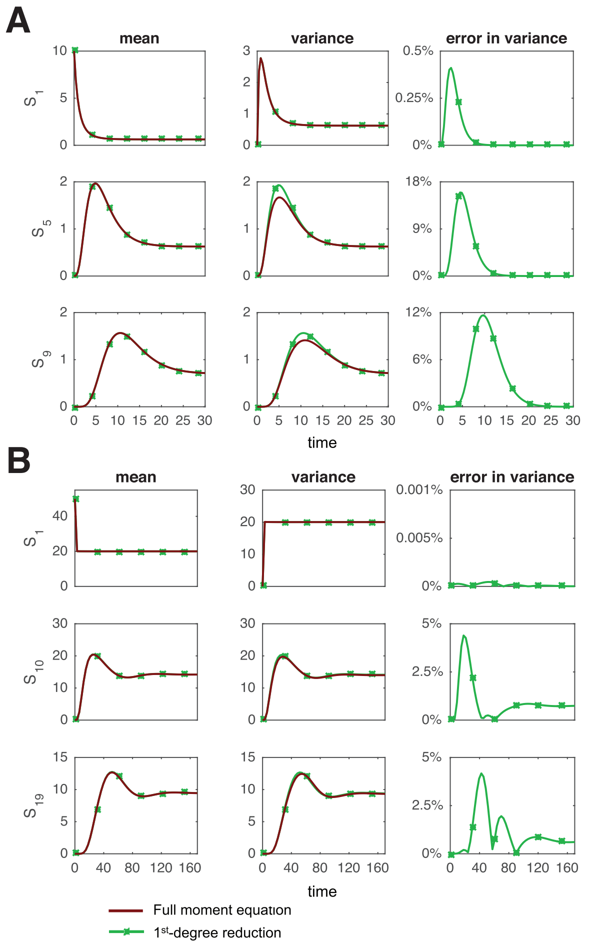

Fig. 5. Approximation accuracy of the s2MA-1 for network motifs. (**A**) The chain of monomolecular reactions with $n = 10$. (**B**) The sequence of bimolecular reactions with $n = 20$. (A, B) Means and variances are depicted along with relative errors in the variances (2MA versus s2MA-1) for several biochemical species

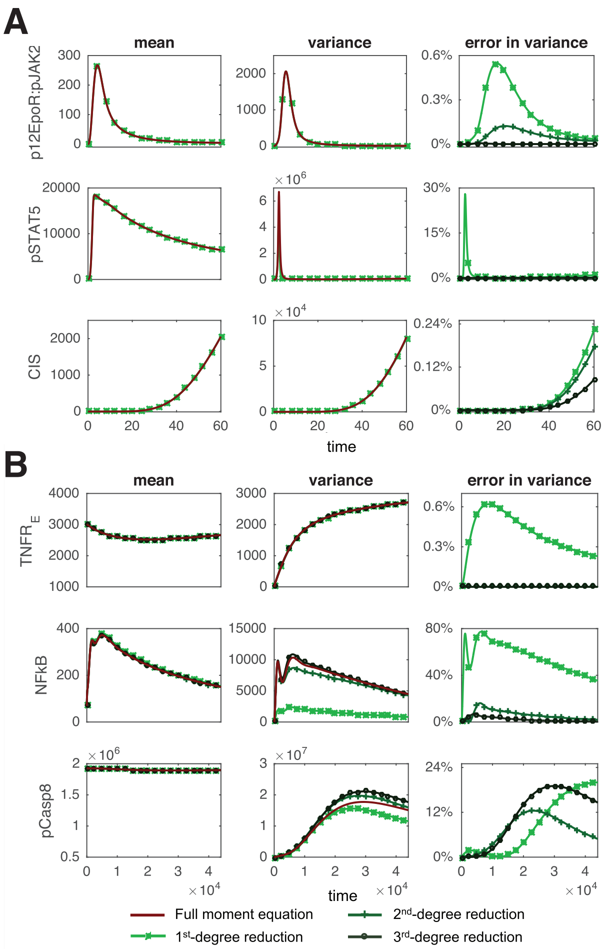

Fig. 6. Approximation accuracy of the s2MA for published pathways. (**A**) The JAK2/STAT5 signalling pathway. (**B**) The TNF signalling network. (A, B) Means and variances computed using the 2MA and the s2MA-1, -2 and -3 are depicted for several biochemical species. For the s2MA of different degrees, the relative error in the variances with respect to the 2MA is provided

response and cancer. The model comprises 47 biochemical species, yielding a 2MA with 1175 state variables. In contrast, the s2MA-1, -2 and -3 possess only 189, 540 and 664 state variables. The numerical simulation of the s2MA-1 was more than 25 times faster than the numerical simulation of the 2MA. The disagreement between s2MA-1 and 2MA, which resulted in a relative error of 100% for some species (Fig. 6B) indicates that also covariances between species which do not interact directly might be required for an accurate description of mean and variances. The comparison of the results for s2MA-1, -2 and -3 confirmed that the approximation error decreases as more covariances are taken into account. For s2MA-3, the relative error is below 15%.

In summary, our analysis of network motifs and published networks revealed that the s2MA yields substantially smaller ODE models than the 2MA, indicating a substantial gain in computational efficiency. Moreover, even for models with many species and non-mass action kinetics, a good approximation accuracy was achieved.

5 Discussion

Stochasticity of biochemical reactions is an inherent property of biological processes. It contributes to the establishment of functional cell-to-cell variability and robust decision-making (Eldar and Elowitz, 2010; Raj and van Oudenaarden, 2008). The analysis of the stochastic processes is, however, restricted by the available analytical and numerical methods. In this manuscript, we introduce the scalable second-order moment-closure approximation, the first method to enable the simulation of statistical moments of large-scale stochastic processes. The s2MA exploits the network structure to construct approximate evolution equations for selected process statistics.

To assess and illustrate the properties of s2MA, we studied network motifs and a large collection of published networks. This comprehensive evaluation, which sets this study apart from other studies of moment-closure approximations (e.g. (Feng *et al.*, 2016; Singh and Hespanha, 2006), verified that in practice the size of the first-degree s2MA (s2MA-1) grows linearly with the network size, a scalability that is similar to the reaction rate equations. Accordingly, the s2MA enables the assessment of stochastic dynamics on a new scale. The achieved scalability, however, comes at the cost of an approximation error. The approximation quality can be easily controlled via the degree of the s2MA.

Beyond scalable moment-closure approximations for the calculation of means and variances, structured-based approaches might be used for the evaluation of third-order moments and conditional moments (Hasenauer *et al.*, 2014). Complementarily, an improvement might be achieved by tailored moment-closure schemes which avoid neglecting a large fraction of covariances. A possible formulation, for instance, could be based on partial correlations (Krumsiek *et al.*, 2011) or convergent moments (Zhang *et al.*, 2016). All of these methods would benefit from *a priori* and *a posteriori* error bounds, which are not yet available for moment-closure approximations, such as the s2MA, but are urgently needed.

In summary, we presented a scalable moment-closure approximation for the simulation of stochastic chemical kinetics. This method is beneficial for application problems that require numerical simulations at low computation cost, e.g. parameter estimation (Fröhlich *et al.*, 2016; Munsky *et al.*, 2009). An implementation of the method is provided in the open-source MATLAB toolbox CERENA to facilitate its application and further extensions. This implementation, as well as the concept of structure-based reduction, is applicable to a broad range of problems and will help to improve the analysis of stochastic chemical kinetics.

Acknowledgements

We acknowledge financial support from the Postdoctoral Fellowship Program (PFP) of the Helmholtz Zentrum München.

Conflict of Interest: none declared.

References

Albert,R. (2005) Scale-free networks in cell biology. *J. Cell. Sci.*, **118**, 4947–4957.

Bachmann,J. *et al.* (2011) Division of labor by dual feedback regulators controls JAK2/STAT5 signaling over broad ligand range. *Mol. Syst. Biol.*, **7**, 516.

Eldar,A. and Elowitz,M.B. (2010) Functional roles for noise in genetic circuits. *Nature*, **467**, 1–7.

Elowitz,M.B. *et al.* (2002) Stochastic gene expression in a single cell. *Science*, **297**, 1183–1186.

Engblom,S. (2006) Computing the moments of high dimensional solutions of the master equation. *Appl. Math. Comp.*, **180**, 498–515.

Feng,C. *et al.* (2016) Automatic moment-closure approximation of spatially distributed collective adaptive systems. In *ACM Transactions on Modeling and Computer Simulation*, Vol. **26**.

Fröhlich,F. *et al.* (2016) Inference for stochastic chemical kinetics using moment equations and system size expansion. *PLoS Comput. Biol.*, **12**, e1005030.

Funahashi,A. *et al.* (2008) CellDesigner 3.5: A versatile modeling tool for biochemical networks. *Proc. IEEE*, **96**, 1254–1265.

Grima,R. (2010) An effective rate equation approach to reaction kinetics in small volumes: Theory and application to biochemical reactions in nonequilibrium steady-state conditions. *J. Chem. Phys.*, **133**, (035101.

Hasenauer,J. *et al.* (2014) Method of conditional moments (MCM) for the chemical master equation. *J. Math. Biol.*, **69**, 687–735.

Hespanha,J. (2008). Moment closure for biochemical networks. In *Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing*, St Julians, 2008, pp. 142–147.

Kazeroonian,A. *et al.* (2016) CERENA: ChEmical REaction Network Analyzer—a toolbox for the simulation and analysis of stochastic chemical kinetics. *PLoS One*, **11**, e0146732.

Krumsiek,J. *et al.* (2011) Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.*, **5**,

Le Novère,N. *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.

Lee,C.H. *et al.* (2009) A moment closure method for stochastic reaction networks. *J. Chem. Phys.*, **130**, 134107.

Matis,H.J. and Kiffe,T.R. (1999) Effects of immigration on some stochastic logistic models: a cumulant truncation analysis. *Theor. Popul. Biol.*, **56**, 139–161.

Munsky,B. *et al.* (2009) Listening to the noise: random fluctuations reveal gene network parameters. *Mol. Syst. Biol.*, **5**, (318.

Raj,A. and van Oudenaarden,A. (2008) Nature, nurture, or chance: Stochastic gene expression and its consequences. *Cell*, **135**, 216–226.

Rubenstein,R. *et al.* (2007) Dynamics of the nucleated polymerization model of prion replication. *Biophys. Chem.*, **125**, 360–367.

Schliemann,M. *et al.* (2011) Heterogeneity reduces sensitivity of cell death for TNF-stimuli. *BMC Syst. Biol*, **5**, (204.

Schnoerr,D. *et al.* (2015) Comparison of different moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.*, **143**, (185101.

Serban,R. and Hindmarsh,A.C. (2005) CVODES: An ODE solver with sensitivity analysis capabilities. *ACM T. Math. Softw.*, **31**, 363–396.

Singh,A. and Hespanha,J. (2007) A derivative matching approach to moment closure for the stochastic logistic model. *Bull. Math. Biol.*, **69**, 1909–1925.

Singh,A. and Hespanha,J.P. (2006). Lognormal moment closures for biochemical reactions. In *Proceedings of the 45th IEEE Conference on Decision and Control (CDC)*, San Diego, CA, 2006, pp. 2063–2068.

van Kampen,N.G. (2007). *Stochastic Processes in Physics and Chemistry, 3rd edn*. North-Holland, Amsterdam.

Zhang,J. *et al*. (2016) A moment-convergence method for stochastic analysis of biochemical reaction networks. *J. Chem. Phys*, **144**, 194109.

Zheng,Y. *et al*. (2012) Total kinetic analysis reveals how combinatorial methylation patterns are established on lysines 27 and 36 of histone H3. *Proc. Natl. Acad. Sci. USA*, **109**, 13549–13554.