Statistical Model Checking Based Calibration and Analysis of Bio-pathway Models*

Sucheendra K. Palaniappan[1], Benjamin M. Gyori[2], Bing Liu[3], David Hsu[1,2], and P.S. Thiagarajan[1,2]

[1] School of Computing, National University of Singapore, 117417, Singapore
[2] NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, 117417, Singapore
[3] Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract. We present a statistical model checking (SMC) based framework for studying ordinary differential equation (ODE) models of bio-pathways. We address cell-to-cell variability explicitly by using probability distributions to model initial concentrations and kinetic rate values. This implicitly defines a distribution over a set of ODE trajectories, the properties of which are to be characterized. The core component of our framework is an SMC procedure for verifying the dynamical properties of an ODE system accompanied by such prior distributions. To cope with the imprecise nature of biological data, we use a formal specification logic that allows us to encode both qualitative properties and experimental data. Using SMC, we verify such specifications in a tractable way, independent of the system size. This further enables us to develop SMC based parameter estimation and sensitivity analysis procedures. We have evaluated our method on two large pathway models, namely, the segmentation clock network and the thrombin-dependent MLC phosphorylation pathway. The results show that our method scales well and yields good parameter estimates that are robust. Our sensitivity analysis framework leads to interesting insights about the underlying dynamics of these systems.

1 Introduction

Biochemical networks–often called bio-pathways–govern a variety of cellular functions. Their malfunctioning can lead to major diseases [1]. Thus it is important to understand their dynamics using mathematical models [2]. However, building and analyzing such models poses considerable challenges. In this paper, we address the particular challenge of accounting for variable behavior across individual cells. A natural way to cater for this is to use a probabilistic system model such as continuous time Markov chains (CTMCs) [3]. However, such models typically track the occurrences of individual reactions. Hence for pathways of realistic size, calibrating these models using experimental data and analyzing them using stochastic simulations is very difficult. The alternative is to use ordinary differential equations (ODEs) to capture the dynamics. This approach is often computationally more tractable, although it requires that the number of molecules of each type involved in the pathway be abundantly present [4]. In this paper our focus

* This research was partially supported by the Singapore MOE ARC grant MOE2011-T2- 2-012.

A. Gupta and T.A. Henzinger (Eds.): CMSB 2013, LNBI 8130, pp. 120–134, 2013.

is on accounting for cell-to-cell variability in the setting of ODE based models. Specifically, our main contribution is a statistical model checking (SMC) based framework, using which a system with such variability can be efficiently calibrated and analyzed.

Variability in a population of cells has at least two major causes. First, as shown in [5], differences in the initial concentrations of proteins are the primary source of variability in response to external stimuli. Second, due to differing internal and external conditions among cells, the values of kinetic rate constants also vary across cells [6, 7]. In our ODE setting the variables will represent the concentrations of the biochemical species (typically proteins) in the pathway, and hence the initial concentrations of these species will constitute the initial values of the variables. Further, the parameters appearing in the equations will consist of the kinetic rate constants governing the reactions. Thus we can capture cell-to-cell variability in the behavior of the bio-pathway by studying the ODE dynamics across a range of values for the initial concentrations and kinetic rate constant values. We do this in a probabilistic setting by assuming initial probability distributions (usually uniform) over an interval of values for the initial concentrations and rate constants. We then show that the resulting space of trajectories can be used to construct a natural probability measure space if the vector field defined by the ODE system is continuously differentiable. In our setting this requirement is easily met.

To analyze the ODE system, we first formalize properties using our specification logic and decide a corresponding confidence level (probability) with which we wish to assess them. Consequently, an SMC procedure –which poses the problem as a hypothesis test– is used to decide approximately, but with statistical guarantees, whether the properties are satisfied with the desired probability. SMC continues to sample and verify trajectories from the ODE system until a decision can be made. It is well-established that SMC is efficient since its complexity does not depend on the size of the system. Moreover, posing the problem as a sequential hypothesis test reduces the overall number of samples needed to make a decision [8]. These components form a principled method for analyzing the dynamics of a bio-pathway in the presence of dynamic variability across a population of cells.

To demonstrate the applicability of our approach, we develop an SMC based parameter estimation method. The unknown model parameters usually consist of initial concentrations and kinetic rate constants. Here, for convenience, we shall assume all the initial concentrations are known but that their nominal values can vary over a cell population. The parameter estimation procedure searches through the value space of the unknown parameters to determine the "best" combination of values that can explain the given data and predict new behaviors [9]. The key step in this procedure is to determine the fitness-to-data of the current set of parameter values. We use our specification logic to encode both experimental time series data and known qualitative trends concerning the dynamics of the pathway. We then use our SMC procedure to determine the goodness of the given set of parameter values, while taking into account that these values can fluctuate across the population of cells that the data is based on. Subsequently, we use a global optimization strategy known as SRES [10] to choose a new set of candidate parameter values according to the SMC based score assigned to the current set.

An important analysis task to be performed on the model is quantifying the influence of different parameters on the model dynamics. The information gained from such

a sensitivity analysis procedure can help in robustness analysis, optimal experimental design and drug target selection [11]. We show how SMC can be used to generate the statistics needed by the global sensitivity analysis method MPSA [12]. Consequently, one can incorporate a rich class of dynamic behaviors–encoded as formulas in our specification logic–to drive our sensitivity analysis method.

We evaluated our method on two pathway models taken from the BioModels database [13]. For both case studies, we assumed that noisy experimental data and qualitative dynamic traits of a few species were known. This data was separated into training and test components. A subset of the rate constants were assumed to be unknown and estimated using our parameter estimation procedure. The first model, the segmentation clock pathway, consists of 16 differential equations and 75 rate constants, out of which 39 were fixed to be unknown. The second model, the thrombin-dependent myosin light chain (MLC) pathway consists of 105 differential equations and 197 rate constants, out of which 100 were fixed to be unknown. Our results (Section 5) show that our SMC based technique is efficient and scales well. We also applied our sensitivity analysis method to obtain interesting insights into the dynamics of these two bio-pathways.

1.1 Related Work

Probabilistic model checking of stochastic models is an active field of research [14–17]. Of particular interest in our context are sampling based methods such as [18, 19], which verify probabilistic properties using a fixed number of sampled trajectories. In contrast, SMC based methods such as [14, 20] adaptively generate a sufficient number of trajectories to determine if the property is satisfied while meeting the strengths of the statistical test specified by the user. Characterizing the behavior of dynamical systems where the initial conditions and the rate parameters are under-determined has also been discussed in [21] with a focus on sampling methods and computing reachable sets.

Turning to parameter estimation using temporal logic constraints, a brute force search of the parameter space is employed in [16] for Petri nets. In the ODE context, parameter estimation combined with model checking appears in [22] using again a brute force sampling based parameter search approach, and in [23], using an evolutionary strategy to guide the search. However, both these techniques only generate a single simulation trace of the ODE to evaluate a proposed set of parameters. A symbolic model checking approach is explored for the restricted class of multi-affine ODEs in [24, 25]. The work reported in [19] deploys a genetic algorithm to search for the best set of parameters. A fixed number of samples–this number is fixed in an ad hoc manner–is generated, and the probability of satisfying a property is calculated to be the fraction of the samples which satisfy the property. In all these studies, the quality of the estimated parameters is not validated using test data (i.e. data that was not used as training data). While [19] does mention identifying critical parameters, we believe that our approach is the first systematic attempt to develop a property-based sensitivity analysis framework using statistical model checking.

In the next section, we introduce ODE models and their dynamics. In Section 3, we discuss our specification logic and the statistical model checking procedure. Subsequently, we present our parameter estimation and sensitivity analysis framework.

Experimental results are reported in Section 5. Additional experimental results are reported in the supplementary material [26].

2 ODE Based Models and Their Behaviors

A popular formalism for describing the dynamics of a biochemical network is a system of ODEs. For each molecular species x_i in the pathway, there will be an equation of the form $dx_i/dt = f_i(\mathbf{x}, \Theta_i)$. Here f_i describes the kinetics of the reactions that produce and consume x_i, $\mathbf{x}$ denotes the concentrations of the molecular species taking part in these reactions, while the vector Θ_i gives the rate constants governing these reactions.

Each x_i is a real-valued function of $t \in \mathbb{R}_+$, where $\mathbb{R}_+$ denotes the set of non-negative reals. We shall realistically assume that $x_i(t)$ takes values in the interval $[L_i, U_i]$, where L_i and U_i are non-negative rationals with $L_i < U_i$. Hence the state space of the system is $\mathbf{V} = [L_1, U_1] \times \ldots \times [L_n, U_n]$, a bounded subset of $\mathbb{R}^n_+$. Let $\Theta = \bigcup_i \Theta_i = \{\theta_1, \theta_2, \ldots, \theta_m\}$ be the set of all rate constants. We again assume that the range of values for each θ_j is $[L^j, U^j]$ for $1 \leq j \leq m$. We shall present the SMC procedure while assuming that all the rate constants are known. In Section 4, it will become clear how unknown rate constants are handled.

An implicit assumption in what follows is that the value of a rate constant, when fixed initially, does not change during the time evolution of the dynamics, although this value can be different for different cells. To capture the cell-to-cell variability regarding the initial states, we define for each variable x_i an interval $[L_i^{init}, U_i^{init}]$ with $L_i \leq L_i^{init} < U_i^{init} \leq U_i$. The actual value of the initial concentration of x_i is assumed to fall in this interval. Similarly, we shall assume that the nominal value of the rate constant θ_j falls in the interval $[L^j_{init}, U^j_{init}]$ with $L^j \leq L^j_{init} < U^j_{init} \leq U^j$. We set $INIT = (\prod_i [L_i^{init}, U_i^{init}]) \times (\prod_j [L^j_{init}, U^j_{init}])$. Thus $INIT$ captures the cell-to-cell variability in the initial concentration and the rate constant values. In what follows we let $\mathbf{v}$ to range over $\prod_i [L_i^{init}, U_i^{init}]$ and $\mathbf{w}$ to range over $\prod_j [L^j_{init}, U^j_{init}]$.

We will represent our system of ODEs in vector form as $d\mathbf{x}/dt = F(\mathbf{x}, \Theta)$ with $F_i(\mathbf{x}, \Theta) := f_i$. Recall that a function $f_i : \mathbf{V} \to \mathbf{V}$ is a C^1 function if f_i', the derivative of f_i, exists at all $\mathbf{v} \in \mathbf{V}$, and is a continuous function. In the setting of biochemical networks, the expressions in f_i will model kinetic laws such as mass law and Michaelis-Menten [4]. Thus it is reasonable to assume that $f_i \in C^1$ for each i and hence $F : \mathbf{V} \to \mathbf{V}$ is also a C^1 function. As a result, for each $(\mathbf{v}, \mathbf{w}) \in INIT$ the system of ODEs will have a unique solution $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ [27]. Further, it will satisfy: $\mathbf{X}_{\mathbf{v},\mathbf{w}}(0) = \mathbf{v}$ and $\mathbf{X}'_{\mathbf{v},\mathbf{w}}(t) = F(\mathbf{X}_{\mathbf{v},\mathbf{w}}(t))$. We are also guaranteed that $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ is a C^0-function (i.e. continuous function) [27], and hence measurable. This fact will be crucial for our SMC procedure.

It will be convenient to define the flow $\Phi_{\mathbf{w}} : \mathbb{R}_+ \times \mathbf{V} \to \mathbf{V}$ for arbitrary initial vectors $\mathbf{v}$. Intuitively, $\Phi_{\mathbf{w}}(t, \mathbf{v})$ is the state reached under the ODE dynamics if the system starts at $\mathbf{v}$ at time 0. The flow will be the C^0-function given by: $\Phi_{\mathbf{w}}(t, \mathbf{v}) = \mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$. Thus $\Phi_{\mathbf{w}}(0, \mathbf{v}) = \mathbf{X}_{\mathbf{v},\mathbf{w}(0)=\mathbf{v}}$ and $\partial(\Phi_{\mathbf{w}}(t, \mathbf{v}))/\partial t = F(\Phi_{\mathbf{w}}(t, \mathbf{v}))$ for all t. We will, in fact, work with $\Phi_{\mathbf{w},t} : \mathbf{V} \to \mathbf{V}$ where $\Phi_{\mathbf{w},t}(\mathbf{v}) = \Phi_{\mathbf{w}}(t, \mathbf{v})$ for every t and every $\mathbf{v} \in \mathbf{V}$. again, $\Phi_{\mathbf{w},t}$ is guaranteed to be a C^0 function.

In our application, the dynamics will be of interest only up to a maximal time point T. Fixing such a T, a *trajectory* starting from $\mathbf{v} \in \mathbf{V}$ at time 0 and with $\mathbf{w}$ as the parameter

values is denoted $\sigma_{\mathbf{v},\mathbf{w}}$. It is the (continuous) function $\sigma_{\mathbf{v},\mathbf{w}} : [0,T] \rightarrow \mathbf{V}$ satisfying: $\sigma_{\mathbf{v},\mathbf{w}}(t) = X_{\mathbf{v},\mathbf{w}}(t)$. The behavior of our dynamical system is the set of trajectories given by $BEH = \{\sigma_{\mathbf{v},\mathbf{w}} \mid (\mathbf{v},\mathbf{w}) \in INIT\}$. Our goal is to develop an SMC procedure to verify the dynamical properties of BEH.

3 Statistical Model Checking of ODE Dynamics

3.1 Bounded Linear Time Temporal Logic

To formally express dynamical properties of BEH, we use formulas in a specification logic. We will use bounded linear time temporal logic (BLTL) since our trajectories will be of finite duration. An atomic proposition in our logic will be of the form (i,ℓ,u) with $L_i \leq \ell < u \leq U_i$. Such a proposition will be interpreted as "the current concentration level of x_i is in the interval $[\ell,u]$", and we fix a finite set of such atomic propositions.

We first introduce the syntax and then the semantics of BLTL formulas. The formulas of BLTL are defined as: (i) Every atomic proposition as well as the constants *true*, *false* are BLTL formulas. (ii) If ψ, ψ' are BLTL formulas then $\neg\psi$ and $\psi \vee \psi'$ are BLTL formulas. (iii) If ψ, ψ' are BLTL formulas and $t \leq T$ is a *positive integer* then $\psi \mathbf{U}^{\leq t}\psi'$ and $\psi \mathbf{U}^{t}\psi'$ are BLTL formulas. We have mildly strengthened BLTL to be able to express that a certain property will hold exactly at t time units from now. This will enable us to encode experimental data in the specification. The derived propositional operators such as $\wedge$, $\supset$, $\equiv$, and the temporal operators $\mathbf{G}^{\leq t}$, $\mathbf{F}^{\leq t}$ are defined in the usual way.

We will interpret the formulas of our logic at the finite set of time points $\mathcal{T} = \{0,1,\ldots,T\}$. Such a discretization is reasonable since experimental data will be available only at a finite number of discrete time points. Further, qualitative properties of interest are expressible in discrete time. We assume that $\mathcal{T}$ has been chosen appropriately and it includes all the relevant time points with respect to the specified properties.

The semantics of the logic is defined in terms of the relation $\sigma, t \models \varphi$, where σ is a trajectory in BEH and $t \in \mathcal{T}$.

- $\sigma, t \models (i,\ell,u)$ iff $\ell \leq \sigma(t)(i) \leq u$ where $\sigma(t)(i)$ is the i^{th} component of the n-dimensional vector $\sigma(t) \in \mathbf{V}$.
- $\neg$ and $\vee$ are interpreted in the usual way.
- $\sigma, t \models \psi \mathbf{U}^{\leq k}\psi'$ iff there exists k' such that $k' \leq k$, $t + k' \leq T$ and $\sigma, t + k' \models \psi'$. Further, $\sigma, t + k'' \models \psi$ for every $0 \leq k'' < k'$.
- $\sigma, t \models \psi \mathbf{U}^{k}\psi'$ iff $t + k \leq T$ and $\sigma, t + k \models \psi'$. Further, $\sigma, t + k' \models \psi$ for every $0 \leq k' < k$.

We now define $models(\psi) = \{\sigma \mid \sigma, 0 \models \psi,\ \sigma \in BEH\}$.

Next, we wish to make statements of the form $P_{\geq r}(\psi)$, where the intended meaning is that the probability that a trajectory in BEH belongs to $models(\psi)$ is at least r. To assign meaning to such statements, we need to define a probability measure over sets of trajectories. Note, however, that the trajectory $\sigma \in BEH$ is completely determined by $\sigma(0)$, the (vector) value it assumes at $t = 0$. Hence we will identify BEH with $INIT$, the set of initial states. To make this explicit, we define the set $Models(\psi) \subseteq INIT$ as:

$(\mathbf{v}, \mathbf{w}) \in Models(\psi)$ iff $\sigma_{\mathbf{v},\mathbf{w}} \in models(\psi)$. We define the formulas of PBLTL as $P_{\geq r}\psi$ and $P_{\leq r'}\psi$ provided $r \in [0, 1)$, $r' \in (0, 1]$ and ψ is a BLTL formula. We shall say that $\mathcal{S}$, the system of ODEs, meets the specification $P_{\geq r}\psi$ – and this is denoted $\mathcal{S} \models P_{\geq r}\psi$ – iff $P(Models(\psi)) \geq r$, while $\mathcal{S} \models P_{\leq r'}\psi$ iff $P(Models(\psi)) \leq r'$. Here, and in what follows, P is the standard probability measure assigned to members of the σ-algebra generated by the open intervals contained in $INIT$. It is easy to show that $Models(\psi)$ is a member of this σ-algebra for every ψ. The only case that requires an argument is the one for atomic propositions, and here the measurability of the solution functions $\mathbf{X}_{\mathbf{v},\mathbf{w}}(t)$ is crucial. The details can be found in the supplementary material [26].

3.2 Statistical Model Checking of PBLTL Formulas

We now introduce a statistical framework for deciding approximately, but with statistical guarantees, whether the model satisfies a property of the form $P_{\geq r}\psi$. Instead of directly approximating the probability of ψ being satisfied [28], we formulate whether $\mathcal{S} \models P_{\geq r}\psi$, as a hypothesis test. According to [29], the test is posed between the null hypothesis H0 : $p \geq r + \delta$ and the alternative hypothesis H1 : $p \leq r - \delta$, where $p = P(Models(\psi))$. Here, δ is supplied by the user and signifies an indifference region in which one cannot decide on either H0 or H1. The *strength* of the statistical test is decided by parameters α and β which bound the probability of verifying the property as false when it is in fact true (Type-I error) and verifying it as true when it is in fact false (Type-II error) respectively. Thus the verification is carried out approximately but with guaranteed confidence levels and error bounds. The test proceeds by generating a sequence of sample trajectories $\sigma_1, \sigma_2, \ldots$ by randomly sampling an initial state from $INIT$. One assumes a corresponding sequence of Bernoulli random variables $y_1, y_2 \ldots$, where each y_k is assigned the value 1 if $\sigma_k, 0 \models \psi$; otherwise, y_k is assigned the value 0. We next construct a sequential test that lets us decide if the number of samples taken are sufficient, or whether more samples need to be taken to guarantee the chosen test strength. For each $m \geq 1$, after drawing m samples, we compute a quantity q_m as:

$$q_m = \frac{[r-\delta]^{(\sum_{i=1}^{m} y_i)}[1-[r-\delta]]^{(m-\sum_{i=1}^{m} y_i)}}{[r+\delta]^{(\sum_{i=1}^{m} y_i)}[1-[r+\delta]]^{(m-\sum_{i=1}^{m} y_i)}} \quad . \tag{1}$$

The ratio q_m serves as a stopping criterion for the sampling process. Hypothesis H0 is accepted if $q_m \geq \widehat{A}$, and hypothesis H1 is accepted if $q_m \leq \widehat{B}$. If neither is the case then another sample is drawn. The constants $\widehat{A}$ and $\widehat{B}$ are chosen such that it results in a test of strength (α, β). In practice, a good approximation is $\widehat{A} = \frac{1-\beta}{\alpha}$ and $\widehat{B} = \frac{\beta}{1-\alpha}$. A detailed account of our *on-line* model checking algorithm (used to verify each trajectory) can be found in the supplementary material [26].

4 Analysis Methods

Here we present our parameter estimation and sensitivity analysis methods. In doing so, we assume the terminology and notations developed in the previous sections. As a first step, we describe how experimental data can be encoded as a BLTL formula.

Assume, without loss of generality, that $O \subseteq \{x_1, x_2, \ldots, x_k\}$ is the set of variables for which experimental data is available, and which has been allotted as training data to be used for parameter estimation. Assume $\mathcal{T}_i = \{\tau_1^i, \tau_2^i, \ldots, \tau_{T_i}^i\}$ are the time points at which the concentration level of x_i has been measured and reported as $[\ell_t^i, u_t^i]$ for each $t \in \mathcal{T}_i$. The interval $[\ell_t^i, u_t^i]$ is chosen to reflect the noisiness, the limited precision and the cell-population based nature of the experimental data. For each $t \in \mathcal{T}_i$, we define the formula $\psi_i^t = \mathbf{F}^t(i, \ell_t^i, u_t^i)$. Then $\psi_{exp}^i = \bigwedge_{t \in \mathcal{T}_i} \psi_i^t$. We then set $\psi_{exp} = \bigwedge_{i \in O} \psi_{exp}^i$. In case the species x_i has been measured under multiple experimental conditions, the above encoding scheme is extended in the obvious way.

Often qualitative dynamic trends will be available–typically from the literature–for some of the molecular species in the pathway. For instance, we may know that a species shows transient activation, in which its level rises in the early time points, and later falls back to initial levels. Similarly, a species may be known to show oscillatory behavior with certain characteristics. Such information can be described as BLTL formulas that we term to be *trend* formulas. Examples of such formulas can be found in Table1. We let ψ_{qlty} to be the conjunction of all the trend formulas.

Finally, we fix the PBLTL formula $P_{\geq r}(\psi_{exp} \wedge \psi_{qlty})$, where r will capture the confidence level with which we wish to assess the goodness of the fit of the current set of parameters to experimental data and qualitative trends. We also fix an indifference region δ and the strength of the test (α, β). The constants r, δ, α and β are to be fixed by the user. In our application, it will be useful to exploit the fact that both ψ_{exp} and ψ_{qlty} are conjunctions, and hence can be evaluated separately. As shown in [29], one can choose the strength of each of these tests to be $(\frac{\alpha}{J}, \beta)$, where J is the total number of conjuncts in the specification. This will ensure that the overall strength of the test is (α, β). Further, the results of individual statistical tests can be used to compute the objective function associated with the global search strategy to be described below.

4.1 Parameter Estimation Based on PBLTL Specification

We assume $\Theta_u = \{\theta_1, \theta_2, \ldots, \theta_K\}$ is the set of unknown parameters. For convenience we will assume that the other parameter values are known and that their nominal values do not fluctuate across the cell population. We will also assume nominal values for the initial concentrations and the range of their fluctuations of the form $[L_i^{init}, U_i^{init}]$ for each variable x_i. Again, for convenience, we fix a constant δ'' so that if the current estimate of the values of the unknown parameters is $\mathbf{w} \in \prod_{1 \leq j \leq K} [L^j, U^j]$ then this value will fluctuate in the range $[\mathbf{w}(j) - \delta'', \mathbf{w}(j) + \delta'']$. Setting $L_{init,\mathbf{w}}^j = \mathbf{w}(j) - \delta''$ and $U_{init,\mathbf{w}}^j = \mathbf{w}(j) + \delta''$ we define $INIT_{\mathbf{w}} = (\prod_i [L_i^{init}, U_i^{init}]) \times (\prod_j [L_{init,\mathbf{w}}^j U_{init,\mathbf{w}}^j])$. The set of trajectories $BEH_{\mathbf{w}}$ is defined accordingly.

To estimate the quality of $\mathbf{w}$, we run our SMC procedure–using $INIT_{\mathbf{w}}$ instead of $INIT$–to verify $P_{\geq r}(\psi_{exp} \wedge \psi_{qlty})$. Depending on the outcome of this test for the various conjuncts in the specification, we assign a score to $\mathbf{w}$ using an objective function detailed below. We then iterate this scheme for various values of $\mathbf{w}$ generated using a suitable search strategy. The objective function consists of two components, evaluating the contribution from the qualitative properties and the experimental data respectively. It evaluates how many statistical tests carried out with $\mathbf{w}$ resulted in acceptance of the

null hypothesis (desired outcome). For the second component, the tests are evaluated species-wise. The corresponding objective value is then composed as a summation of normalized contribution from each species.

Let J^i_{exp} $(= T_i)$ be the number of conjuncts in ψ^i_{exp}, and J_{qlty} the number of conjuncts in ψ_{qlty}. Let $J^{i,+}_{exp}(\mathbf{w})$ be the number of formulas of the form ψ^t_i (a conjunct in ψ^i_{exp}) such that the statistical test for $P_{\geq r}(\psi^t_i)$ accepts the null hypothesis (that is, $P_{\geq r}(\psi^t_i)$ holds) with the strength $(\frac{\alpha}{J}, \beta)$, where $J = \sum_{i \in O} J^i_{exp} + J_{qlty}$. Similarly, let $J^+_{qlty}(\mathbf{w})$ be the number of conjuncts in ψ_{qlty} of the form $\psi_{\ell,qlty}$ that pass the statistical test $P_{\geq r}(\psi_{\ell,qlty})$ with the strength $(\frac{\alpha}{J}, \beta)$. Then $\mathcal{G}(\mathbf{w})$ is computed via:

$$\mathcal{G}(\mathbf{w}) = J^+_{qlty}(\mathbf{w}) + \sum_{i \in O} \frac{J^{i,+}_{exp}}{J^i_{exp}} \tag{2}$$

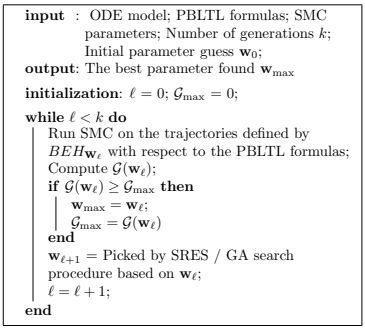
Thus the goodness to fit of $\mathbf{w}$ is measured by how well it agrees with the qualitative properties as well as the number of experimental data points with which there is acceptable agreement. To avoid over-training the model, we do not insist that every qualitative property and every data point must fit well with the dynamics predicted by $\mathbf{w}$.

The search strategy to evolve candidate parameters will use the values $\mathcal{G}(\mathbf{w})$ to traverse the parameter value space. Global search methods such as Genetic Algorithms (GA) [30], and Stochastic Ranking Evolutionary Strategy (SRES) [10] are computationally more intensive than local methods, but are much better at avoiding local minima. The overall structure of our parameter estimation procedure is presented in Algorithm 1. In practice, one usually maintains a *population* of parameter value vectors in each round, and a round is usually called a *generation*. For convenience, we have assumed that each population is a singleton in the description of Algorithm 1. We use the SRES strategy in our work since it is known to perform well in the context of pathway models [9]. The particular choice of search algorithm, however, is orthogonal to our proposed method.

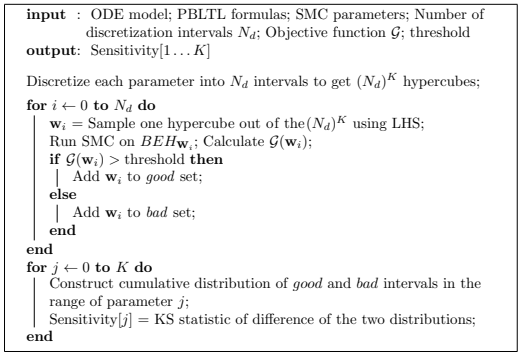
4.2 Sensitivity Analysis Based on PBLTL Specification

As another application of our SMC procedure, we have constructed a property based sensitivity analysis method by coupling our SMC routine with the global sensitivity analysis technique called multi-parametric sensitivity analysis (MPSA) [12]. We assume we have specified a set of properties (encoded as PBLTL formulas), and are interested in knowing which parameters, when changed, affect these properties significantly. The MPSA procedure involves sampling a large number of parameter combinations from their valid ranges. For each sampled combination, one calculates the objective value with respect to the PBLTL properties according to Equation 2. The objective values allow us to assess the extent to which each parameter affects the model's behavior to the given formulas. Intuitively, if the objective value shows strong dependence on the value of a parameter (over its range) then the output is sensitive to that parameter. The MPSA method employs statistical tests to quantify this dependence, which can be directly interpreted as a measure of sensitivity. The sensitivity is based on computing the Kolmogorov-Smirnov (KS) test to compare the two profiles consisting of (a) the cumulative appearance of *good* intervals along the value space of the parameter and

(b) the same for the *bad* intervals. If these profiles differ significantly then the system is more sensitive to this parameter, and the KS test will assign a higher score to this parameter. Our procedure is outlined in Algorithm 2.



Algorithm 1. Parameter estimation



Algorithm 2. Sensitivity analysis

5 Results

We applied our SMC based analysis framework to pathway models taken from the BioModels database [13]. These models have nominal point values for all the rate constants and initial concentrations. We first verified a few properties of the two pathways using SMC. Then, for parameter estimation, we formulated qualitative trends for some species, and generated synthetic experimental data for some other species as follows. We set a $\pm 5\%$ range around the nominal value for the initial concentration of each species and assumed a uniform distribution over the resulting set of initial states. To mimic western blot data, which is cell population based, we averaged 10^4 random trajectories generated by sampling these initial concentration intervals. We then added noise to the data and used a major portion of it for training, and reserved the rest as test data. Finally, we fixed a subset of rate constants to be unknown, and ran our parameter estimation procedure. We let the variability in parameters (δ'') to be 0.5% of the proposed value.

We implemented our method using MATLAB and C++ on a PC with a 3.4Ghz Intel Core i7 processor with 8GB RAM. ODE systems were numerically solved using the SUNDIALS CVODE package [31, 32]. The source code is available at [26]. The code has been optimized to take advantage of the multi-core architecture; all experimental results were run on 8 threads. The parameters used for the statistical model checking algorithm were $r = 0.9$, $\alpha = \beta = \delta = 0.05$ for all our experiments. The choice of these parameters were made so that the probability of satisfaction of the formulas was sufficiently high, and the errors were sufficiently low. The dependence of the performance of the statistical test on the parameters of SMC is well established, we refer the interested reader to [29] for more details. To show the goodness of our estimated parameters (taking into account the variability concerning the initial states and reaction rates),

Table 1. Statistical model checking based verification - The PBLTL Formulas

Pathway	Property	Formula	Result
Thrombin-MLC	sustained activation	$P_{\geq 0.9}(([\text{Phospho MLC} \leq 1]) \wedge (F^{\leq 20}(G^{\leq 20}([\text{Phospho MLC} \geq 3]))))$	*false*
Thrombin-MLC	transient activation	$P_{\geq 0.9}(([\text{Phospho MLC} \leq 1]) \wedge F^{\leq 20}(([\text{Phospho MLC} \geq 3]) \wedge F^{\leq 20}(G^{\leq 20}((\text{Phospho MLC} \leq 1)))))$	*true*
Segmentation clock	oscillations	$P_{\geq 0.9}(([\text{Lunatic fringe mRNA} \leq 0.4]) \wedge (F^{\leq 40}([\text{Lunatic fringe mRNA} \geq 2.2] \wedge F^{\leq 40}([\text{Lunatic fringe mRNA} \leq 0.4] \wedge F^{\leq 40}([\text{Lunatic fringe mRNA} \geq 2.2] \wedge F^{\leq 40}([\text{Lunatic fringe mRNA} \leq 0.4])))))))$	*true*

we generated 1000 trajectories and plotted these to show that the estimated parameters result in a good fit to the data. In each case, experimental data is plotted along with the tolerance interval used in constructing the specification.

For the experiments reported in this section, we used an SRES based global strategy to guide the search. Here we present only the highlights of our experimental results. Many further details including the results obtained using a Genetic Algorithm based search can be found in the supplementary material [26].

5.1 The Case Studies

The segmentation clock network An oscillating segmentation clock governs the segmentation pattern of the spine in developing vertebrate embryos. It couples signaling pathways of FGF, Notch and Wnt, whose periodic behaviors are produced by negative feedback loops. The ODE model consists of 16 differential equations and 75 kinetic rate parameters. Simulation time (T) was fixed at 200 minutes assumed to be observable at 40 equally spaced time points.

The thrombin-dependent MLC phosphorylation pathway Endothelial cells form a dynamic barrier between blood/lymph and the underlying connective tissue, and their contraction is crucial to physiological and pathological processes. Agonists such as thrombin play an important role in the contraction function through phosphorylation of MLC, while Rho-kinase is crucial for the sustained contraction of endothelial cells. The pathway model with 105 differential equations and 197 kinetic parameters is considerably large. Simulation time was fixed at 1000 seconds assumed to be observable at 20 equally spaced time points.

5.2 Statistical Model Checking Based Verification

First, we used our SMC framework to verify pathway properties expressed in PBLTL. We used the nominal models (all rate parameter values known, taken from the BioModels database) to verify if they conformed to properties expressed in our logic with high probability. We describe a few such properties along with their BLTL formulas and the result of verification in Table1. For instance, for the MLC phosphorylation pathway, it is known experimentally that the concentration of phosphorylated MLC starts at a low level, and then reaches a high steady state value. Our SMC method shows that the

nominal model does not satisfy the property, instead, phosphorylated MLC exhibits a transient profile. This discrepancy has been studied in [33], and attributed to missing components and interactions in the proposed model.

5.3 Parameter Estimation

For the segmentation clock pathway, we assumed 39 of the rate parameters as unknown. We used a combination of dynamic trends and quantitative experimental data. Specifically, we synthesized population based experimental time series data for Axin2 mRNA measured at 14 time points up to 165 minutes. For 5 other species {Notch protein, nuclear NicD, Lunatic fringe mRNA, active ERK and Dusp6 mRNA}, we encoded the dynamic trends as properties in our logic. The dynamic trend of 2 species (cytosolic NicD and Dusp6 protein) were used as test data. Parameter estimation was done with a population of 200 per generation and for 300 generations. The time taken by SRES based search was 2.3 hours. Figure 1 shows simulation profiles with the estimated parameters. Figure 1(a) shows that the model fits training data consisting of the experimental data of Axin2 mRNA and qualitative trends for 3 other species. Figure 1(b) shows dynamic trends of cytosolic NicD used for testing. The simulated time profiles fit the specified test properties (see [26]).

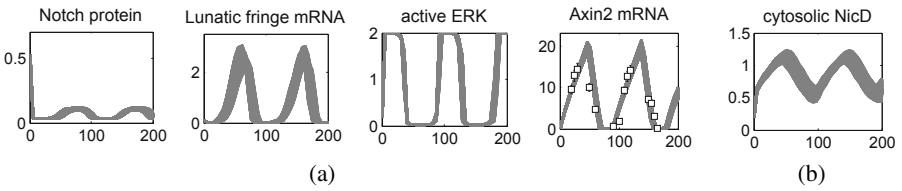


Fig. 1. Parameter estimation results of the segmentation clock pathway. (a) Training data including the experimental data for Axin2 mRNA and the dynamic trends for 3 species), and (b) the test data for one of the species.

To illustrate the scalability of our approach, for the thrombin pathway, we assumed 100 of the kinetic parameters to be unknown. We synthesized population based experimental time series data for 10 species including RGS_2, Rho.GTP, PKC.DAG, MLC_2,

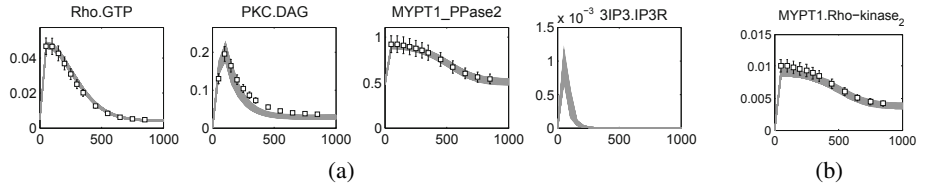


Fig. 2. Parameter estimation results of Thrombin-dependent MLC phosphorylation pathway. (a) Training data, including experimental data of 3 species and dynamic trends of one species, and (b) the test data for one of the species.

CPI-17, Ca-super-2-plus, p115RhoGEF-GTP-alpha, MYPT1-PPase, Rho-kinase.MLC, MYPT1.Rho-kinase$_2$. For thrombinR-active and 3IP3.IP3R we assumed that only the dynamic trend is known. The data of Rho-kinase.MLC and MYPT1.Rho-kinase$_2$ were reserved as test data to evaluate the quality of our parameter estimates, while the data of all other species was used to calibrate the model. Parameter estimation was done with a population of 100 per generation and for 1000 generations. The time taken by SRES based search was 48.8 hours. Figure 2 shows the fit to data of the simulation profiles with the best predicted parameter values for both the training data (Figure 2(a)) and the test data (Figure 2(b)).

5.4 Property Based Sensitivity Analysis

Here we report results just for the segmentation clock pathway (due to the space constraints). We evaluated the sensitivity of parameters against all properties used for parameter estimation. The results are shown in Figure 3(a). It can be seen that the most sensitive parameters are *ksDusp*, *kcDusp*, *VMsMDusp*, *VMdMDusp*, *VMaX*, *VMdX*. This also indicates that the reactions involving Dusp6 degradation and transcription affect the overall dynamics most. Since all these parameters belong to the FGF pathway, we hypothesize that FGF pathway is the most crucial component that drives the behavior of the system.

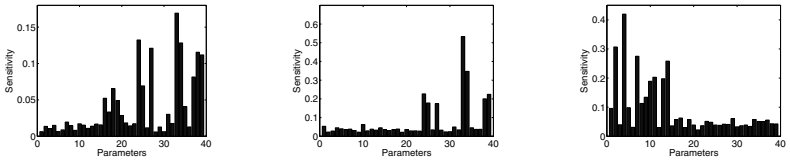


Fig. 3. Sensitivity analysis results. (a-c) Parameter sensitivities of the segmentation clock pathway with respect to (a) all properties, (b) Dusp6mRNA profile, and (c) nuclear nicD profile.

We next searched for parameters affecting the oscillatory property of *Dusp6 mRNA* alone. We found that the same set of parameters as above are the most crucial (see Figure 3(b)). However, when evaluating the oscillatory property of nuclear NicD (Figure 3(c)), we find that the parameters *vsN*, *kt1*, *VdNan* are the most significant. This suggests that although the Notch synthesis (*vsN*), and nuclear NicD transportation (*kt1*) and degradation (*VdNan*) do not significantly affect the overall dynamics, they play a dominant role in segmentation patterning.

6 Conclusion

We have proposed an SMC based approach for studying ODE based bio-pathway models. We have used the temporal logic BLTL to encode both quantitative experimental data and qualitative properties of pathway dynamics. To cater for variability among

cells, we assume a uniform distribution over a set of initial states and kinetic rate constants–and impose a reasonable continuity restriction–and show how the probability of the property being met by the behavior of the model can be assessed using an SMC procedure. By combining this method with a global search strategy, we arrive at a parameter estimation procedure as well as a sensitivity analysis technique.

We have demonstrated the applicability of our method with the help of two ODE based bio-pathway models: the segmentation clock network and the thrombin-dependent MLC phosphorylation pathway. Our method successfully obtained good parameter estimates using noisy cell-population data and qualitative knowledge. The results show that our method scales well and can cope with large biological networks. We also show results for performing property based sensitivity analysis, and thereby gain interesting insights about the pathway dynamics that would be difficult to obtain using conventional approaches.

Our parameter estimation method is a generic one and has the potential to be applied to model classes such as continuous time Markov chain (CTMC) models and stochastic differential equation (SDE) models [3]. We plan to explore this in our future work. Another interesting direction will be to develop a GPU-based implementation of our method to exploit the inherent massive parallelism in generating trajectories through numerical integration. In this connection, the platform-aware implementation of a related systems biology application presented in [17] promises to offer helpful pointers.

References

1. De Ferrari, G.V., Inestrosa, N.C.: Wnt signaling function in Alzheimer's disease. Brain Res. Rev. 33, 1–12 (2000)
2. Aldridge, B.B., Burke, J.M., Lauffenburger, D.A., Sorger, P.K.: Physicochemical modelling of cell signalling pathways. Nat. Cell Biol. 8(11), 1195–1203 (2006)
3. Wilkinson, D.: Stochastic modelling for systems biology. CRC Press (2011)
4. Klipp, E., Herwig, R., Kowald, A., Wierling, C., Lehrach, H.: Systems biology in practice: concepts, implementation and application. Wiley-VCH, Weinheim (2005)
5. Spencer, S., Gaudet, S., Albeck, J., Burke, J., Sorger, P.: Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis. Nature 459(7245), 428–432 (2009)
6. Snijder, B., Pelkmans, L.: Origins of regulated cell-to-cell variability. Nature Reviews Molecular Cell Biology 12(2), 119–125 (2011)
7. Weiße, A., Middleton, R., Huisinga, W.: Quantifying uncertainty, variability and likelihood for ordinary differential equation models. BMC Systems Biology 4(1), 144 (2010)
8. Younes, H.L.S., Kwiatkowska, M., Norman, G., Parker, D.: Numerical vs. statistical probabilistic model checking. International Journal on Software Tools for Technology Transfer 8, 216–228 (2006)
9. Moles, C.G., Mendes, P., Banga, J.R.: Parameter estimation in biochemical pathways: A comparison of global optimization methods. Genome Res. 13(11), 2467–2474 (2003)
10. Runarsson, T., Yao, X.: Stochastic ranking for constrained evolutionary optimization. IEEE T. Evolut. Comput. 4, 284–294 (2000)
11. Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., Tarantola, S.: Global sensitivity analysis: the primer. Wiley-Interscience (2008)

12. Cho, K.H., Shin, S.Y., Kolch, W., Wolkenhauer, O.: Experimental design in systems biology, based on parameter sensitivity analysis using a Monte Carlo method: A case study for the TNFα-mediated NF-κB signal transduction pathway. Simulation 79(12), 726–739 (2003)
13. Le Novere, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., Hucka, M.: BioModels Database: A free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. 34, D689–D691 (2006)
14. Jha, S.K., Clarke, E.M., Langmead, C.J., Legay, A., Platzer, A., Zuliani, P.: A bayesian approach to model checking biological systems. In: Degano, P., Gorrieri, R. (eds.) CMSB 2009. LNCS, vol. 5688, pp. 218–234. Springer, Heidelberg (2009)
15. Heath, J., Kwiatkowska, M., Norman, G., Parker, D., Tymchyshyn, O.: Probabilistic model checking of complex biological pathways. Theor. Comput. Sci. 391(3), 239–257 (2008)
16. Li, C., Nagasaki, M., Koh, C.H., Miyano, S.: Online model checking approach based parameter estimation to a neuronal fate decision simulation model in Caenorhabditis elegans with hybrid functional Petri net with extension. Mol. Biosyst. 7(5), 1576–1592 (2011)
17. Liu, B., Hagiescu, A., Palaniappan, S.K., Chattopadhyay, B., Cui, Z., Wong, W., Thiagarajan, P.S.: Approximate probabilistic analysis of biopathway dynamics. Bioinformatics 28(11), 1508–1516 (2012)
18. Donaldson, R., Gilbert, D.: A monte carlo model checker for probabilistic ltl with numerical constraints. University of Glasgow, Dep. of CS, Tech. Rep. (2008)
19. Donaldson, R., Gilbert, D.: A model checking approach to the parameter estimation of biochemical pathways. In: Heiner, M., Uhrmacher, A.M. (eds.) CMSB 2008. LNCS (LNBI), vol. 5307, pp. 269–287. Springer, Heidelberg (2008)
20. Clarke, E.M., Faeder, J.R., Langmead, C.J., Harris, L.A., Jha, S.K., Legay, A.: Statistical model checking in *BioLab*: Applications to the automated analysis of T-cell receptor signaling pathway. In: Heiner, M., Uhrmacher, A.M. (eds.) CMSB 2008. LNCS (LNBI), vol. 5307, pp. 231–250. Springer, Heidelberg (2008)
21. Maler, O.: On under-determined dynamical systems. In: Proceedings of the Ninth ACM International Conference on Embedded Software, pp. 89–96. ACM (2011)
22. Calzone, L., Chabrier-Rivier, N., Fages, F., Soliman, S.: Machine learning biochemical networks from temporal logic properties. T. Comput. Syst. Biol. VI, 68–94 (2006)
23. Rizk, A., Batt, G., Fages, F., Soliman, S.: On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology. In: Heiner, M., Uhrmacher, A.M. (eds.) CMSB 2008. LNCS (LNBI), vol. 5307, pp. 251–268. Springer, Heidelberg (2008)
24. Batt, G., Page, M., Cantone, I., Goessler, G., Monteiro, P., de Jong, H.: Efficient parameter search for qualitative models of regulatory networks using symbolic model checking. Bioinformatics 26(18), i603–i610 (2010)
25. Barnat, J., Brim, L., Krejci, A., Streck, A., Safranek, D., Vejnar, M., Vejpustek, T.: On parameter synthesis by parallel model checking. IEEE/ACM T. Comput. Bi. 9(3), 693–705 (2012)
26. Supplementary information and source code, `http://www.comp.nus.edu.sg/~rpsysbio/SMC/`
27. Hirsch, M., Smale, S., Devaney, R.: Differential equations, dynamical systems, and an introduction to chaos. Academic Press (2012)
28. Hérault, T., Lassaigne, R., Magniette, F., Peyronnet, S.: Approximate probabilistic model checking. In: Steffen, B., Levi, G. (eds.) VMCAI 2004. LNCS, vol. 2937, pp. 73–84. Springer, Heidelberg (2004)
29. Younes, H.L.S., Simmons, R.G.: Statistical probabilistic model checking with a focus on time-bounded properties. Inform. Comput. 204, 1368–1409 (2006)
30. Goldberg, D.: Genetic algorithms in search, optimization, and machine learning. Addison-Wesley (1989)

31. Hindmarsh, A., Brown, P., Grant, K., Lee, S., Serban, R., Shumaker, D., Woodward, C.: SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. ACM T. Math. Software 31(3), 363–396 (2005)
32. Vanlier, J., Tiemann, C., Hilbers, P., van Riel, N.: An integrated strategy for prediction uncertainty analysis. Bioinformatics 28(8), 1130–1135 (2012)
33. Maedo, A., Ozaki, Y., Sivakumaran, S., Akiyama, T., Urakubo, H., Usami, A., Sato, M., Kaibuchi, K., Kuroda, S.: Ca^{2+}-independent phospholipase A2-dependent sustained Rho-kinase activation exhibits all-or-none response. Genes Cells 11, 1071–1083 (2006)