Research

Authors for correspondence:
Eszter Lakatos
e-mail: e.lakatos13@imperial.ac.uk
Michael P. H. Stumpf
e-mail: m.stumpf@imperial.ac.uk

THE ROYAL SOCIETY
PUBLISHING

Control mechanisms for stochastic biochemical systems via computation of reachable sets

Eszter Lakatos and Michael P. H. Stumpf

Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Biosciences, Imperial College London, London SW7 2AZ, UK

MPHS, 0000-0002-3577-1222

Controlling the behaviour of cells by rationally guiding molecular processes is an overarching aim of much of synthetic biology. Molecular processes, however, are notoriously noisy and frequently nonlinear. We present an approach to studying the impact of control measures on motifs of molecular interactions that addresses the problems faced in many biological systems: stochasticity, parameter uncertainty and nonlinearity. We show that our reachability analysis formalism can describe the potential behaviour of biological (naturally evolved as well as engineered) systems, and provides a set of bounds on their dynamics at the level of population statistics: for example, we can obtain the possible ranges of means and variances of mRNA and protein expression levels, even in the presence of uncertainty about model parameters.

1. Introduction

Much of the research in synthetic and systems biology in the last decade has focused on the study of elementary biological systems. This has often been with the aim of controlling and modifying them in order to achieve new functional modules exhibiting novel and useful behaviour [1], such as sustained oscillations [2] and bistability [3]. It is now becoming possible to use engineered biological systems made of characterized components to solve specific problems, such as information processing, energy production or production of chemicals.

However, biological systems are inherently noisy and probabilistic in nature, which can pose significant difficulty for one aiming for a reliable, well-characterized module. Although several external control techniques have been developed which

are able to avoid some of the variability in a population [4,5], noise at the level of molecular processes is often unavoidable and does, for example, affect quite profoundly how information is transmitted along the molecular networks underlying cell function [6–8]. Furthermore, due to the unreliability of measured quantities, our understanding of the underlying mechanisms might be mistaken leading to sub-optimal analysis and design. Therefore, we have to assess the practical limits on the amount of noise in a general biological module, in order to evaluate the efficiency of a control design or the reliability of a mechanistic model.

Reachability analysis has been widely used in control design and engineering for applications such as verification of electrical or mechanical networks and hybrid automata [9–13]. The analysis focuses on the computation of the subset of the state-space that can be reached within a certain time-limit, given some starting position of the system and external inputs. The technique can also be used to verify that a certain undesired state is not reached under realistic operating condition, or that the behaviour of the system is robust and qualitatively/quantitatively holds for different conditions, including different realistic inputs to the system.

To this end, reachability is generally calculated under varying levels of uncertainty regarding details of the system—such as initial state, input signal and rate parameters—usually formalized by assuming that these parameters come from a set of plausible values. Therefore, unless analytical solutions are derived for some abstraction of the system, the applicability of the computation heavily relies on the choice of the set representation [11,14,15]. Some representations might prove computationally expensive and hence impractical for high-dimensional systems, while a simple shape representation can lead to crude over-approximation of the reachable set. Methods have been proposed using several techniques such as polygonal projections [16], oriented hyper-rectangles, special polyhedra [17], ellipsoids [15] or level sets [11]. In this work, we use zonotopes [9], a centrally symmetric type of polytopes that can be conveniently represented by a list of vectors.

Although there is already a substantial body of work on reachable set computation for problems in engineering, including highly nonlinear cases [18], hybrid automata [19] and differential-algebraic equations [20], the complexity, frequent nonlinear behaviour and strict constraints on many of the model parameters of biochemical systems require the development of specialized analysis methods. There are already a few applications of reachability techniques to biological examples with special emphasis on the treatment of the nonlinearity of the system, either through direct computation [21] or through hybridization-based methods using either static or dynamic partitioning of the state-space [22–24]. The work by Dang *et al.* [24] has also been expanded to take into account the possible lack of knowledge of parameter values [25]. The stochasticity in biological systems has been tackled in even more diverse ways: through computing bounds on the probability function [26], using stochastic hybrid systems [27], or analytically deriving invariant sets for linear equations obtained from the stochastic model [28].

In this work, we propose a computationally efficient and flexible method to compute the reachable set (in a zonotopic representation) of stochastic biochemical systems; besides stochasticity we also consider possibly nonlinear rate laws; controlled or uncertain input signals; and uncertainty about model parameters, including those which we may need to control. The main steps behind our derivation are the following: (i) we first obtain an ordinary differential equation (ODE) representation of the system's mean and (co-)variances; (ii) then use an iterative procedure to obtain a conservative approximation of consecutive reachable sets up to a final time of interest. Here, we primarily use the linear noise approximation (LNA) for the first step, but also consider the moment expansion approximation [29–31] to demonstrate that other methods with different applicability can be used equally well in order to generate equations for the second step. We derive a new method to tightly approximate realistic biological input signals, and piece-wise temporal linearization is employed to deal with common nonlinearities. We also give conservative approximation formulae for the reachable set when rate parameters of the system are not precisely known—which is very often the case in biochemical systems [32].

The method is first demonstrated on two elementary modules fundamental to mathematical models and regulatory designs of biochemical networks. The first system presents the use of reachability analysis for the study of noisy biochemical reactions and evaluating a control on the levels of cell heterogeneity; the second example considers the task of model (in)validation for cases when high cell-to-cell variability poses a challenge to estimating the system's true behaviour. Finally, we explore the limitations of our method and how it can be used as a quick indicator of 'challenging' dynamics.

2. Material and methods

2.1. Zonotopes

A zonotope [33] is described by the position of its centre (c) and a set of generator vectors ($g_1, \ldots, g_p$) as

$$\mathcal{Z} := \left\{ x \in \mathbb{R}^n \mid x = c + \sum_{i=1}^{p} a_i g_i, \; -1 \le a_i \le 1 \right\}. \tag{2.1}$$

Two example zonotopes are shown in the electronic supplementary material, figure S1(a). In the following, we use the shortened notation $(c; G)$ to represent $\mathcal{Z}$, where $G \in \mathbb{R}^{n \times p}$ is a generator matrix formed from the generator column vectors. Zonotopes are a convenient representation as they are closed under Minkowski-addition and affine transformations, the two key operations in reachability analysis [9]. Furthermore, the above can be calculated through simple matrix-vector operations: given zonotopes $\mathcal{Z} = (c; G)$ and $\mathcal{W} = (d; H)$, $(H \in \mathbb{R}^{n \times q})$ and the affine transformation $\mathbb{T}(x) = Ax + b$,

$$\mathcal{Z} \oplus \mathcal{W} = \{x + y \mid x \in \mathcal{Z}, \; y \in \mathcal{W}\} = (c + d; [G, H]) \tag{2.2a}$$

and

$$\mathbb{T}(\mathcal{Z}) = \{Ax + b \mid x \in \mathcal{Z}\} = (Ac + b; AG), \tag{2.2b}$$

where $[G, H]$ is the concatenation of the 'vector lists'.

2.2. Linear noise approximation

Given a stochastic system with N-dimensional state variable $\chi(t)$ describing the abundance of modelled species at time t. We assume that the system is defined by (i) a stoichiometry matrix, S, in which each element (S_{ij}) describes the amount of i molecules gained through reaction j; and (ii) a collection of reaction propensities, $F = [a_1, a_2, \ldots, a_r]^T$ that quantifies the probability of each reaction happening, similarly to reaction speeds in classical deterministic systems. A detailed illustration of the stochastic formalism is given in the first example in §4.

In the LNA [34], we divide the state variable into a macroscopic part, and random fluctuations, as $\chi(t) = \phi(t) + \xi(t)$. This way the original system, which is typically described by the time-evolution of the entire probability distribution [35], is converted into an ordinary and a stochastic differential equation (in the Ito form),

$$\left. \begin{aligned} \frac{d\phi(t)}{dt} &= S \cdot F(\phi(t)) \\[2mm] \frac{d\xi(t)}{dt} &= S \cdot D(\phi(t))\xi(t) + \sum_{j=1}^{r} S_{\cdot,j} W_j(a_j(\phi(t))), \end{aligned} \right\} \tag{2.3}$$

and

where $W_j(a_j(\phi(t)))$ are independent Wiener processes ($\forall j$) with normally distributed increments, $\mathcal{N}(0, a_j(\phi(t)))$; and $\{D(\phi)\}_{ik} = \partial a_j(\phi)/\partial \phi_k$, hence $D = \partial F/\partial \phi$ is the Jacobian of the system.

In the applications we aim to investigate, instead of trajectories of individual realization (i.e. single cells), we are interested in the population-level behaviour. Therefore, equation (2.3) is used to obtain equations that describe how the mean and variance of the probability distribution of values of $\chi(t)$ changes. The decomposition of the state variable makes it straightforward to follow the change in the mean of the system, as the macroscopic part corresponds to the mean behaviour and hence the evolution it is already given by equation (2.3). The evolution of the covariance matrix ($\Sigma \in \mathbb{R}^{N \times N}$) is calculated as

$$\frac{d\Sigma}{dt} = SD\Sigma + \Sigma(SD)^T + S \cdot \text{diag}\{F(\phi)\} \cdot S^T, \tag{2.4}$$

where $\text{diag}\{F\}$ represents the $r \times r$ matrix, whose diagonal elements are those of F (i.e. $\text{diag}\{F\}_{ii} = F_i$) and all other elements are zero.

Thus a set of ordinary differential equations can be derived that follows the time-dependent change of the mean values and (co)variances of all species, summarized in the $x(t) \in \mathbb{R}^n$ vector with $n = N(N+3)/2$

elements. We obtain $x(t)$ by concatenating $\phi(t)$ and the vectorized form of the upper triangle of $\Sigma(t)$ denoted as $(\Sigma(t))^v$,

$$\frac{dx}{dt} = \frac{d}{dt} \begin{bmatrix} \phi(t) \\ (\Sigma(t))^v \end{bmatrix} = \begin{bmatrix} S \cdot F(\phi(t)) \\ (SD\Sigma(t))^v + (\Sigma(t)(SD)^T)^v + (S\,\mathrm{diag}\{F(\phi(t))\}S^T)^v \end{bmatrix} = f(x(t)),$$

where the upper triangle vectorization operator stands for row-continuous listing of all elements in and above the diagonal: for example $(\Sigma(t))^v = [\Sigma_{11}, \Sigma_{12}, \ldots, \Sigma_{1N}, \Sigma_{22}, \ldots, \Sigma_{2N}, \Sigma_{33}, \ldots, \Sigma_{NN}]$. We give an $N = 2$ dimensional example of this computation in the first example of §4.

2.3. Moment expansion approximation

The LNA is based on the assumption that the system noise is well described by a normal distribution, and molecules are present in at least moderately high amounts. In many cases, it offers a good approximation even when these conditions are not met. However, to handle more problematic cases, we also use moment expansion approximation [29,30], a less constrained moment generating method, to obtain ordinary differential equations.

In brief, given the aforementioned stochastic system formed by $\chi(t)$, S and F, one can use the moment generating function of the species' probability distribution to obtain expressions of the time-evolution of kth order moments. This can be derived and evaluated to any arbitrary value of k. However, for nonlinear systems every moment equation would depend on moments of subsequent orders, leading to an infinite set of equations. Therefore, we also need to apply moment closure formulae (mc^k) to substitute the highest order terms ($\mathbb{E}(\chi^k)$) with expressions of means ($\phi(t)$) and (co)variances ($\Sigma(t)$): $\mathbb{E}(\chi^k) \longleftarrow \mathrm{mc}^k(\phi(t), \Sigma(t))$.

$$\mathbb{E}(\chi^k) \longleftarrow \mathrm{mc}^k(\phi(t), \Sigma(t)).$$

A major difference between applying the LNA or moment expansion techniques is that the latter can be applied to any arbitrary order or moments, such as skewness, kurtosis, etc. However, for reachability analysis our primary interest is in the mean and variance of molecules of the system, and moment expansion can be used in three ways to derive these. We can (i) set the expansion to be only up to second-order moments; (ii) use the step in the above equation to substitute all higher-order moments, not just the highest order ones; or (iii) use the whole system of k moments as our state vector, to make sure no essential influence on means and variances is omitted.

There has been considerable discussion of moment equations and their applicability to the analysis of stochastic systems, see e.g. [36–38]. We, therefore, refer to [29,30] for the applicability and limitations of this particular method and mathematical details of the derivation. Furthermore, we suggest the use of the Python-based implementation, MEANS [31], for automatic computation of the step converting the stochastic system into a deterministic ODE. The package provides a user-friendly framework that only requires the stochastic description to output the desired equation set, and offers both the linear noise and the moment expansion approximation for equation generation.

2.4. State-space representation

Next we consider the representation of the deterministic system obtained via either of the aforementioned approximation techniques. The general system is described by the evolution of the n-dimensional state variable, $x(t)$ as in [39]

$$\dot{x}(t) = f(x(t), u(t)); \quad x(0) = x_0, \tag{2.5}$$

where $u(t) \in \mathbb{R}^m$ is an input signal, and $f(x, u)$ is a generally nonlinear but time-independent transition function. We focus on the case where the input-dependence in the above equation can be separated as

$$f(x, u) = g(x) + Bu,$$

where $B \in \mathbb{R}^{n \times m}$ is the input matrix, which specifies which states are affected by the inputs. In terms of the original stochastic system, u can be understood, for example, as a zeroth-order (constitutive production) reaction, as we will discuss through the first example in §4.

We further assume that instead of x_0, a set containing all possible initial values in the state-space, $\mathcal{I}$, is given—for example, because we observe a range of expression values in a population of cells. Similarly, the input signal may also come from a set, $\mathcal{U}$, which is bound by some value $\mu \in \mathbb{R}$ so that $\|u(t)\| \le \mu \ \forall t$.

Then equation (2.5) is presented as a differential inclusion [40] of the form

$$\dot{x}(t) \in g(x(t)) \oplus B\mathcal{U}; \quad x(0) \in \mathcal{I}. \tag{2.6}$$

The reachable set of this system at any time t is defined as

$$\mathcal{R}_t := \{y \in \mathbb{R}^n \mid \exists(x, u); \; x(0) \in \mathcal{I}, \; u(s) \in \mathcal{U}, \; x(s) \text{ follows } (2.6) \; \forall s \le t, \; y = x(t)\}, \tag{2.7}$$

i.e. all the states we can achieve by having the above described system start from a possible initial state and affected by plausible input signals. Similarly, we can define $\mathcal{R}_{[0,t]}$ as all the states reachable within the time-interval $[0, t]$, by computing all individual reach sets: $\mathcal{R}_{[0,t]} = \bigcup_{s \in [0,t]} \mathcal{R}_s$.

3. Reachable set computation

We start our derivation by considering the case when f is a linear function of x; under such condition the system can be represented in the linear time-invariant (LTI) form,

$$\dot{x}(t) = Ax(t) + Bu(t); \quad x(0) = x_0, \tag{3.1}$$

where $A \in \mathbb{R}^{n \times n}$ is the state (transition) matrix. We focus on single-input systems, where B is an $n \times 1$ matrix (i.e. a column vector); however, all results can be easily generalized to other values of input dimension, m. The solution of this system is generally given as

$$x(t) = e^{At}x_0 + \int_0^t e^{As} Bu(t - s) \, \mathrm{d}s, \tag{3.2}$$

which can be further simplified through assuming a constant input signal and an invertible transition matrix

$$\int_0^t e^{As} \, \mathrm{d}s \cdot Bu = A^{-1}(e^{At} - I)Bu, \tag{3.3}$$

where I is the $n \times n$ identity matrix. In the following for convenience, we use the notation

$$\psi(A, t) = A^{-1}(e^{At} - I).$$

We consider an equidistant partitioning of the time horizon into N intervals, with time-step $\tau = T/N$. We start the derivation of reachable sets from $\mathcal{I}$ to $\mathcal{R}_T$ by handling the case of the autonomous system, i.e. when $u \equiv 0$.

Given a reachable set at an arbitrary time-point, $\mathcal{R}_t = (c; G)$, the new set reached under zero input can be calculated according to equation (3.2), as

$$\mathcal{R}_{t+\tau} = e^{A\tau} \mathcal{R}_t, \tag{3.4}$$

since the second term becomes zero for autonomous systems. This operation is in fact an affine transformation $x \rightarrow Mx$ (with $M = e^{A\tau}$) and hence the resulting set will also be a zonotope, as given by equation (2.2b),

$$\mathcal{R}_{t+\tau} = (e^{A\tau}c; \; e^{A\tau}G).$$

As the same equation applies for all values of t, the transition matrix can be applied iteratively to propagate $\mathcal{I}$ up to the final reach set, $\mathcal{R}_T$. If the aim is to map the entire space the system explores between times 0 and T, the reachable sets of the time-intervals, $\mathcal{R}_{[i\tau, (i+1)\tau]}$ has to be derived from the two end-points. Generally, this can be done by computing the convex hull of $\mathcal{R}_{i\tau}$ and $\mathcal{R}_{(i+1)\tau}$ and then extending this set to contain all affine solutions for one time-step (see [14]; electronic supplementary material, figure S1(b) for an illustration). However, zonotopes are not closed under this operation and it is desirable to avoid the exact computation of a convex hull. We use the conservative (over-)approximation presented in [9] to derive a set enclosing all points reachable between 0 and τ. All following intervals can be propagated from this set, as for all $t \in [0, \tau]$, $\mathcal{R}_t \in \mathcal{R}_{[0,\tau]}$ and hence $e^{A\tau}\mathcal{R}_t \in e^{A\tau}\mathcal{R}_{[0,\tau]}$.

The above computation is easily extended to account for a constant input signal, u. According to equations (3.3) and (3.4), $\mathcal{R}_{t+\tau}$ is given by

$$\mathcal{R}_{t+\tau} = e^{A\tau}\mathcal{R}_t + A^{-1}(e^{A\tau} - I)Bu,$$

which is conveniently also a zonotope, as $A^{-1}(e^{A\tau} - I)Bu = \psi(A, \tau)$ is a constant vector. However, in systems with uncertainty (so, for example, where the input signal is unknown) we need to transform and enlarge, or *bloat*, the set $\mathcal{R}_{i\tau}$ in each iteration, to account for the system dynamics under all admissible input/parameter values. This corresponds to taking all combinations of an unknown initial

value (contained in $\mathcal{I}$) and unknown parameter (input signal or reaction rate value) and obtaining every state the system can thus possibly reach. This is generally done by taking the Minkowski sum of the reachable set propagated from $\mathcal{I}$, $\mathcal{R}_{t+\tau}$; and the 'bloating' sets, β_μ and β_δ, accounting for the uncertainty introduced in one time-step by the input or rate parameter set, respectively. In the following sections, we derive these bloating sets in a zonotope formalism for convex, one-dimensional sets of input and parameter values. In addition, see electronic supplementary material, figure S2 for demonstration of the computational steps.

To summarize, using the short form $\mathcal{R}_i$ to denote either $\mathcal{R}_{i\tau}$ or $\mathcal{R}_{[(i-1)\tau,i\tau]}$ given in a zonotope form, the next reachable set is calculated as

$$\mathcal{R}_{i+1} = (e^{A\tau} c_i; e^{A\tau} G_i) \oplus \beta_\mu \oplus \beta_\delta. \tag{3.5}$$

As zonotopes can be stored as a set of column vectors summarized in a matrix, all the above computational steps are typically executed very efficiently, especially in languages dedicated for matrix-vector operations. Additionally, functions of the transition matrix and the constant time-step—such as $e^{A\tau}$—have to be calculated only once through the initialization of the algorithm, providing a further gain in computational speed.

3.1. Input for biological systems

In the algorithm proposed by Girard [9] the input set, $\mathcal{U}$ is taken to be an n-dimensional hypercube enclosing all points between $[-\mu, \mu]$ in all dimensions: equivalent to the radius μ ball in the infinity norm. Such a generalized approximation of the input set is sometimes necessary, as the analysis is motivated by determining either a target or an avoidable set in the state-space, and the input set is meant to capture all variations induced by noise in a physical system.

In the case of biological inputs, however, we usually have more detailed information about $\mathcal{U}$, meaning that the aforementioned general input set would contain many implausible signals, and hence the over-approximation would be too loose to provide useful information on the actual reachable sets. For example, in the typical case the control input is implemented as a certain type of molecular species added to the system; this cannot take negative values and it is also reasonable to assume an upper limit on the number of molecules injected at any time-point, which can be viewed as the bound μ. Furthermore, the input matrix, B, determines how much each variable is affected by the input signal and hence cannot be neglected.

As $\mathcal{U}$ is typically not centred around 0, we divide the input effect into a drift term, corresponding to the effect of the centre of $\mathcal{U}$ (u_c) and an uncertainty term representing our lack of knowledge about the exact input value, i.e. the maximal difference, u_d, between the centre and possible values of $\mathcal{U}$. Thus, the whole input set is taken into account as $\mathcal{U} = [u_c - u_d; u_c + u_d]$; for instance, in the example above, $u_c = \mu/2$ and $u_d = \mu/2$. The drift and uncertainty terms are calculated using equation (3.3), and the bloating set will be the zonotope

$$\beta_\mu = (\psi(A, \tau)Bu_c; \psi(A, \tau)Bu_d). \tag{3.6}$$

This calculation in this exact form is only possible if A is invertible—in the singular case we can use an approximation of equation (3.3), based on the integral of the Taylor-series of e^{As} and compute a bloating factor β_ν to correct for the small error thus introduced [20]:

$$\psi(A, t) = \sum_{i=0}^{\nu} \frac{A^i t^{i+1}}{(i+1)!} \oplus \underbrace{\left(0; e^{|A|t} - \sum_{i=0}^{\nu} \frac{|A|^i t^i}{i!}\right)}_{\beta_\nu}$$

3.2. Parameter uncertainty

Often we have to make predictions but only have approximate values of the parameters contributing to matrix A, either due to imperfect knowledge of reaction rates [41], or because we can only control a reaction through some interaction that is not or cannot be modelled explicitly. Therefore, we derive a way to account for uncertainty in case the matrix A is also drawn from a set or ensemble of matrices. For example, we can consider the case where a single reaction rate, k, possibly affecting more than one element of A, is not known precisely or controlled. We assume that k has some plausible upper and lower bounds and hence it can be considered as coming from an interval centred at the nominal value $\hat{k}$, i.e. $k \in [\hat{k} - \delta, \hat{k} + \delta]$. We approach the problem by following the nominal dynamics using the previously

defined matrix, A, and defining a bloating term to enclose all solutions arising from admissible parameter values, so that

$$\mathcal{R}_{i+1}(A(k)) \subseteq \mathcal{R}_{i+1}(A(\hat{k})) \oplus \beta_\delta,$$

where $A(k)$ denotes the matrix formed using the parameter value k. For positive x, we can bound $A(k)x$ with $A(\hat{k})x + Dx$, where D is the $n \times n$ matrix computed as $|A(\hat{k} + \delta) - A(\hat{k})|$. To make Dx independent of the state variable—and thus derive a general formula—we use a conservative estimation of x: $Dx \leq D|x|_{\max}$, where $|x|_{\max}$ can be approximated generally for the whole algorithm, or, more practically, in each time-step based on the states in the current reachable set. We derive $|x|_{\max}$ from the zonotope definition in equation (2.1): given a reachable set in the form (c, G), the ith coordinate of the maximal vector is computed by the formula

$$\{|x|_{\max}\}_i = \{c\}_i + \sum_{k=1}^{p} |\{g_k\}_i|. \tag{3.7}$$

Note that $|x|_{\max}$ might not be in the reachable set, but there is at least one point for each coordinate for which $\{x\}_i = \{|x|_{\max}\}_i$ holds (see electronic supplementary material, figure S2(b)). $|x|_{\max}$ is derived as $|c| + \sum |G|$ with all summations carried out *by rows*. We also take into account a rough estimate of the next reachable set, so that in non-converging cases $|x|_{\max}$ is not underestimated: $|x|_{\max} = \max\{|c| + \sum |G|, e^{A\tau}(|c| + \sum |G|)\}$.

For each coordinate of vector $D|x|_{\max}$ the difference between dynamics under any k and the centre, $\hat{k}$, can be enclosed in the zero-centred set,

$$\forall d_i = \{A(k)x\}_i - \{A(\hat{k})x\}_i; \quad d_i \in (0; \{D|x|_{\max}\}_i), \tag{3.8}$$

from which a zonotope considering all coordinates can be obtained as $(0; \text{diag}\{D|x|_{\max}\})$. From here, we proceed considering this zonotope as an input set centred at 0—in agreement with the fact that trajectories of the nominal value, $\hat{k}$, are already calculated through the transition matrix. Therefore, the bloating set accounting for parameter uncertainty, β_δ, can be computed as

$$\beta_\delta = (\underline{0}; \psi(A, \tau)\text{diag}\{D|x|_{\max}\}). \tag{3.9}$$

3.3. Nonlinear systems

Biological systems are often nonlinear, as even the simplest dimer-formation requires second-order rate laws that cannot be eliminated from the system. Nonlinear systems can lead to complex and unpredictable behaviour and their analysis can be difficult. The most popular way to overcome this is by performing a (piece-wise) linearization of the system. Here, we adopt this strategy, and at each step we linearize the system around its current centre using a first-order Taylor expansion, as described in [42]. This results in a system which is piece-wise linear, and has LTI properties in any interval, but which is overall time-varying; of course, this is only an approximation of the real underlying dynamics. The linearized system calculated at time $i\tau$ is

$$\dot{x}(t) = \left.\frac{\partial f}{\partial x}\right|_{x=c_i} (x(t) - c_i) + f(c_i) = A_i x(t) - A_i c_i + f(c_i). \tag{3.10}$$

This system can be used to obtain the next reachable set together with a bloating factor that accounts for the difference between the original and the linearized dynamics

$$\mathcal{R}_{i+1} = (c_i + \psi(A_i, \tau)f(c_i); e^{A_i\tau}G_i) \oplus \beta_\epsilon, \tag{3.11}$$

where β_ϵ can be computed iteratively as proposed in [42]. Here, we use the zonotopic set representation to obtain a tight estimate of the effect of nonlinearity.

Consider the error function $\epsilon(x) = |f(x) - [A_i(x - c_i) + f(c_i)]|$—using a substitution of x by $(c_i + dx)$, we obtain ϵ as a function of dx, the distance between a point and the centre. We define an upper bound for dx, as in equation (3.7): $\{|dx|_{\max}\}_i = \sum_{k=1}^{p} |\{g_k\}_i|$, and approximate $\epsilon(x)$ with a function monotonously increasing with dx. Hence, we obtain an estimation for which $\epsilon(x) \leq \hat{\epsilon}(|dx|_{\max})$ and construct a generator set

$$G_l = \text{diag}\{\hat{\epsilon}(|dx|_{\max})\}, \tag{3.12}$$

which represents a bound on the nonlinearity of the system.

To obtain the reach set of the time-interval $[i\tau, (i+1)\tau]$ further approximations have to be applied to enclose all trajectories between the two time-points. As the piece-wise linear system is time-varying, the estimation of the time-interval $[0, \tau]$ is not sufficient. Instead, as mentioned before (see also in electronic
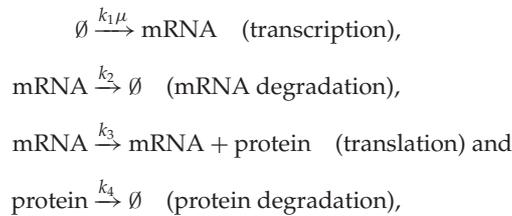
supplementary material, figure S1(b)), the convex hull of two delimiting reach sets is calculated and bloated, the bloating factor for which can be derived in several ways, e.g. as in [9] or [42].

Although the computation in equation (3.11) assumes a constant time-step, it can be modified to incorporate variable steps, τ_i, without loss of efficiency, as in the nonlinear case the matrix exponential, $e^{A_i\tau}$, should anyway be computed in each iteration. Consequently, the first-order linearization used above can be combined with a higher-order, adaptive method by computing the central trajectory (c_i) and each time-step (τ_i) with the higher-order numerical technique, followed by reachable set computation based on these quantities. This way stiff systems with rapidly changing dynamics can be tackled without significant additional computing time, and slowly changing systems can gain a speed-up compared with fixed time-step calculations. Alternatively, we could replace the linearization with a more flexible description of the stochasticity than the LNA, such as moment closures [30] or finite-state projection methods [43].

4. Example applications

4.1. Control of gene expression noise

The first example considered is a controlled stochastic gene expression system [28,44]. In spite of its simplicity, this model is one of the most important examples in practice as protein production is a necessary and elementary building module in real and synthetic biological systems. Many applications might rely on the steady operation of such a unit, hence it is of great practical interest to know what the achievable noise levels and possible states of operation are. Let us take a simple depiction of protein production that focuses on two macromolecules, mRNA and protein, captured in two quantities of interest (mRNA and protein copy numbers, m and p, respectively). The overall process is broken up into four reactions, as follows:

$$\emptyset \xrightarrow{k_1\mu} \text{mRNA} \quad \text{(transcription)},$$

$$\text{mRNA} \xrightarrow{k_2} \emptyset \quad \text{(mRNA degradation)},$$

$$\text{mRNA} \xrightarrow{k_3} \text{mRNA} + \text{protein} \quad \text{(translation) and}$$

$$\text{protein} \xrightarrow{k_4} \emptyset \quad \text{(protein degradation)},$$

where the symbol $\emptyset$ represents that a molecule is produced from (or degraded into) some substance not modelled in the representation. This system translates into the stoichiometry matrix

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix},$$

with each column standing for molecular abundance changes during a reaction of the ones listed above. The propensity of each reaction is formed by its characteristic rate and the molecular species' abundances,

$$a_1 = k_1\mu; \quad a_2 = k_2m; \quad a_3 = k_3m \quad \text{and} \quad a_4 = k_4p.$$

In this example, the dimensionless multiplier in the first reaction, μ, represents our control over the transcription rate, assuming either discrete ($\mu = 0$ or $\mu = 1$) or continuous ($\mu \in [0,1]$) values. A typical control signal would be a sequence of zeros and ones, delivered through, e.g. a light-sensitive set-up [45], switching transcription on and off for some time period. After performing the LNA, we obtain a set of five ordinary differential equations, determining the time-evolution of the mean of m and p, the variance of m (σ_m), the covariance of the two species (σ_{mp}) and finally the variance of the protein abundance (σ_p),

$$\frac{\mathrm{d}}{\mathrm{d}t} \begin{bmatrix} m \\ p \\ \sigma_m \\ \sigma_{mp} \\ \sigma_p \end{bmatrix} = \begin{bmatrix} k_1\mu - k_2m \\ k_3m - k_4p \\ -2k_2\sigma_m + k_1\mu + k_2m \\ k_3\sigma_m - (k_2 + k_4)\sigma_{mp} \\ 2k_3\sigma_{mp} - 2k_4\sigma_p + k_3m + k_4p \end{bmatrix}.$$

The system is started from $\mathcal{I} = \underline{0}$, corresponding to no previously produced target molecules in the system and we investigate the reachable sets up to time $T = 10$ with time-step $\tau = 0.01$. The time-units are normalized according to the protein's average turnover time, so that we gain a general model that
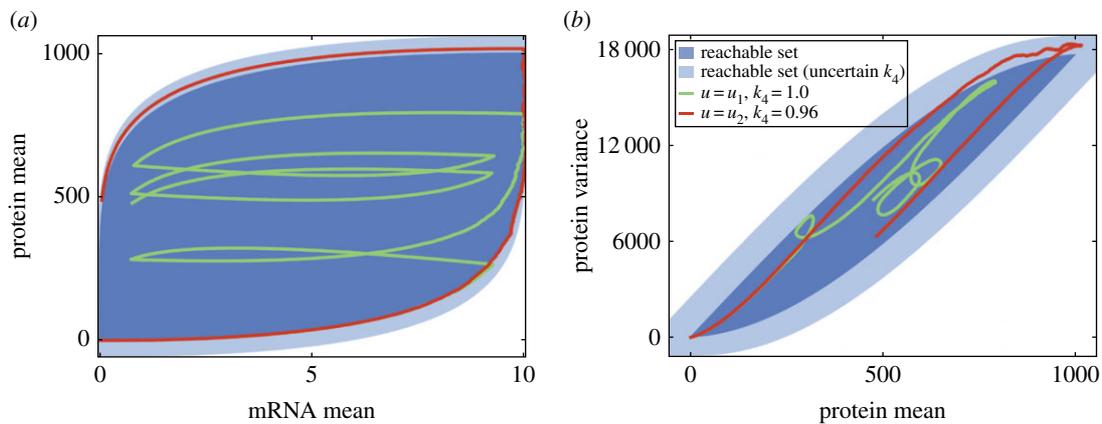
Figure 1. Reachable states of the stochastic gene expression system with controlled transcription and additional uncertainty. Blue-shaded regions show projection of the final reachable set to (*a*) the mRNA mean–protein mean plane and (*b*) the protein mean–protein variance plane. Dark and light blue shades indicate reachable sets without and with 5% uncertainty in parameter k_4. Red and green example trajectories are calculated from 10 000 exact simulations, under the input sequences $u_1 = [1, 0, 1, 1, 1, 0, 1, 0, 1, 0]$ and $u_2 = [1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 1]$, respectively, with protein degradation values as indicated in legend.

might be later scaled to reflect a specific macromolecule. Therefore, k_1 has units of [molecules/time-unit (t.u.)], and all other k_i are in [1/t.u.]. We choose the reaction rates $\underline{k} = [100, 5, 100, 1]$ to show biologically relevant ratios of reaction speeds and steady states: under constitutive production there will be on average 20 mRNA and 2000 protein molecules.

Figure 1 shows projections of the final reachable set together with two sample trajectories. The samples are computed as the mean and variance of a population consisted of 10 000 direct realizations of the original stochastic system. The input signal $\mu(t)$ is defined as a piece-wise constant function with values randomly drawn from {0, 1} and switching every 60 or 30 s (u_1 and u_2, respectively); the signal is kept the same for all realizations contributing to a particular trajectory in figure 1. We also take into account some uncertainty regarding the protein degradation rate, i.e. that the value of k_4 is unknown or can be externally changed to values within 5% of the nominal value, 1. The light blue regions in figure 1 show the estimate of the reachable set under such uncertainty. Interestingly, while a substantial part of the mRNA—protein mean space can be covered, the reachable states on the protein mean—variance plane are limited to a very narrow band. Therefore, if relying on the production of a protein with this module we will have to make a compromise: either have a low number of proteins produced, or a high amount but with great variability (as may be expected given the Poisson nature of this process). In addition, we find that uncertainty in the degradation rates has a profound effect on the general uncertainty of both the mean and variance levels (see electronic supplementary material, figure S3)—in agreement with previous findings on the importance of degradation [46,47].

4.2. Model validation

In our second application, we consider a chain of three molecules affecting each others' production and demonstrate how our reachability analysis can contribute to model validation [48] for stochastic systems. The molecules A, B and C in figure 2 can represent any chain of interacting species with similar reaction networks; for example, a simple model of transcription factors, where only protein levels are modelled explicitly. The regulatory effect of these molecules is through the production rate of their target species. Three different wiring schemes are considered: (i) molecule A induces molecule B, which in turn activates the production of C; (ii) molecule A also activates molecule C, such that this effect is more profound than the activation via B; (iii) molecule C feeds back onto, and induces the production of A. In each model, we simplify the mathematical description by assuming equal degradation rates for all species; hence the models can be summarized by five parameters: k_{AB}, k_{BC}, k_{AC}, k_{CA} and k_{deg}, where the subscript XY refers to the activation of Y by X. In all models $k_{AB} = 1$ and $k_{deg} = 0.8$, and the other parameters are chosen to reflect the connections of the model and produce similar maximal values in the output, C. All reaction rate values have units of [1/t.u.], where the arbitrary unit of time relates to the speed of changes in the system. However, the general shape of dynamics is determined by the relative weights of
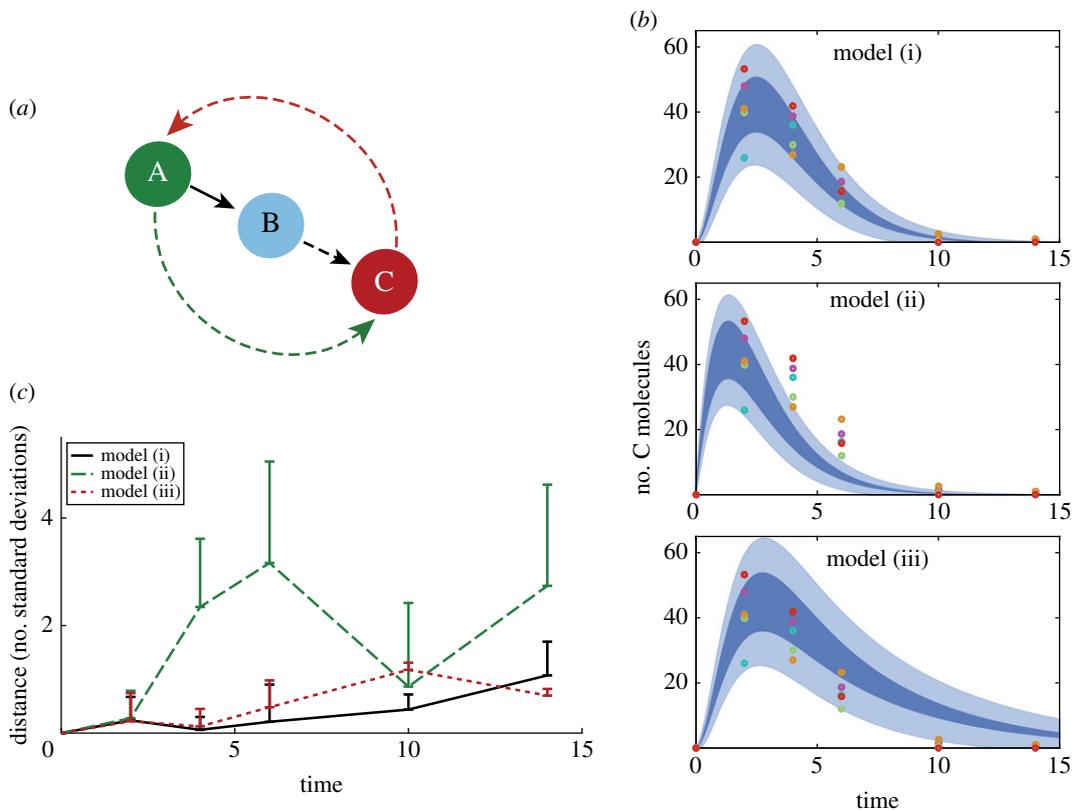
Figure 2. Model evaluation by comparing reachable sets and single measurements. (*a*) Schematics of the three reaction chain models. Models differ in rates corresponding to dashed arrows. (i) $k_{BC} = 1$, $k_{AC} = k_{CA} = 0$. (ii) $k_{BC} = 0.1$, $k_{AC} = 0.9$, $k_{CA} = 0$. (iii) $k_{BC} = 1$, $k_{AC} = 0$, $k_{CA} = 0.2$. (*b*) Reachable region over time of the output (molecule C) starting from the initial set $80 \le A_0 \le 120$, $B_0 = C_0 = 0$. Dark areas represent reachable values of the mean, light blue shades are the ± 1 s.d. region computed with the maximal reachable value of the variance. Coloured circles are sample points taken from single exact simulations of model (i). (*c*) Distance of observation points from the reachable set of mean values. Lines show the average distance of observed data points at each time of measurement, the top of error bars depict the maximal distance at the evaluation points (colours as indicated in legend). Distances for each model are normalized by the maximal reachable standard deviation value. Time in (*b*) and (*c*) is in arbitrary units based on the macromolecular production rate, such that with maximal interaction strength, on average one product molecule is created in 1 t.u.

certain reactions, which the above simplified rate values show more clearly. These can be later scaled to represent a particular system with known time-scales. For example, if it is known that the production of molecule B through the activation of A results in approximately 500 molecules each hour, k_{AB} and k_{deg} should become 500 and 400 [1/hour], but the following analysis would only differ in the scale of the horizontal axis.

In order to model measurements, we generate five individual stochastic trajectories from model (i), with parameter values $k_{BC} = 1$, $k_{AC} = k_{CA} = 0$ and an initial A value randomly chosen from the range [80, 120]. We then sample the amount of molecule C, associated with the experimentally measurable output, from these simulations at six time-points; and evaluate the distance of each measurement point from the reachable set of mean values at that specific time. The distance is normalized by the maximum reachable value of standard deviation and consequently zero for all points within the reachable set. As figure 2*b*,*c* show, it is very unlikely that the observations arose from model (ii). Although the true model shows the best correspondence to the data, model (iii)—representing a feedback system—cannot be discarded due to the wide range of values molecule C can reach in that specific wiring scheme. If the relative noise level is reduced by raising the average initial abundance (to approx. 300 molecules), or measurements from another species (e.g. molecule A) is also available, the distinction between different models becomes clear with model (i) unambiguously fitting the data best.

This should only serve as an illustration of the potential use of reachability analysis for model (in)validation. Coupled with, e.g. experimental design methods [49,50], we can, for example, iteratively rule out candidate models.

4.3. Bistable system

Our last example is a model exhibiting bistability, on which we demonstrate limitations of zonotope-based reach set computation, but also how reachability analysis can be used for *qualitative* exploration of biological system dynamics. As bistability plays a crucial role in cellular decision making [51], we choose a concrete stem cell differentiation model to examine this phenomenon. The model describes the interaction of the Oct4-Sox2 complex with stem cell marker Nanog [52]. Although it is a simplified representation of the underlying network that summarizes numerous reactions into rates used in modelling, it still produces bistable behaviour. The model has two variables (N and OS) and five reactions,

$$a_1 = k_1 \cdot \left(\frac{k_2 OS^{\gamma_{os}}}{(K + OS^{\gamma_{os}})k_{\deg}} \right)^2 ; \quad a_2 = k_{\deg} OS;$$

$$a_3 = \frac{k_3 OS^{\gamma_{os}}}{K + OS^{\gamma_{os}}}; \quad a_4 = \frac{k_4 N^{\gamma_n}}{K + N^{\gamma_n}} \quad \text{and} \quad a_5 = k_{\deg} N,$$

where the reactions stand for auto-activated production of the Oct4-Sox2 complex, degradation of the complex, Oct4-Sox2 activated production of Nanog, auto-activated production and degradation of Nanog, respectively. Accordingly, the stoichiometry matrix of the system is

$$S = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}.$$

We take a set of parameter values from [52] that makes the model bistable,

$$[k_1, k_2, k_3, k_4, k_{\deg}, K, \gamma_{OS}, \gamma_N] = [0.03, 50, 0.1, 14, 1, 10, 1, 2].$$

Here, we focus on how bistability affects our prediction of the average behaviour. Therefore, in practice, our analysis is equivalent to computing the reachable sets of the corresponding deterministic system. We fix rate parameters to the above values, and introduce uncertainty through allowing a range of initial values. As noise in the initial Oct4-Sox2 concentration, OS_0, has no significant effect on the dynamics, we fix its value to 60. Similarly, variance and covariance values do not influence mean Nanog level, hence their value is also without variability, $\sigma_{os} = 10$, $\sigma_{os-n} = 1$, $\sigma_n = 0.1$. On the other hand, as the system is in a bistable regime, Nanog initial values determine which of the two fixed points—corresponding to high-Nanog, stem cell phenotype or low-Nanog, differentiated phenotype—is reached by the system. Therefore, we explore different initial values of Nanog, N_0, with a 25% uncertainty in each case.

The value 0.6 is close to the boundary between the two basins of attraction (c.f. fig. 2A in [52]). Hence, the initial set $\mathcal{I} = [0.45, 0.75]$ inevitably contains a sample of states from both basins. We first compute the reachable set up until 15 arbitrary time units. As zonotopes are limited to describe centrally symmetric and convex reach sets, the set enclosing both high- and low-Nanog steady states also includes a set of implausible, negative values that arise as the symmetric counterparts of the highly positive (greater than 10) Nanog values. We exclude this range of values as it is clearly an artefact of set representation and concentrate on the strictly positive subset of reachable states. Figure 3 shows Nanog mean values in blue on a logarithmic scale as a function of time. As a comparison, we also conduct reachability analysis restricted to the stem cell or differentiated cell regime. The former is started from the initial set $\mathcal{I} = [1.5, 2.5]$ and the latter from Nanog values $N_0 \in [0.3, 0.5]$. The collective evolution of trajectories is shown with red and green reach sets in figure 3.

It is clear from the comparison of the shaded regions in figure 3 that the final reachable set obtained with $N_0 = 0.6$ indeed contains both steady states. Furthermore, the upper limit of the reach set is not only conservative, but also tight. Therefore, the only crude over-approximation is due to symmetry and the high-Nanog state is well outlined even in this analysis. In addition, obtaining negative reachable sets for variables that should not assume negative values can be used as an indicator of discontinuous behaviour within the analysis. After such an observation, restricting the set of initial values—similarly to how the second and third reachable sets are obtained in figure 3—can shed light on the underlying dynamics.

5. Discussion

In this work we have introduced a method to compute the states that a stochastic biochemical network can access under a control input signal. The method is particularly applicable to determine reachable mean–variance values of the investigated species, and hence estimate noise levels of a system. Our aim
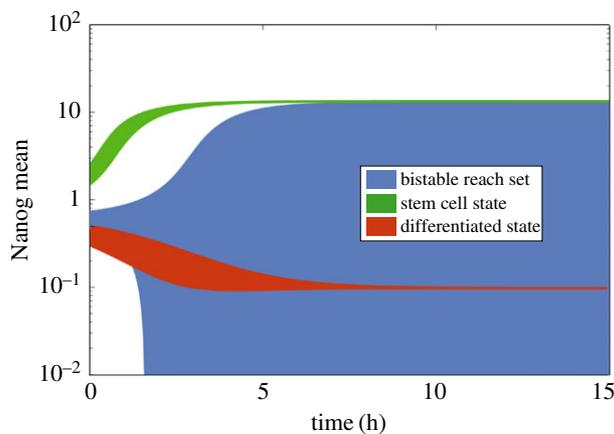
Figure 3. Investigating bistability via reachable sets of a stem cell differentiation model. Shaded regions show a conservative estimate of the achievable mean Nanog levels over time. The dark blue region is an over-approximation of the reachable set of the bistable system, when the analysis is started from an initial set of bistability, i.e. a set enclosing states leading to both fixed points. $N_0 = 0.6 \pm 0.15$. Red and green shades show reachable sets computed from 'monostable' initial sets, from which all trajectories converge to the differentiation fixed point (low-Nanog level, in red) or to the stem cell fixed point (high-Nanog level, in green). Initial Nanog values are 0.4 ± 0.1 and 2 ± 0.5, respectively.

is to provide a computationally efficient tool that could be used in the search for the best modelling description or regulatory design of stochastic networks.

Here, we used the linear noise and moment expansion approximations to extract deterministic equations of the mean and covariance matrix that served as a basis for our reachable set computation. Both can be accessed on GitHub as parts of our Python package, MEANS [31]. We have presented a general algorithm for linear and nonlinear reachability analysis based on zonotopes, together with formulae for handling relevant input types: addition of molecules and control of a reaction rate. All steps are implemented in Matlab to semi-automatically generate the reachable sets for a problem defined in terms of a transition function, input matrix and bounds, and parameter uncertainty. The algorithm can be executed on any given set of ODEs describing some characteristics of the system, hence it is applicable regardless of the approximation method used for the generation of equations and also to deterministic biological models.

We demonstrate the method on three schematic models of biological macromolecules: a controlled gene expression system; a cascade of modifying molecules, e.g. transcription factors; and a bistable model of stem cell differentiation. In all examples the approximation is conservative, hence trajectories randomly generated from a set of admissible signals and initial values are confined within the set our method predicts. In linear cases, if we have precise knowledge of the parameter values and input bounds, the predicted reachable set will be exact as well, i.e. each point in it can be actually accessed by an appropriate input sequence. This cannot be ensured for nonlinear systems (or uncertain rate values) as the conservative estimates used for the nonlinearity (or reaction rate effect) are obtained from approximations the system might not take. However, our estimation reflects the general characteristics of the system—e.g. if some variables of a generally nonlinear system have linear equations—and hence unnecessary over-approximations are avoided to provide a tight estimate. This is crucial in the evaluation of various control designs on the basis of mean–variance values reachable through them.

Our method is based on a zonotope representation of reachable, initial and parameter sets. Set operations can be efficiently computed using this representation method, which allows us to apply our reachability analysis to many-variable, complicated biological networks. However, the number of generator vectors in the reachable set increases in every iteration, and for a long time horizon or small time-step the algorithm can become expensive regarding memory space. In smaller systems, such as our examples, this effect is still negligible, but for high-dimensional cases the increase is more significant. For such cases one can turn to the zonotope-reduction technique presented by Girard [9] to limit the size of zonotopes to an adjustable value.

Note also that zonotopes, and hence the sets in our analysis, are by definition convex and centrally symmetric. Therefore, systems where the actual reachable set is concave or consisted of multiple sets—especially networks with bi- and multi-stability—will be largely over-approximated. In our last example,

we have tested this aspect of the algorithm on a differentiation model with bistable dynamics. As the two expression level states in such cases are typically of different magnitudes, our method is forced to enclose a high proportion of negative values besides also incorporating a non-accessible region between the two states. However, we also found that this drop in approximation quality is a good indicator of the system entering a bistable regime of initial conditions/input signal. A similar over-approximation case can arise for imprecise parameter values of the transition matrix: although admissible parameter sets are symmetric sets given as a nominal value with error bounds, the influence of parameters is usually not symmetric on the reachable set. Therefore zonotopes lead to a rough approximation when some rates show high uncertainty. Just like in the case of multi-stability, a preliminary analysis using our general method can reveal these issues, and point one towards a more exhaustive study using a series of reachable sets for fixed values (or small intervals) of initial values or rate parameters to cover the range of interest and extrapolate for tighter approximation.

As often the case in the analysis of biochemical systems, incomplete information about the model structure and crucial constants of the system makes a detailed and informative analysis impossible. Our method is not applicable for the exploration of differences in model structures; this problem can be treated implicitly by either doing independent analysis of all candidate models or choosing rate parameter uncertainty such that zero (i.e. no interaction) is among the admissible values. The latter is likely to give rise to an over-generalized approximation, as described above; while the first one is only advisable for a small number of possible models, like in our second example. Furthermore, limited measurements and high levels of noise can also influence the method's power for validation, as figure 2c demonstrates: in such cases flexible models might be favoured even over the true model. Therefore, we advise the use of other tools, such as topological sensitivity analysis [53] coupled with our method for a more thorough investigation of model space.

References

1. Khalil AS, Collins JJ. 2010 Synthetic biology: applications come of age. *Nat. Rev. Genet.* **11**, 367–379. (doi:10.1038/nrg2775)

2. Elowitz MB, Leibler S. 2000 A synthetic oscillatory network of transcriptional regulators. *Nature* **403**, 335–338. (doi:10.1038/35002125)

3. Gardner TS, Cantor CR, Collins JJ. 2000 Construction of a genetic toggle switch in *Escherichia coli*. *Nature* **403**, 339–342. (doi:10.1038/35002131)

4. Milias-Argeitis A, Summers S, Stewart-Ornstein J, Zuleta I, Pincus D, El-Samad H, Khammash M, Lygeros J. 2011 *In silico* feedback for *in vivo* regulation of a gene expression circuit. *Nat. Biotechnol.* **29**, 1114–1116. (doi:10.1038/nbt.2018)

5. Menolascina F, Fiore G, Orabona E, De Stefano L, Ferry M, Hasty J, di Bernardo M, di Bernardo D. 2014 *In-vivo* real-time control of protein expression from endogenous and synthetic gene networks. *PLoS Comput. Biol.* **10**, e1003625. (doi:10.1371/journal. pcbi.1003625)

6. McMahon SS, Lenive O, Filippi S, Stumpf MPH. 2015 Information processing by simple molecular motifs and susceptibility to noise. *J. R. Soc. Interface* **12**, 20150597. (doi:10.1098/rsif.2015.0597)

7. Filippi S *et al.* 2016 Robustness of MEK-ERK dynamics and origins of cell-to-cell variability in MAPK signaling. *Cell Rep.* **15**, 2524–2535. (doi:10.1016/j.celrep.2016.05.024)

8. Fu AQ, Pachter L. 2016 Estimating intrinsic and extrinsic noise from single-cell gene expression measurements. *Stat. Appl. Genet. Mol. Biol.* **15**, 447–471. (doi:10.1515/sagmb-2016-0002)

9. Girard A. 2005 Reachability of uncertain linear systems using zonotopes. In *Hybrid systems: computation and control* (eds M Morari, L Thiele), pp. 291–305. Berlin, Germany: Springer.

10. Audemard G, Bozzano M, Cimatti A, Sebastiani R. 2005 Verifying industrial hybrid systems with MathSAT. *Electron. Notes Theor. Comput. Sci.* **119**, 17–32. (doi:10.1016/j.entcs.2004.12.022)

11. Cross EA, Mitchell IM. 2008 Level set methods for computing reachable sets of systems with differential algebraic equation dynamics. In *American Control Conf., Seattle, WA, 11–13 June*, pp. 2260–2265.

12. Prandini M, Hu J. 2009 Application of reachability analysis for stochastic hybrid systems to aircraft conflict prediction. *IEEE Trans. Automat. Control* **54**, 913–917. (doi:10.1109/TAC.2008.2011011)

13. Scacchioli A, Bayen AM, Stojadinovic B. 2014 Assessment of uncertainty propagation in the dynamic response of single-degree-of-freedom structures using reachability analysis. *J. Eng. Mech.* **140**, 1–12. (doi:10.1061/(ASCE)EM.1943-7889. 0000676)

14. Althoff M, LeGuernic C, Krogh BH. 2011 Reachable set computation for uncertain time-varying linear systems. In *Proc. of the 14th Int. Conf. on Hybrid Systems: Computation and Control, HSCC '11*, pp. 93–102. New York, NY: ACM.

15. Kurzhanski AB, Vályi I. 1997 *Ellipsoidal calculus for estimation and control*. Laxenburg, Austria: IIASA.

16. Greenstreet MR, Mitchell I. 1999 *Reachability analysis using polygonal projections*, vol. 1569, pp. 103–116. Berlin, Heidelberg: Springer.

17. Han Z, Krogh BH. 2006 Reachability analysis of large-scale affine systems using low-dimensional polytopes. In *Hybrid systems: computation and control* (eds J Hespanha, A Tiwari), pp. 287–301, Berlin, Heidelberg: Springer.

18. Asarin E, Dang T, Girard A. 2003 Reachability analysis of nonlinear systems using conservative approximation. In *Hybrid systems: computation and control* (eds O Maler, A Pnueli). LNCS, vol. 2623, pp. 20–35, Berlin, Heidelberg: Springer.

19. Mitchell IM, Templeton JA. 2005 A toolbox of Hamilton-Jacobi solvers for analysis of nondeterministic continuous and hybrid systems. In *Hybrid systems: computation and control* (eds M Morari, L Thiele), pp. 480–494, Berlin, Heidelberg: Springer.

20. Althoff M, Krogh BH. 2014 Reachability analysis of nonlinear differential-algebraic systems. *IEEE Trans. Automat. Control* **59**, 371–383. (doi:10.1109/TAC. 2013.2285751)

21. Szederkényi G, Kovács M, Hangos KM. 2002 Reachability of nonlinear fed-batch fermentation processes. *Int. J. Robust Nonlin. Control* **12**, 1109–1124. (doi:10.1002/rnc.686)

22. Brim L, Fabriková J, Drazan S, Safránek D. 2011 Reachability in biochemical dynamical systems by quantitative discrete approximation. *Comput. Res. Repository*. (http://arxiv.org/abs/1107.5924)

23. Halász ÁM, Kumar V, Imieliński M, Belta C, Sokolsky O, Pathak S, Rubin H. 2007 Analysis of lactose metabolism in *E. coli* using reachability analysis of hybrid systems. *IET Syst. Biol.* **1**, 130–148. (doi:10.1049/iet-syb:20060035)

24. Dang T, Le Guernic C, Maler O. 2011 Computing reachable states for nonlinear biological models. *Theor. Comput. Sci.* **412**, 2095–2107. (doi:10.1016/j.tcs.2011.01.014)

25. Testylier R, Dang T. 2012 Analysis of parametric biological models with non-linear dynamics. In *Proc. of the First Int. Workshop on Hybrid Systems and Biology* (eds E Bartocci, L Bortolussi). *Electron Proc. Theor. Comput. Sci.* **92**, 16–29. Open Publishing Association.

26. El-Samad H, Fazel M, Liu X, Papachristodoulou A, Prajna S. 2006 Stochastic reachability analysis in complex biological networks. In *American Control Conf., Minneapolis, MN, 14–16 June*, pp. 6–12.

27. Riley D, Koutsoukos X, Riley K. 2007 Verification of biochemical processes using stochastic hybrid systems. In *IEEE Int. Symp. on Intelligent Control*, pp. 100–105.

28. Parise F, Valcher ME, Lygeros J. 2014 On the reachable set of the controlled gene expression system. In *IEEE Conf. on Decision and Control*. Los Angeles, CA, USA.

29. Ale A, Kirk P, Stumpf MPH. 2013 A general moment expansion method for stochastic kinetic models. *J. Chem. Phys.* **138**, 174101. (doi:10.1063/1.4802475)

30. Lakatos E, Ale A, Kirk PDW, Stumpf MPH. 2015 Multivariate moment closure techniques for stochastic kinetic models. *J. Chem. Phys.* **143**, 094107. (doi:10.1063/1.4929837)

31. Fan S, Geissmann Q, Lakatos E, Lukauskas S, Ale A, Babtie AC, Kirk PDW, Stumpf MPH. 2016 MEANS: python package for moment expansion approximation, iNference and simulation. *Bioinformatics* **32**, 2863–2865. (doi:10.1093/bioinformatics/btw229)

32. Kirk PDW, Babtie AC, Stumpf MPH. 2015 Systems biology (un)certainties. *Science* **350**, 386–388. (doi:10.1126/science.aac9505)

33. Fukuda K. 2004 From the zonotope construction to the Minkowski addition of convex polytopes. *J. Symb. Comput.* **38**, 1261–1272. (doi:10.1016/j.jsc.2003.08.007)

34. VanKampen. NG. 2011 *Stochastic processes in physics and chemistry*. North-Holland Personal Library. Elsevier Science.

35. McQuarrie DA. 1967 Stochastic approach to chemical kinetics. *J. Appl. Probab.* **4**, 413–478. (doi:10.1017/S002190020002547X)

36. Ruess J, Lygeros J. 2015 Moment-based methods for parameter inference and experiment design for stochastic biochemical reaction networks. *ACM Trans. Model Comput. Simul.* **25**, 8:1–8:25. (doi:10.1145/2688906)

37. Schnoerr D, Sanguinetti G, Grima R. 2014 Validity conditions for moment closure approximations in stochastic chemical kinetics. *J. Chem. Phys.* **141**, 084103. (doi:10.1063/1.4892838)

38. Schnoerr D, Sanguinetti G, Grima R. 2015 Comparison of different moment-closure approximations for stochastic chemical kinetics. *J. Chem. Phys.* **143**, 185101. (doi:10.1063/1.4934990)

39. Szederkényi G, Lakner R, Gerzson M. 2006 *Intelligent control systems: an introduction with examples*. Springer US. (doi:10.1007/b101833)

40. Aubin J-P, Cellina A. 1984 Differential inclusions: set-valued maps and viability theory. In *Grundlehren der mathematischen Wissenschaften* (eds A Cherciner, J Coates, SRS Varadhan), vol. 264. Berlin, Heidelberg: Springer.

41. Erguler K, Stumpf MPH. 2011 Practical limits for reverse engineering of dynamical systems: a statistical analysis of sensitivity and parameter inferability in systems biology models. *Mol. Biosyst.* **7**, 1593–1602. (doi:10.1039/c0mb00107d)

42. Han Z, Krogh BH. 2006 Reachability analysis of nonlinear systems using trajectory piecewise linearized models. In *American Control Conf., Minneapolis, MN, 14–16 June*, pp. 1505–1510.

43. Munsky B, Khammash M. 2006 The finite state projection algorithm for the solution of the chemical master equation. *J. Chem. Phys.* **124**, 044104. (doi:10.1063/1.2145882)

44. Ingram PJ, Stumpf MPH, Stark J. 2008 Nonidentifiability of the source of intrinsic noise in gene expression from single-burst data. *PLoS Comput. Biol.* **4**, e1000192. (doi:10.1371/journal.pcbi.1000192)

45. Tabor JJ, Levskaya A, Voigt CA. 2011 Multichromatic control of gene expression in *Escherichia coli*. *J. Mol. Biol.* **405**, 315–324. (doi:10.1016/j.jmb.2010.10.038)

46. Komorowski M, Mikisz J, Stumpf MPH. 2013 Decomposing noise in biochemical signaling systems highlights the role of protein degradation. *Biophys. J.* **104**, 1783–1793. (doi:10.1016/j.bpj.2013.02.027)

47. Lakatos E, Salehi-Reyhani A, Barclay M, Stumpf MPH, Klug DR. 2017 Protein degradation rate is the dominant mechanism accounting for the differences in protein abundance of basal p53 in a human breast and colorectal cancer cell line. *PLoS ONE* **12**, e0177336. (doi:10.1371/journal.pone.0177336)

48. Anderson J, Papachristodoulou A. 2009 On validation and invalidation of biological models. *BMC Bioinform.* **10**, 132. (doi:10.1186/1471-2105-10-132)

49. Liepe J, Filippi S, Komorowski M, Stumpf MPH. 2013 Maximizing the information content of experiments in systems biology. *PLoS Comput. Biol.* **9**, e1002888. (doi:10.1371/journal.pcbi.1002888)

50. Silk D, Kirk P, Barnes CP, Toni T, Stumpf MPH. 2014 Model selection in systems biology depends on experimental design. *PLoS Comput. Biol.* **10**, e1003650. (doi:10.1371/journal.pcbi.1003650)

51. Ferrell Jr JE. 2012 Bistability, bifurcations, and Waddington's epigenetic landscape. *Curr. Biol.* **22**, R458–66. (doi:10.1016/j.cub.2012.03.045)

52. Glauche I, Herberg M, Roeder I. 2010 Nanog variability and pluripotency regulation of embryonic stem cells—insights from a mathematical model analysis. *PLoS ONE* **5**, e11238. (doi:10.1371/journal.pone.0011238)

53. Babtie AC, Kirk P, Stumpf MPH. 2014 Topological sensitivity analysis for systems biology. *Proc. Natl Acad. Sci. USA* **111**, 18 507–18 512. (doi:10.1073/pnas.1414026112)