Parameter discovery in stochastic biological models using simulated annealing and statistical model checking

Faraz Hussain* and Sumit K. Jha

Computer Science Department,
University of Central Florida,
Orlando, FL 32816, USA
Email: fhussain@cs.ucf.edu
Email: jha@cs.ucf.edu
*Corresponding author

Susmit Jha

Intel Strategic CAD Labs,
Portland, OR 9712, USA
Email: susmit.jha@intel.com

Christopher J. Langmead

Lane Center for Computational Biology,
Carnegie Mellon University,
Pittsburgh, PA 15213, USA
and
Computer Science Department,
Carnegie Mellon University,
Pittsburgh, PA 15213, USA
Email: cjl@cs.cmu.edu

Abstract: Stochastic models are increasingly used to study the behaviour of biochemical systems. While the structure of such models is often readily available from first principles, unknown quantitative features of the model are incorporated into the model as parameters. Algorithmic discovery of parameter values from experimentally observed facts remains a challenge for the computational systems biology community. We present a new parameter discovery algorithm that uses simulated annealing, sequential hypothesis testing, and statistical model checking to learn the parameters in a stochastic model. We apply our technique to a model of glucose and insulin metabolism used for in-silico validation of artificial pancreata and demonstrate its effectiveness by developing parallel CUDA-based implementation for parameter synthesis in this model.

Keywords: parameter discovery; biochemical systems; computational systems biology; behavioural specifications; statistical hypothesis testing; parameter synthesis; probabilistic verification; SPRT; temporal logic; statistical model checking; stochastic models; machine learning; bioinformatics; CUDA; glucose-insulin model; biomedical devices; cyber-physical systems; CPS.

Copyright © 2014 Inderscience Enterprises Ltd.

Reference to this paper should be made as follows: Hussain, F., Jha, S.K., Jha, S. and Langmead, C.J. (2014) 'Parameter discovery in stochastic biological models using simulated annealing and statistical model checking', *Int. J. Bioinformatics Research and Applications*, Vol. 10, Nos. 4/5, pp.519–539.

Biographical notes: Faraz Hussain is a doctoral candidate in the EECS department at the University of Central Florida. His primary research interest is in developing formal methods techniques for analysing probabilistic computational models.

Sumit K. Jha is an Assistant Professor in the EECS department at the University of Central Florida (UCF). He has a PhD from the Carnegie Mellon University and has been on the UCF faculty since 2010.

Susmit Jha is a Research Scientist at Strategic CAD Labs at Intel. He earned his PhD from the University of California at Berkeley and his undergraduate degree from the Indian Institute of Technology, Kharagpur.

Christopher J. Langmead is an Associate Professor in the School of Computer Science at Carnegie Mellon University. He has been on the Carnegie Mellon faculty since 2004.

This paper is a revised and expanded version of a paper entitled 'Parameter discovery for stochastic biological models against temporal behavioral specifications using an SPRT based metric for simulated annealing' presented at the '2nd IEEE International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)', Las Vegas, NV, USA, 23–25 February 2012.

1 Introduction

The computational modelling of the precise dynamics of biochemical systems involves the modelling and analysis of complex Continuous-Time Markov Chain (CTMC) models. Such detailed biochemical models are often reduced to readily analysable and succinct models like Stochastic Differential Equations (SDEs) and Discrete-Time Markov Chains (DTMCs). Recent interest in the development of Computer-Aided Design (CAD) techniques for biomedical devices has led to the development of heterogeneous models that include a stochastic biochemical model coupled with an external deterministic controller. The design and formal verification of such biomedical devices require the ability to model, analyse and verify complex stochastic models of cyber-physical systems.

The structure of such stochastic models can often be determined from the first principles and a survey of existing biochemical literature. However, several quantitative features of such models cannot easily be obtained from the literature or inferred from experimental data. The discovery of such quantitative parameters of computational models from observed experimental facts remains the subject of ongoing research in computational systems biology.

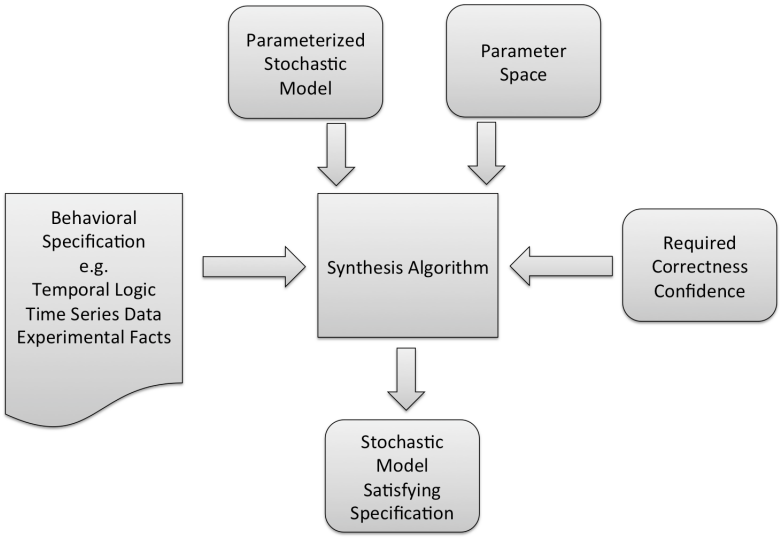
In this paper, we make two contributions to the discovery of parameters of stochastic models of biochemical systems and biomedical devices:

1 We present an algorithm for discovering parameters of stochastic computational models from experimental observations that uses a combination of simulated

annealing and the Sequential Probability Ratio Test (SPRT) to reduce the number of samples required for discovering the correct parameter values. Figure 1 summarises the problem of parameter discovery in stochastic models from experimental facts and time-series data. Our algorithm uses the fact that during the simulated annealing-based parameter exploration process for a parameterised model, *if the SPRT rejects the null hypothesis, the expected number of samples required is proportional to the probability with which the model satisfies the given specification*.

2 We apply our proposed algorithm to a complex model of glucose–insulin metabolism for testing artificial pancreata. We show that the model is capable of reproducing pancreatic-induced diabetes by a suitable reparameterisation of three parameters related to the pancreas in the model.

Figure 1 Problem definition. Given the behavioural specifications about the system being modelled, the parameterised stochastic models, the parameter space and a correctness confidence, our algorithm seeks to discover the value of the parameters that enables the given model to satisfy the given probabilistic behavioural specifications



2 Related work

The field of stochastic modelling has emerged to overcome the inherent limitations of deterministic modelling (Astrom, 2006; Busch et al., 2008; Kulkarni, 1995; Maybeck, 1979). Such a modelling approach permits randomness in the behaviour of the system, and often uses sampling-based statistical inference to find the probability distribution of potential outcomes. Research in systems biology has greatly benefited from the existing literature in stochastic modelling, and has also inspired new scientific expeditions into the realm of stochastic modelling, analysis and verification. Indeed, *in silico* modelling has been particularly useful in developing a systems view of biology. Models for whole-cell analysis (Roberts et al., 2011), drug discovery for tuberculosis (Arenas et al., 2011),

control of type-1 diabetes (Kovatchev et al., 2009), sequence analysis (Higo et al., 1998), enzyme interaction with drugs (de Graaf et al., 2005) and prediction of blood-secretory proteins (Liu et al., 2010) are some of the success stories in computational systems biology. However, several components of these models are not available from the first principles or using experimental data. In such a scenario, model designers include missing information as parameters of the model. The number of parameters increases with the size and complexity of the model, and it becomes increasingly difficult to determine the value of these parameters for large and detailed models of biological systems.

The discovery of parameters for stochastic models has been carried out using various approaches (Amin et al., 2010; Caragea et al., 2010; Saul and Filkov, 2007; Frehse et al., 2008; Jha, 2008; Jha et al., 2007; Kalyanaraman et al., 2011). Different estimation techniques have been adopted by researchers for finding parameters of stochastic biochemical reactions (Gillespie, 1977; Reinker et al., 2006; Salis and Kaznessis, 2005; Turner et al., 2004). Estimators used for deterministic models have also been extended to stochastic models (Wilkinson, 2006, 2011). Considerable research has been directed towards the use of statistical hypothesis testing for verification of stochastic models (Jha and Langmead, 2011; Jha et al., 2011; Younes and Simmons, 2002; Younes and Simmons, 2006), including those arising in systems biology. Researchers have also adopted Bayesian frameworks (Jha et al., 2012; Nikolova et al., 2007) for parameter identification in stochastic models such as stochastic gradient descent (Bottou, 2003), simulated annealing (Gonzalez et al., 2007) and evolutionary computing (Busch et al., 2008). Recently, researchers have also been successful in synthesising parameters in real-time systems such as parametric timed automata (André and Soulat, 2013).

3 Background

In this section, we discuss the various classes of stochastic models that can benefit from our parameter discovery algorithm. We also present a specification formalism for representing facts observed from experimental data that describes the properties that the stochastic model with the synthesised parameters must satisfy. Finally, we briefly survey the literature on the SPRT and its relationship to statistical estimation.

3.1 Stochastic models

Our proposed algorithm can be applied to several classes of parameterised stochastic models, including CTMCs, SDEs, jump diffusion processes and heterogeneous models consisting of stochastic processes interacting with deterministic models like Ordinary Differential Equations (ODEs).

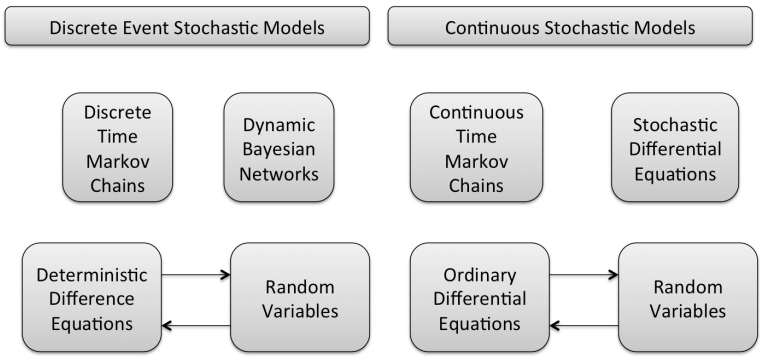
Figure 2 illustrates the various types of stochastic models whose parameters can be discovered using our algorithm. DTMCs are models whose state-space can be indexed by a countable set. Each transition between states of a DTMC is associated with a finite probability. Dynamic Bayesian Networks (DBNs) are representations of probabilistic models amenable to analysis using statistical inference and machine learning. All of these models are discrete event stochastic systems, where the behaviour of the system over a period of time can be completely described by a finite or countable numerical sequence of values assigned to system variables. Another interesting class of systems is formed by

a set of deterministic differential equations interacting with a set of random variables or stochastic processes. CTMCs and SDEs represent stochastic systems that evolve continuously in time. Two kinds of stochastic models that are often used to model biochemical and cyber-physical systems are of special interest to us and merit deeper discussion:

Continuous-time Markov chains: Biochemical systems consisting of a set of biochemical reactions in a homogeneous well-mixed volume can be precisely modelled using CTMCs. CTMC models are often simulated using Gillespie's (1977) stochastic simulation algorithm. However, the values of the rate constants or kinetic parameters that determine the transition probability in CTMC models are very difficult to obtain from the first principles. In many cases, they are also difficult to measure in an *in vivo* setting. Our technique can be used to discover kinetic parameters for CTMC models of biochemical systems from data gathered by empirical observations.

ODEs and random variables: A biomedical cyber-physical system is often modelled using a system of deterministic ODEs interacting with random variables to represent the biochemical system and the external controller (DallaMan et al., 2007). However, several parameters and variables in such models are either unknown in biological literature or they vary substantially from one individual to another in a population. In both these cases, such parameters and variables are modelled as random variables or stochastic processes. The resulting complex cyber-physical model is a continuous time stochastic system. The algorithm presented in this paper can also be used to discover parameters of models of such complex cyber-physical biomedical systems.

Figure 2 Stochastic models. Our algorithm is applicable to both discrete and continuous time stochastic models. CTMCs are particularly important for studying biochemical systems, while ODEs interacting with random variables naturally model cyber-physical biomedical systems



3.2 Specifications

The behavioural specification of stochastic biochemical and biomedical models requires complex temporal reasoning. We use Probabilistic Bounded Linear Temporal Logic (PBLTL) to specify the behaviour of such systems. Temporal logics can be used to formally describe the use of tense and various forms of causality in natural languages.

We first define the syntax and semantics of Bounded Linear Temporal Logic (Finkbeiner and Sipma, 2001; Lamport, 1982; Pnueli, 1977). A bounded linear temporal logic (BLTL) specification is a set of predicates connected using Boolean and temporal operators. The syntax of the logic is given by the following grammar:

$$\phi ::= x \le v | x \ge v | (\phi_1 \vee \phi_2) | (\phi_1 \wedge \phi_2) | \neg \phi_1 | (\phi_1 \mathbf{U}^t \phi_2)$$

where $\mathcal{V}$ is a set of discrete-valued variables, $x \in \mathcal{V}$, $v \in \mathbb{R}$ and $t \in \mathbb{R}_{\ge 0}$ denotes time. We can define additional temporal operators such as $\mathbf{F}^t \psi = \mathbf{True}\mathbf{U}^t \psi$ and or $\mathbf{G}^t \psi = \neg \mathbf{F}^t \neg \psi$. The formula $\mathbf{F}^t \psi$ implies that ψ holds sometime within t time units. The formula $\mathbf{G}^t \psi$ implies that ψ holds at all moments for the next t time units into the future. The fact that a path $\wp$ satisfies the BLTL property ϕ is denoted by $\wp \vDash \phi$. Let $\wp = (y_0, \tau_0), (y_1, \tau_1), \ldots$ be an execution of the model along states $y_0, y_1, \ldots$ with durations $\tau_0, \tau_1, \ldots \in \mathbb{R}$. We denote the path starting at state i by $\wp^i$ (in particular, $\wp^0$ denotes the original execution $\wp$). The value of the state variable x in $\wp$ at the state i is denoted by $V(\wp, i, x)$.

Specifications about biochemical and biomedical systems are often probabilistic in nature. For example, it may be required that a respirator will not allow the oxygen level in the blood to fall below 90% of its average value for more than 4 seconds with 99.9999% probability. Such behaviours are naturally expressed as probabilistic specifications. If ϕ is a BLTL specification, $Pr_{\ge \rho}(\phi)$ is a probabilistic bounded linear temporal logic (PBLTL) specification. The model $\mathcal{M}$ is said to satisfy the PBLTL specification $Pr_{\ge \rho}(\phi)$ if at least ρ fraction of independently drawn random behaviours observed from the model satisfies the BLTL specification ϕ.

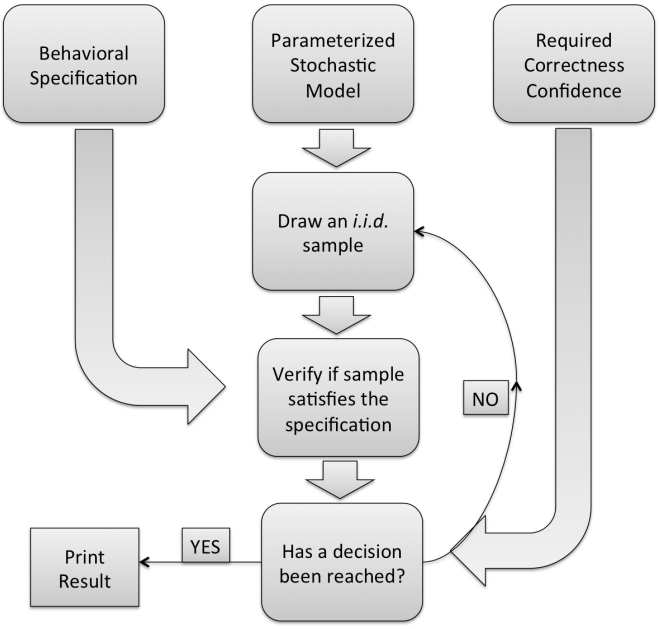
3.3 Sequential Probability Ratio Test (SPRT)

Given a stochastic model $\mathcal{M}$ and a specification ϕ, let p be the unknown probability with which the model satisfies the specification. Let ρ be the probability threshold with which the model is expected to satisfy a specification ϕ (i.e. $\mathcal{M} \vDash Pr_{\ge \rho}(\phi)$). The SPRT (Wald, 1945; Younes and Simmons, 2006) is used to decide which of the following hypotheses is true: (i) null hypothesis H_0: $p \ge p_0$, (ii) alternate hypothesis H_1: $p \le p_1$. The SPRT uses a sequential sampling procedure (Wald, 1945; Younes and Simmons, 2002; Younes and Simmons, 2006), where the number of observations is determined by the outcome of the observations themselves. The goal of a sequential testing procedure is to reduce the number of samples required to reject the null or the alternate hypothesis. The number p_0 is chosen as $\rho + \varepsilon$ and p_1 is chosen as $\rho - \varepsilon$, where ρ is the probability threshold with which the model is expected to satisfy the specification and 2ε is the value of the indifference region (Younes and Simmons, 2006). Also, note that $0 \le p_1 < p_0 < 1$.

The SPRT is a sequential sampling test in which each observations can lead to three outcomes: (i) the null hypothesis is accepted; (ii) the alternate hypothesis is accepted; (iii) additional observations are needed to accept either hypothesis and hence, the procedure continues. This process of generating additional samples continues until a decision is made to accept one of the hypotheses. Our SPRT-based parameter synthesis technique is illustrated in Figure 3. The result of the SPRT procedure is correct in a probabilistic

sense. There can be two kinds of errors in the answer produced by the SPRT procedure: Type I and Type II. A Type I error is the condition of rejecting the null hypothesis H_0 when it is actually true. Accepting the null hypothesis H_0 when the alternate hypothesis H_1 is true results in a Type II error. The SPRT bounds the probability of making Type I and Type II errors during hypothesis testing by two constants α and β that are used to parameterise the test, denoted $S(A,B)$. Note that A and B are calculated using α and β as described in Wald (1945).

Figure 3 Our SPRT-based parameter discovery procedure. Given a probabilistic behavioural specification and a parameterised stochastic model, the SPRT procedure decides if the model satisfies the specification



3.3.1 Optimality for simple hypotheses

Our motivation for using SPRT-based hypothesis testing is to reduce the number of points in the parameter space that needs to be explored in order to discover parameters in the stochastic model so as to synthesise a model that meets the given probabilistic behavioural specifications.

3.3.2 Comparison with statistical estimation

In statistical inference and decision theory, judgement about a population is often based on sampling properties of a randomly chosen subset of the population. Such an inference can be accomplished using two different but interrelated approaches: (i) statistical estimation theory and (ii) hypothesis testing.

Estimation theory makes use of empirical random data to suggest an estimate of a parameter value. Standard estimators for decision theory include Maximum Likelihood (MLE), Bayes estimator, Kalman filter and Markov chain Monte Carlo (MCMC) procedures.

In *statistical hypothesis testing*, an assumption is made about the distribution of the parameter being studied. The assumption or hypothesis is approved or rejected based on a probabilistic belief computed by observing a series of data points. Some of the popular methods for hypothesis testing are Bayesian tests, likelihood ratio tests, sequential tests and union-intersection and intersection-union tests.

Sequential analysis (Ghosh and Sen, 1991; Wald, 1947) is a technique for statistical hypothesis testing where the number of samples required to reach a conclusion regarding a claim is not fixed, but is itself a random variable. The SPRT that we have used in our algorithmic parameter discovery algorithm is a sequential analysis technique (Wald, 1945).

4 Our approach

We present a new parameter discovery algorithm for stochastic models that brings together the SPRT and the simulated annealing procedure. The algorithm uses fewer samples than a traditional approach based on statistical estimation and simulated annealing (Gonzalez et al., 2007). The correctness proof of our algorithm is based on the fact that the number of samples used by the SPRT algorithm is related to the fitness value at a parameter point during the simulated annealing-based exploration of the parameter space. We later present the formal correctness proof of our claim.

4.1 Algorithm

The parameter discovery problem is presented in Figure 1. There are four inputs to our algorithm:

1 *Behavioural specification*: the user provides a probabilistic specification about the behaviour of the stochastic biological system. The specification may be provided as PBLTL formula, or it may be available as extreme-scale time series data collected by observing a number of experiments. Various classes of experimental data and observed facts can be translated into variants of temporal logic.

2 *Parameterised stochastic model*: the algorithm discovers the parameter values for parameterised stochastic models. In this setting, a parameter is a model variable whose value does not evolve during the execution of the model. As discussed earlier, our approach can be applied to a number of stochastic models, including those useful for biochemical and biomedical applications.

3 *Parameter space*: our algorithm also requires that the possible space of all possible parameter values for the model be defined. A bounded (but not necessarily finite) parameter space is a requirement to ensure that the algorithm eventually terminates.

4 Type-1 and type-2 error bounds for the SPRT.

Our parameter discovery algorithm builds on the classical simulated annealing approach for parameter discovery (see Figure 4). Simulated annealing (Bertsimas and Tsitsiklis, 1993) is a stochastic optimisation method for finding the global minimum of a system

that possibly has several local minima. It is a probabilistic version of the gradient descent algorithm where, instead of moving along the gradient, the algorithm decides the optimisation steps stochastically. A global minimum of a system is located by moving stochastically through the function space, based on the value of an objective function. This process continues until the algorithm converges to one of the global minima.

In most real optimisation problems, an estimation of the objective function is made using statistical estimation and other approaches (Maybeck, 1979). Several heuristic estimation methods have been used for computing the probability (objective function) of a model satisfying a specification during simulated annealing. However, these estimation-based methods require a large number of samples and cannot be practically used for parameter discovery.

Figure 4 illustrates the classical simulated annealing algorithm. Given a parameter space Ω, the algorithm seeks to find the parameter value ω^* such that $E(\omega^*)$ represents a global minimum. The algorithm always accepts a better value for ω^* and also accepts a worse value with probability $e^{-(E(\omega')-E(\omega))/t}$. The probability of accepting a worse parameter value gets smaller with time.

Figure 4 Classic simulated annealing algorithm. The figure shows the classic algorithm that uses the exact probability calculation for simulated annealing

Require: Parameter space Ω, Objective function $E : \Omega \to \mathbb{R}$,
Temperature Cooling Schedule $\mathrm{T} : \mathbb{N} \to (0, \infty)$,
Starting Temperature t, Stopping Temperature t_0.
ω = Pick a random point in Ω.
$E(\omega) = \infty$
while $t \geq t_0$ **do**
  Select a neighbor ω' randomly
  if $E(\omega') \leq E(\omega)$ **then**
    $\omega \leftarrow \omega'$
  end if
  if $E(\omega') > E(\omega)$ **then**
    $\omega \leftarrow \omega'$ with probability $e^{-(E(\omega') - E(\omega))/t}$
  end if
  $t = T(t)$
end while
Ensure: Algorithm stops at ω^* that minimizes $E(\omega)$.

The above-mentioned algorithm reaches an optimal value when $E(\omega^*)$ and $E(\omega)$ are estimated correctly. The estimation procedures are tedious and challenging for stochastic systems due to the presence of randomness and the consequent necessity of observing millions of model simulations before the value of the fitness function at a given parameter point can be adequately estimated.

However, for our parameter discovery technique, such a precise estimation of the probability values is not useful for most of the parameter space being explored. In reference to statistical model checking, Younes and Simmons (2006) also note that exact calculations can be replaced by asking weaker questions and thus making the verification procedure more efficient. We propose the use of the SPRT for deciding if a parameter value is interesting *without* explicitly estimating the value of the fitness function at this

parameter point. It also enables us to construct a computationally inexpensive objective function for simulated annealing. This approach helps in efficiently comparing the various states of the system towards obtaining a global minimum.

Figure 5 uses the number of simulations needed by the SPRT procedure as a metric for guiding the simulated annealing-based parameter synthesis algorithm. We show that such an approach is correct. In particular, we establish a relationship between the number of samples used by the SPRT procedure and the probability with which a parameterised model satisfies a behavioural specification.

Claim 1: *Let p be the probability with which the model $\mathcal{M}$ satisfies the specification ϕ. Given the null hypothesis H_0: $p \geq p_0$ and the alternate hypothesis H_1: $p \leq p_1$, and error thresholds α and β, if SPRT(A,B) rejects the null hypothesis, the average number of samples observed by our SPRT-based algorithm increases as p increases. Note that A,B are calculated from the α, β as given in Wald (1945).*

Figure 5 SPRT-based simulated annealing algorithm for parameter discovery. Our new algorithm for parameter discovery that combines the SPRT with simulated annealing

Require: Parameter space Ω, Probabilistic Behavioral Specification $Pr_{\geq\rho}(\phi)$
Require: *Temperature Cooling Schedule* $\mathrm{T} : \mathbb{N} \rightarrow (0, \infty)$
Require: Starting Temperature t, Stopping Temperature t_0
Require: Bounds on Type-I/II errors: α, β.
Ensure: Algorithm synthesizes ω such that $\mathcal{M}(\omega) \models Pr_{\geq\rho}(\phi)$ or prints "No parameter found."
  $H_0 : \mathcal{M}(\omega) \models Pr_{\geq\rho}(\phi)$
  $H_1 : \mathcal{M}(\omega) \not\models Pr_{\geq\rho}(\phi)$
  $A = \frac{1-\beta}{\alpha}$
  $B = \frac{\beta}{1-\alpha}$
  $\omega =$ Pick a random point in Ω.
  if SPRT(A, B) at $\mathcal{M}(\omega)$ accepts H_0 **then**
    print ω
    return
  else
    $N(\omega) \leftarrow$ number of samples needed by the SPRT to reject H_0
  end if
  while $t \geq t_0$ **do**
    Select a neighbor $\omega^{'}$ randomly
    if SPRT(A, B) at $\mathcal{M}(\omega^{'})$ accepts H_0 **then**
      print $\omega^{'}$
      return
    else
      $N(\omega^{'}) \leftarrow$ number of samples needed by the SPRT to reject H_0
    end if
    if $N(\omega^{'}) \geq N(\omega)$ **then**
      $\omega \leftarrow \omega^{'}; N(\omega) \leftarrow N(\omega^{'})$
    else
      $\omega \leftarrow \omega^{'}; N(\omega) \leftarrow N(\omega^{'})$ with probability $e^{-(N(\omega^{'}) - N(\omega))/t}$
    end if
    $t = T(t)$
  end while
  print "No parameter found."
  return

A proof of this claim is shown in a subsequent section. We note that the algorithm in Figure 5 does not actually use the exact value of the probability but merely needs to compare the probability for different choices of parameter values for which it uses the number of samples required by the SPRT instead of calculating the actual probabilities. As such, any other (easily) computable metric that preserves the ordering relationship among parameter values would be sufficient. Thus, the algorithm shown in Figure 5 replaces the computation of the fitness value at a given point with the computation of the number of samples required by the hypothesis testing procedure.

5 Experimental results

In this section, we report the performance of our algorithm using experiments conducted on a parallel implementation. We used a 12 core 16 GB machine with an NVIDIA Compute Unified Device Architecture (CUDA) card to generate multiple simulations in parallel. We also performed an experimental comparative study of the SPRT and fixed sample size statistical estimation. Our results show that our parameter synthesis algorithm (that uses a simulated annealing approach for exploring the state space and the SPRT for calculating the value of the objective function) is more scalable than another technique that uses fixed-sample size estimation for discovering parameters of stochastic biological and biomedical models against behavioural specifications.

5.1 Comparison of SPRT with statistical estimation

The focus of our algorithm is to replace the statistical estimation of the fitness value calculation during simulated annealing style exploration of the parameter space of a stochastic model with the SPRT. Using our theoretical result that the average number of samples required by the SPRT-based algorithm, if the SPRT rejects the null hypothesis, is related to the probability of a specification being true on a parameterised stochastic model, we used a simulated annealing-based parameter synthesis technique that avoids the use of statistical estimation altogether.

In Table 1, we study the number of samples required by a statistical estimation algorithm to compute the probability with which the model satisfies a behavioural specification. We used the z-test to determine the number of samples. It is well known that the random variable representing the sum of n Bernoulli trials, where the success of probability in each trial is p, yields a binomial distribution. Further, when the number of samples n is not too small, the binomial distribution can be approximated by a normal distribution with mean np and variance $np(1-p)$.

Table 1 Number of samples needed by a statistical probability estimation algorithm

Confidence level	*Confidence interval*	*No. of samples*
95%	0.01	9604
95%	0.001	960,400
99%	0.01	16,641
99%	0.001	1,664,100

The SPRT (Wald, 1945) is an *adaptive* procedure where the number of samples is a function of both the actual probability with which the system satisfies the specification

(say p) and the probability with which the behavioural specification requires the system to satisfy the specification (say ρ). In Table 2, we demonstrate the effect of varying ρ when p is fixed. We note that the number of samples is much smaller compared to the statistical estimation approach when there is a large difference between the actual probability with which a behaviour is true for the stochastic model and the probability with which the specification requires the behaviour to be true.

Table 2 Number of samples needed by the SPRT-based algorithm

ρ	*Confidence interval*	*No. of samples*
95%	0.001	377
20%	0.001	1825
5%	0.001	355
95%	0.01	36
5%	0.01	33

Note: $p = 0.5$.

5.2 Benchmark example: studying the influence of pancreas on glucose–insulin metabolism

The synthesis of parameters for biochemical and biomedical models is the primary focus of our research. We developed a parallel CUDA-based implementation of a well-studied glucose–insulin model (DallaMan et al., 2007) that is used to perform in silico validation of artificial pancreata. We simulated a population of patients whose glucose intake was modelled as a normal distribution. We note that our approach does not require us to fix *a priori* the size of the in silico patient population. Instead, the size of the in silico population depends on the region of the parameter space that the algorithm is exploring. Such an adaptive use of in silico population size has not been reported before to the best of our knowledge. Three parameters of the glucose–insulin metabolism model determine the influence of the pancreas on the glucose–insulin dynamics:

- pancreatic responsivity to glucose rate of change;
- delay between glucose signal and insulin secretion;
- pancreatic response to glucose.

We synthesised parameters that ensure that the glucose–insulin subsystem model spends at least 20 minutes in a diabetic scenario, where the glucose concentration in the blood is above 140 or below 80. The results of our synthesis algorithm are presented in Figures 6–8. Thus, we show that the model is capable of reproducing pancreatic-induced diabetes by a suitable parameterisation of three parameters related to the pancreas in the model.

In Table 3, we compare our algorithm to lower bounds of an approach based on statistical estimation. We notice that the statistical estimation approach needs a *uniform* number of samples throughout the synthesis algorithm, while our approach based on the SPRT is adaptive and the number of samples is large only when the parameterised model reaches close to a correctly parameterised model.

Figure 6 Results of the parameter synthesis algorithm (I). The figure shows the first parameter (namely pancreatic responsivity to glucose rate of change) of the glucose-insulin model whose value was discovered by our algorithm (see online version for colours)

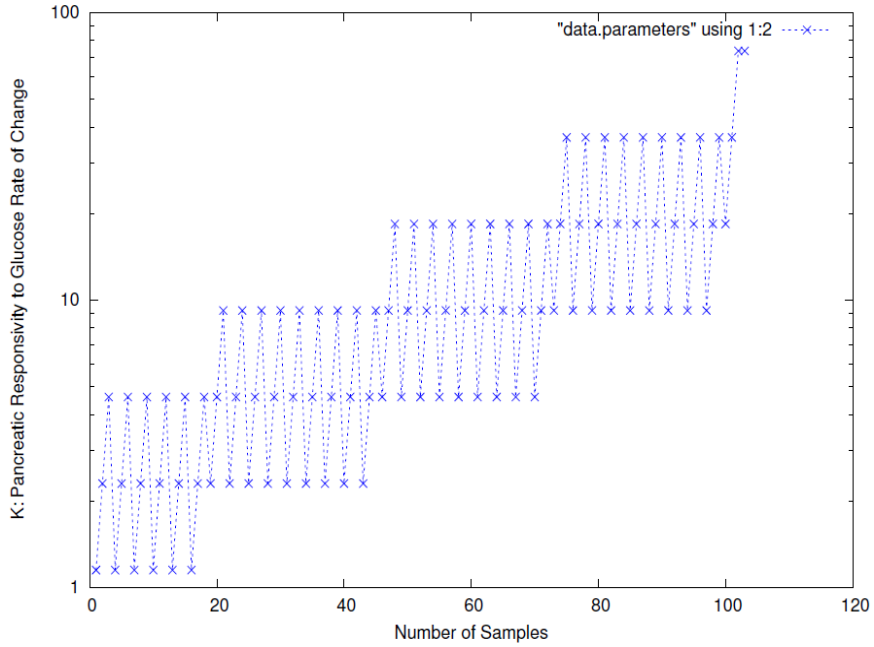


Figure 7 Results of the parameter synthesis algorithm (II). The figure shows the second parameter (namely delay between the glucose signal and insulin secretion) of the glucose-insulin model whose value was discovered by our algorithm (see online version for colours)

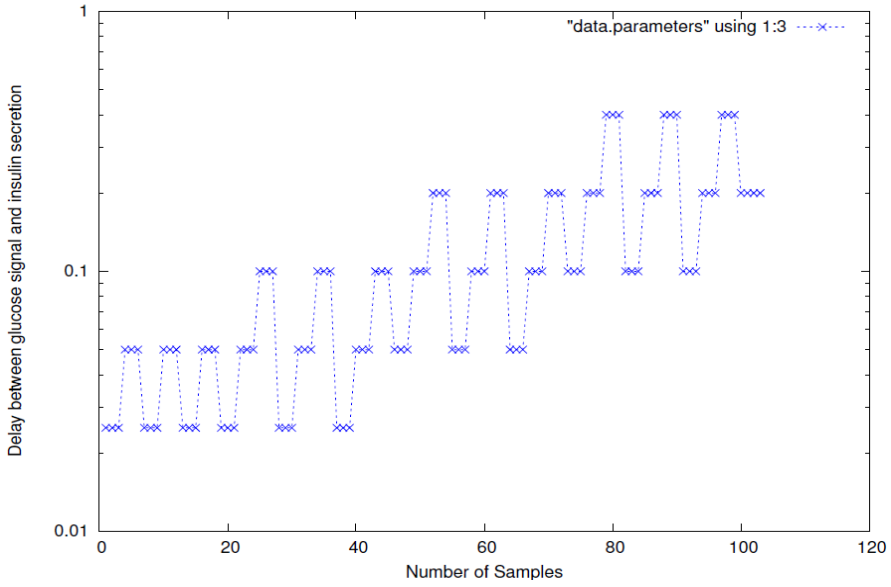


Figure 8 Results of the parameter synthesis algorithm (III). The figure shows the third parameter (namely pancreatic responsivity to glucose) of the glucose-insulin model whose value was discovered by our algorithm (see online version for colours)

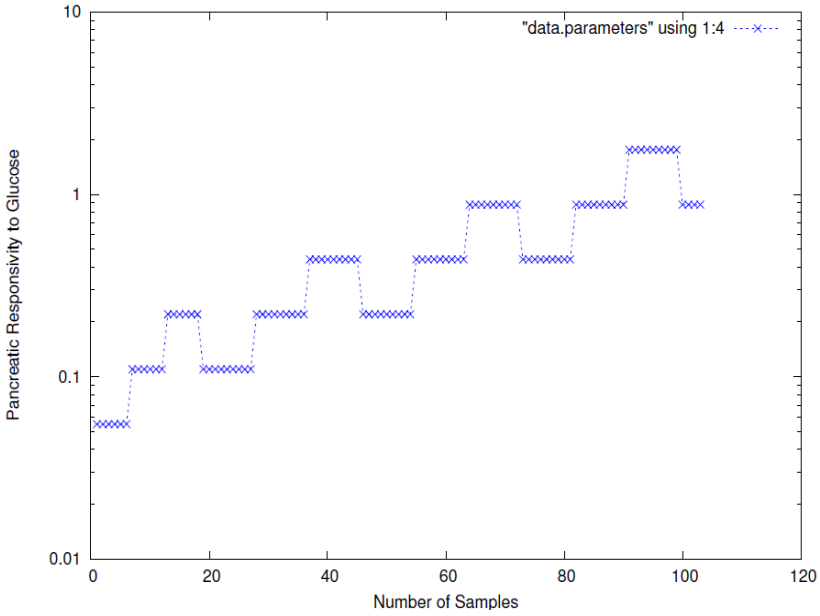


Table 3 Probability estimation vs. SPRT. Estimation assumes a confidence interval of 0.001 and a confidence level of 95%. SPRT assumed a Type-I/II error probability of 10^{-40} and an indifference region of 0.000001

Step	*No. of samples (probability estimation)*	*No. of samples (SPRT)*
1	$>9.6 \times 10^5$	133
10	$>9.6 \times 10^6$	1330
50	$>4\times 10^7$	6743
100	$>8 \times 10^7$	14123

6 Proof of correctness

In this section, we study the relationship between the number of samples required by our parameter synthesis algorithm and the probability of a model satisfying a given specification. Let p be the probability with which the model $\mathcal{M}$ satisfies the specification ϕ and ρ is the minimum desired probability with which the model is required to satisfy the specification.

The SPRT algorithm determines which of the following two hypotheses should be rejected:

- Null hypothesis: $H_0: \ p \geq p_0$;
- Alternate hypothesis: $H_1: \ p \leq p_1$.

Note that $0 \le p_1 < p_0 < 1$, $p_0 = \rho + \varepsilon$ and $p_1 = \rho - \varepsilon$, where 2ε is the indifference region (Younes and Simmons, 2006). Given independent and identically distributed (i.i.d.) samples x_i from the model $\mathcal{M}$, the SPRT procedure defines the following auxiliary quantities:

$$\overline{z_i} = \log\left(\frac{1-p_1}{1-p_0}\right) + x_i \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)$$

$$Z_n = \sum_{i=1}^{n} \overline{z_i}$$

The SPRT(A,B) test accepts the null hypothesis H_0 if $Z_n \le \log(B)$, rejects the null hypothesis H_0 if $Z_n \ge \log(A)$ (where $A = \frac{1-\beta}{\alpha}$ and $B = \frac{\beta}{1-\alpha}$) and continues making i.i.d. observations otherwise (Wald, 1945; Younes and Simmons, 2006). After n Bernoullis trials $x_1, x_2 \ldots x_n$ with the successes defined as $m = \Sigma_{i=1}^{n} x_i$, the SPRT calculates the following quantity:

$$f_n = \frac{\binom{n}{m} p_1^m (1-p_1)^{n-m}}{\binom{n}{m} p_0^m (1-p_0)^{n-m}}$$

$$= \left(\frac{p_1}{p_0}\right)^m \left(\frac{1-p_1}{1-p_0}\right)^{n-m}$$

$$= \left(\frac{p_1}{p_0}\right)^m \left(\frac{1-p_1}{1-p_0}\right)^n \left(\frac{1-p_1}{1-p_0}\right)^{-m}$$

$$= \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^m \left(\frac{1-p_1}{1-p_0}\right)^n$$

$$\log(f_n) = \log\left(\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^m \left(\frac{1-p_1}{1-p_0}\right)^n\right)$$

$$= \log\left(\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^m\right) + \log\left(\left(\frac{1-p_1}{1-p_0}\right)^n\right)$$

$$= m \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) + n \log\left(\frac{1-p_1}{1-p_0}\right)$$

$$= \sum_{i=1}^{m}\left[\log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)\right] + \sum_{i=1}^{n}\left[\log\left(\frac{1-p_1}{1-p_0}\right)\right]$$

$$
\begin{aligned}
&= \sum_{i=1}^{n}\left[I[x_i = 1] \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) \right] \\
&\quad + \sum_{i=1}^{n}\left[\log\left(\frac{1-p_1}{1-p_0} \right) \right] \\
&= \sum_{i=1}^{n}\left[x_i \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) \right] \\
&\quad + \sum_{i=1}^{n}\left[\log\left(\frac{1-p_1}{1-p_0} \right) \right] \\
&= \sum_{i=1}^{n}\left[x_i \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) + \log\left(\frac{1-p_1}{1-p_0} \right) \right] \\
&= \sum_{i=1}^{n} \overline{z_i} \\
&= Z_n
\end{aligned}
$$

We note that $I[x_i = 1]$ is the random variable indicator (Cormen et al., 2009) that indicates whether the random variable x_i has the value 1. Next, we state the theorem relating the average number of samples with the probability of a model satisfying a specification.

Theorem 1: *Let p be the probability with which the model $\mathcal{M}$ satisfies the specification ϕ. Given the null hypothesis H_0: $p \geq p_0$ and the alternate hypothesis H_1: $p \leq p_1$, and error thresholds α and β, if SPRT(A, B) rejects the null hypothesis, the average number of samples observed by our SPRT-based algorithm increases as p increases. Note that A,B are calculated from the α, β as given in Wald (1945).*

Proof:

$$
\begin{aligned}
&E[Z_n \mid p] \\
&= E\left[\sum_{i=1}^{n} \overline{z_i} \mid p \right] \\
&= \sum_{i=1}^{n} E\left[\overline{z_i} \mid p \right] \\
&= \sum_{i=1}^{n} E\left[\log\left(\frac{1-p_1}{1-p_0} \right) + x_i \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) \mid p \right] \\
&= \sum_{i=1}^{n} E\left[\log\left(\frac{1-p_1}{1-p_0} \right) | p \right] + \sum_{i=1}^{n} E\left[x_i \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) | p \right] \\
&= \sum_{i=1}^{n} \log\left(\frac{1-p_1}{1-p_0} \right) + \sum_{i=1}^{n} \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) E[x_i \mid p] \\
&= n \log\left(\frac{1-p_1}{1-p_0} \right) + \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)} \right) np \\
&= nX + nYp
\end{aligned}
$$

Note that $X = \log\left(\frac{1-p_1}{1-p_0}\right)$ and $Y = \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)$ are constants. Further,

$$\begin{aligned}
&\frac{\mathrm{d}E[Z_n \mid p]p}{\mathrm{d}p} \\
&= \frac{\mathrm{d}\left(n\log\left(\frac{1-p_1}{1-p_0}\right) + \log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)np\right)p}{\mathrm{d}p} \\
&= n\log\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) \\
&< 0
\end{aligned}$$

This shows that $E[Z_n \mid p]$ is monotonically decreasing in p. Further, we reject H_0 only if $Z_n \geq \log(A)$. Thus, the SPRT procedure with higher values of p takes a larger number of samples (n) for Z_n to cross the threshold A and, hence, to reject the null hypothesis H_0.

7 Discussion

In this section, we discuss the termination of our statistical model checking-based parameter discovery algorithm (see Figure 5). The heart of our algorithm is the idea of avoiding the computationally expensive, repeated and unnecessary calculation of the precise fitness function value, i.e. the exact probability with which the stochastic model $\mathcal{M}$ satisfies the given behavioural specification ϕ. In Figure 9, we show two parameter points V_s and V_t in the parameter space and the corresponding probabilities p_s and p_t. Recall that the null hypothesis is H_0: $p \geq p_0$. From Figure 9, it is clear that during the exploration process, we should move from V_t to V_s. We have demonstrated that our algorithm ensures that we can replace the actual calculation of the exact probabilities with the expected number of samples required by the SPRT in order to reject the null hypothesis. Figure 10 is a pictorial depiction showing that *if the null hypothesis is rejected, the number of samples required by the SPRT increases with the probability with which the model satisfies the specification*.

The algorithm will terminate when any of the following conditions holds:

1 the current temperature t falls below a threshold temperature t_0,

2 the SPRT-based statistical model checking algorithm itself terminates.

Condition 1 is true because of our monotonically decreasing temperature schedule. Condition 2 is known to be true almost surely for well-framed hypothesis testing queries.

Figure 9 Exploring the parameter space. Our aim is to find a parameterised model that satisfies the null hypothesis $H : p \geq p_0$. While exploring the parameter space, we move to the parameter that is closer to satisfying the null hypothesis (see online version for colours)

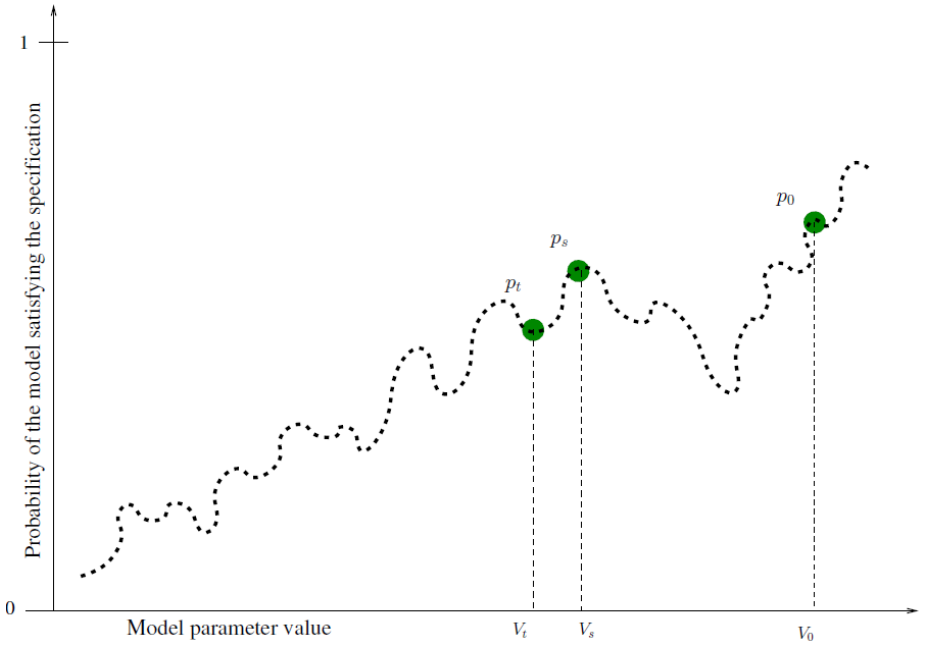
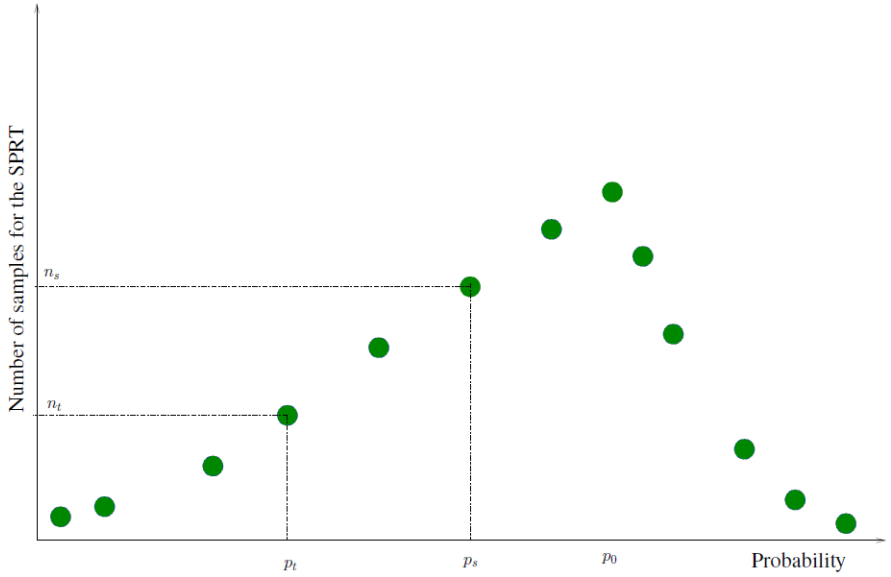


Figure 10 Our SPRT-based algorithm for parameter discovery. While exploring the parameter space, our algorithm uses the number of samples required by the SPRT (instead of calculating the actual probability at each point of the parameterised model satisfying the specification) in order to determine which parameter is better (see online version for colours)



8 Conclusion

We have designed a new SPRT-based parameter discovery algorithm for complex stochastic models of biochemical and biomedical systems. While the traditional approach to simulated annealing-based parameter synthesis for stochastic models requires the precise fitness value calculation by an estimation of the exact probability of a given stochastic model satisfying a given behavioural specification, we showed that such an estimate is computationally expensive to obtain. Further, we argued that the computation of such an estimate is not needed. We presented theoretical results to show that the expected number of samples required during the simulated annealing-based parameter exploration procedure that uses the SPRT for verifying whether a model satisfies a given probabilistic behavioural specification is a good surrogate for the actual estimate of the probability itself. We also argued that the SPRT algorithm requires fewer samples than the statistical estimation algorithm. Finally, we presented experimental evidence to demonstrate the computational attractiveness of our proposed approach.

Several exciting directions for future work remain open. Our proposed algorithm draws new samples whenever a parameter value changes. We believe that a reparameterisation of a stochastic model may not require new samples. Instead, change of measure arguments (Jha and Langmead, 2011) may be used to reuse existing samples with modified probability measures. Another interesting area of research is the use of unbounded temporal specifications. This would permit the specification of interesting properties including cyclic behaviour and periodic oscillations. The curse of dimensionality in parameter discovery is well known, and the proposed algorithm should be merged with existing ideas on model order reduction and sensitivity analysis before it can be deployed in a readily usable parameter discovery tool.

Acknowledgements

The authors thank Bernd Losert (Department of Mathematics, University of Central Florida) for comments on a draft of this paper and Raj Dutta for discussions on the working of the glucose–insulin model. The authors also thank the anonymous reviewers for their comments and suggestions.

References

Amin, M.S., Bhattacharjee, A., and Finley Jr., R.L. and Jamil, H. (2010) 'A stochastic approach to candidate disease gene subnetwork extraction', *Proceedings of the 25th ACM Symposium on Applied Computing*, 22–26 March, Sierre, Switzerland, pp.1534–1538.

André, É. and Soulat, R. (2013) *The Inverse Method*, ISTE Ltd and John Wiley & Sons Inc., London and Hoboken, NJ.

Arenas, N.E., Salazar, L.M., Soto, C.Y., Vizcaino, C., Patarroyo, M.E., Patarroyo, M.A. and Gomez, A. (2011) 'Molecular modeling and in silico characterization of Mycobacterium tuberculosis TlyA: possible misannotation of this tubercle bacilli-hemolysin', *BMC Structural Biology*, Vol. 11, No. 16.

Astrom, K.J. (2006) *Introduction to Stochastic Control Theory*, Dover Publications, Mineola, NY.

Bertsimas, D. and Tsitsiklis, J. (1993) 'Simulated annealing', *Statistical Science*, Vol. 8, No. 1, pp.10–15.

Bottou, L. (2003) 'Stochastic learning', in Bousquet, O. and von Luxburg, U. (Eds): *Advanced Lectures on Machine Learning*, Springer Verlag, Berlin, pp.146–168.

Busch, H., Camacho-Trullio, D., Rogon, Z., Breuhahn, K., Angel, P., Eils, R. and Szabowski, A. (2008) 'Gene network dynamics controlling keratinocyte migration', *Molecular Systems Biology*, Vol. 19, pp.3147–3162.

Caragea, C., Silvescu, A., Caragea, D. and Honavar, V. (2010) 'Abstraction augmented Markov models', *Proceedings of the IEEE 10th International Conference on Data Mining*, 13–17 December, Sydney, NSW, pp.68–77.

Cormen, T.H., Leiserson, C.E., Rivest, R.L. and Stein, C. (2009) *Introduction to Algorithms*, 3rd ed., MIT Press, Cambridge, MA.

DallaMan, C., Rizza, R.A. and Cobelli, C. (2007) 'Meal simulation model of the glucose-insulin system', *IEEE Transactions on Bio-Medical Engineering*, Vol. 54, No. 10, pp.1740–1749.

de Graaf, C., Vermeulen, N.P.E. and Feenstra, K.A. (2005) 'Cytochrome P450 in silico: an integrative modeling approach', *Journal of Medicinal Chemistry*, Vol. 48, No. 8, pp.2725–2755.

Finkbeiner, B. and Sipma, H. (2001) 'Checking finite traces using alternating automata', *Electronic Notes in Theoretical Computer Science*, Vol. 55, No. 2, pp.147–163.

Frehse, G., Jha, S.K. and Krogh, B.H. (2008) 'A counterexample-guided approach to parameter synthesis for linear hybrid automata', *Hybrid Systems: Computation and Control*, Vol. 4981, pp.187–200.

Ghosh, B.K.B.K. and Sen, P.K. (Eds) (1991) *Handbook of Sequential Analysis (Statistics: textbooks and monographs)*, M. Dekker, New York.

Gillespie, D.T. (1977) 'Exact stochastic simulation of coupled chemical reactions', *The Journal of Physical Chemistry*, Vol. 81, No. 25, pp.2340–2361.

Gonzalez, O.R, Küper, C., Jung, K., Naval, P.C. and Mendoza, E. (2007) 'Parameter estimation using simulated annealing for S-system models of biochemical networks', *Bioinformatics*, Vol. 23, No. 4, pp.480–486.

Higo, K., Ugawa, Y., Iwamoto, M. and Higo, H. (1998) 'PLACE: a database of plant cis-acting regulatory DNA elements', *Nucleic Acids Research*, Vol. 26, No. 1, pp.358–359.

Jha, S.K. (2008) 'd-IRA: a distributed reachability algorithm for analysis of linear hybrid automata', *Hybrid Systems: Computation and Control*, Vol. 4981, pp.618–621.

Jha, S.K. and Langmead, C.J. (2011) 'Exploring behaviors of SDE models of biological systems using change of measures', *Proceedings of the IEEE 1st International Conference on Computational Advances in Bio and Medical Sciences*, 3–5 February, Orlando, FL, pp.111–116.

Jha, S.K., Dutta, R.G., Langmead, C.J., Jha, S. and Sassano, E. (2012) 'Synthesis of insulin pump controllers from safety specifications using Bayesian model validation', *Proceedings of the 10th AsiaPacific Bioinformatics Conference*, 17–19 January, Melbourne, Australia.

Jha, S.K., Krogh, B.H., Weimer, J.E. and Clarke, E.M. (2007) 'Reachability for linear hybrid automata using iterative relaxation abstraction', *Hybrid Systems: Computation and Control*, pp.287–300.

Jha, S.K., Langmead, C.J., Mohalik, S. and Ramesh, S. (2011) 'When to stop verification?: statistical trade-off between expected loss and simulation cost', *Proceedings of the Design, Automation & Test in Europe Conference & Exhibition*, 14–18 March, Grenoble, France, pp.1309–1314.

Kalyanaraman, A., Cannon, W.R., Latt, B. and Baxter, D.J. (2011) 'MapReduce implementation of a hybrid spectral library-database search method for large-scale peptide identification', *Bioinformatics*, Vol. 27, pp.3072–3073.

Kovatchev, B.P., Breton, M., DallaMan, C. and Cobelli, C. (2009) 'In silico preclinical trials: a proof of concept in closed-loop control of type 1 diabetes', *Journal of Diabetes Science and Technology*, Vol. 3, No. 1, pp.44–55.

Kulkarni, V.G. (1995) *Modeling and Analysis of Stochastic Systems*, Chapman & Hall, Ltd., London, UK.

Lamport, L. (1982) 'An assertional correctness proof of a distributed algorithm', *Science of Computer Programming*, Vol. 2, No. 3, pp.175–206.

Liu, Q., Cui, J., Yang, Q. and Xu, Y. (2010) 'In-silico prediction of blood-secretory human proteins using a ranking algorithm', *BMC Bioinformatics*, Vol. 11, p.250.

Maybeck, P.S. (1979) *Mathematics in Science and Engineering (Vol. 141): Stochastic Models, Estimation, and Control*, Academic Press, Inc., New York.

Nikolova, O., Zola, J. and Aluru, S. (2007) 'A parallel algorithm for learning Bayesian networks', *Computer Engineering*, Vols. 2–6.

Pnueli, A. (1977) 'The temporal logic of programs', *Proceedings of the 18th IEEE Symposium on Foundations of Computer Science*, 31 October–2 November, Providence, RI, USA, pp.46–67.

Reinker, S., Altman, R.M. and Timmer, J. (2006) 'Parameter estimation in stochastic biochemical reactions', *Systems Biology*, Vol. 153, No. 4, pp.168–178.

Roberts, E., Magis, A., Ortiz, J.O., Baumeister, W. and Luthey-Schulten, Z. (2011) 'Noise contributions in an inducible genetic switch: a whole-cell simulation study', *PLoS Computational Biology*, Vol. 7, No. 3, p.e1002010.

Salis, H. and Kaznessis, Y. (2005) 'Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions', *The Journal of Chemical Physics*, Vol. 122, No. 5, p.54103.

Saul, Z.M. and Filkov, V. (2007) 'Exploring biological network structure using exponential random graph models', *Bioinformatics*, Vol. 23, No. 19, pp.2604–2611.

Turner, T.E., Schnell, S. and Burrage, K. (2004) 'Stochastic approaches for modelling in vivo reactions', *Computational Biology and Chemistry*, Vol. 28, No. 3, pp.165–178.

Wald, A. (1945) 'Sequential tests of statistical hypotheses', *The Annals of Mathematical Statistics*, Vol. 16, No. 2, pp.117–186.

Wald, A. (1947) *Sequential Analysis*, 1st ed., John Wiley and Sons, New York.

Wilkinson, D.J. (2006) *Stochastic Modelling for Systems Biology (Chapman & Hall/CRC Mathematical & Computational Biology)*, 1st ed., Chapman and Hall/CRC, London, UK.

Wilkinson, D.J. (2011) 'Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology (with discussion)', in Bernardo, J.M. (Ed.): *Bayesian Statistics*, Vol. 9, Oxford University Press, Oxford, pp.679–705.

Younes, H.L.S. and Simmons, R.G. (2002) 'Probabilistic verification of discrete event systems using acceptance sampling', *Computer Aided Verification*, pp.223–235.

Younes, H.L.S. and Simmons, R.G. (2006) 'Statistical probabilistic model checking with a focus on time-bounded properties', *Information and Computation*, Vol. 204, No. 9, pp.1368–1409.