Parameter Discovery for Stochastic Biological Models against Temporal Behavioral Specifications using an SPRT based Metric for Simulated Annealing

Faraz Hussain, Raj Gautam Dutta, Sumit Kumar Jha
Electrical Engineering and Computer Science Department
University of Central Florida
Orlando, FL
Email: fhussain,rdutta,jha@cs.ucf.edu

Christopher James Langmead
Computer Science Department
Carnegie Mellon University
Pittsburgh, PA
cjl@cs.cmu.edu

Susmit Jha
EECS Department
UC Berkeley
Berkeley, CA
jha@cs.berkeley.edu

***Abstract*—Stochastic models are often used to study the behavior of biochemical systems and biomedical devices. While the structure of such models is often readily available from first principles, several quantitative features of the model are not easily determined. These quantitative features are often incorporated into the model as parameters. The algorithmic discovery of parameter values from experimentally observed facts (including extreme-scale data) remains a challenge for the computational systems biology community. In this paper, we present a new parameter discovery algorithm based on Wald's sequential probability ratio test (SPRT). Our algorithm uses a combination of simulated annealing and sequential hypothesis testing to reduce the number of samples required for parameter discovery of stochastic models. We use probabilistic bounded linear temporal logic (PBLTL) to express the desired behavioral specification of a model. We also present theoretical results on the correctness of our algorithm, and demonstrate the effectiveness of our algorithm by studying a detailed model of glucose and insulin metabolism.**

I. Introduction

The computational modeling of the precise dynamics of biochemical systems involves the modeling and analysis of complex continuous time Markov chain models. Such detailed biochemical models are often reduced into readily analyzable and succinct models like stochastic differential equations (SDEs). The relatively small size of SDE models and the rich literature on the symbolic analysis of stochastic differential equations make such models a focus of attention in the area of computational systems biology. 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 requires the ability to model, verify, and diagnose complex stochastic models of cyber-physical systems.

The structure of such stochastic models can often be determined from first principles and a survey of existing biochemical literature. However, several quantitative aspects of such models are not readily obtained 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 new contributions to the discovery of parameters for stochastic models of biochemical systems and biomedical devices:

1) We present a new algorithm for discovering parameters of stochastic computational models from experimental observations. *Our algorithm uses a combination of simulated annealing and the sequential probability ratio test (SPRT) to reduce the number of samples required for discovering parameter values.* The reduction in the number of samples stems from the fact that the best statistical hypothesis testing algorithm is at least as efficient as the best statistical estimation algorithm.
2) Our algorithm uses the fact that *the expected number of samples required by the sequential probability ratio test (SPRT) to accept the null hypothesis for a given parameterized model is proportional to the probability with which the parameterized model satisfies the given specification*. Our proof does not explicitly compute the number of samples required by the SPRT procedure.

We apply our proposed algorithm to a detailed model of glucose-insulin metabolism. The metabolic model is important for *in silico* verification of artificial pancreata. We show that the model is capable of reproducing pancreata-induced diabetes by a re-parameterization of three parameters related to the pancreata in the model.

II. Related Work

The field of stochastic modeling has emerged to overcome the inherent limitations of deterministic modeling [1]. Such a modeling approach permits randomness in the behavior of the system, and permits the development of models with multiple outcomes. One often uses sampling based statistical inference algorithms to determine the probability distribution of potential outcomes. The emerging field of system biology has greatly benefited from existing literature in stochastic modeling, and has also inspired new scientific expeditions into the realm of stochastic modeling.

978-1-4673-1321-6/12/$31.00

Indeed, *in silico* modeling has been particularly useful in developing a systems view of biology. Models for whole cell analysis, drug discovery, control of Type 1 diabetes, sequence analysis, studying enzyme interaction with drugs, and prediction of blood-secretory proteins are some of the success stories in computational systems biology. However, several components of these models are not available from first principles. In such a scenario, model designers include missing information as parameters in the model. The number of such parameters increases with the size and the complexity of the model, and it is difficult to determine the values of these parameters for large and detailed model of biological systems.

The discovery of parameters for stochastic models has been carried out using various approaches [2], [3]. Different estimation techniques have been adopted by researchers for finding parameters of stochastic biochemical reactions. Estimators used for deterministic models have also been extended to stochastic models. Recently, researchers have adopted Bayesian frameworks [4], [5] for parameter identification in stochastic models such as stochastic gradient descent, simulated annealing and evolutionary algorithms. Considerable research has been directed in the use of statistical hypothesis testing for verification of stochastic models [6], [7], including those arising in systems biology. In particular, we have studied the in silico synthesis of insulin pumps from behavioral specifications using Bayesian hypothesis testing [4].

III. Background

In this section, we discuss the various classes of stochastic models that benefit from our parameter discovery algorithm. We also present specification formalisms for representing facts observed from experimental data. Finally, we briefly survey the literature on sequential probability ratio test (SPRT) and its relationship to statistical estimation.

A. Stochastic Models

Our proposed algorithm can be applied to several classes of parameterized stochastic models, including continuous time Markov chains, stochastic differential equations, jump diffusion processes, and heterogeneous models consisting of stochastic processes interacting with deterministic models like ordinary differential equations.

Figure 1 illustrates the various types of stochastic models whose parameters can be discovered using our algorithm. Discrete time Markov chains (DTMCs) are finite state space models where each state is labeled by a given value for each variable, and each transition between the states of a DTMC is associated with a finite probability. Dynamic Bayesian networks (DBNs) are a factored and succinct representation of Markov-like models. All of these models are discrete event stochastic systems, where the behavior of the system over a finite period of time can be completely

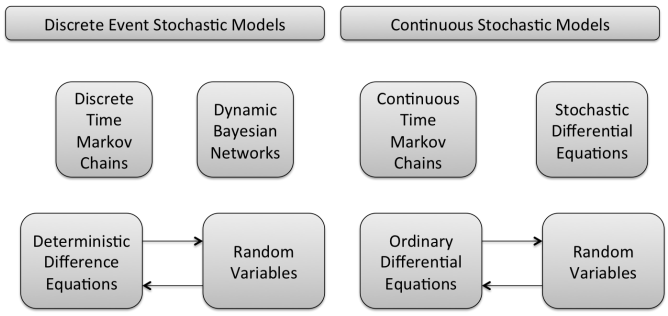

Figure 1. Our algorithm is applicable to both discrete and continuous time stochastic models. CTMCs are particularly important for studying biochemical systems, while ODEs interacting with random variables naturally model a number of cyber-physical biomedical devices.

described by a finite numerical sequence of values assigned to system variables. Continuous time Markov chains and stochastic differential equations represent stochastic systems that evolve continuously in time. Another interesting class of systems is formed by a set of deterministic differential equations interacting with a set of random variables or stochastic processes. It is often difficult to obtain value of parameter and variables associated with such models using first principle. The algorithm presented in this paper can be used to discover parameter values for models of such complex systems.

B. Specifications

The specification of the correct behavior of stochastic biochemical and biomedical models requires complex temporal reasoning. We suggest the use of probabilistic bounded linear temporal logic (PBLTL) to specify the correct behavior 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* (BLTL) [8].

A Bounded Linear Temporal Logic specification is a set of predicates connected using Boolean and temporal operators. The syntax of the logic is given by the following grammar:

$$\phi ::= x \leq v \mid x \geq v \mid (\phi_1 \vee \phi_2) \mid \neg\phi_1 \mid (\phi_1 \mathbf{U^t} \phi_\mathbf{2}),$$

where $\mathcal{V}$ is set of continuous-valued variables, $x \in \mathcal{V}$, $v \in \mathbb{R}$, and t denotes time. We can define additional temporal operators such as $\mathbf{F^t}\psi = \mathbf{True}\,\mathbf{U^t}\,\psi$, or $\mathbf{G^t}\psi = \neg\mathbf{F^t}\neg\psi$. Intuitively, the formula $\mathbf{F^t}\psi$ implies that ψ holds sometime within t time units. Similarly, 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 $\wp \models \phi$.

Specifications about biochemical and biomedical systems are often probabilistic in nature. Such behaviors are naturally expressed as probabilistic bounded linear temporal logic (BLTL) specifications. If ϕ is a bounded linear temporal logic (BLTL) specification, $Pr_{\geq\rho}(\phi)$ is a probabilistic

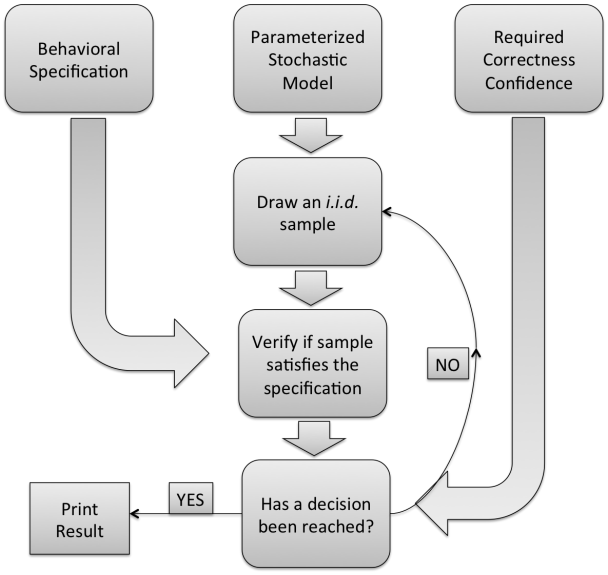

Figure 2. The SPRT procedure: Given a behavioral specification and a parameterized stochastic model, the SPRT procedure decides if the model satisfies the behavioral specification.

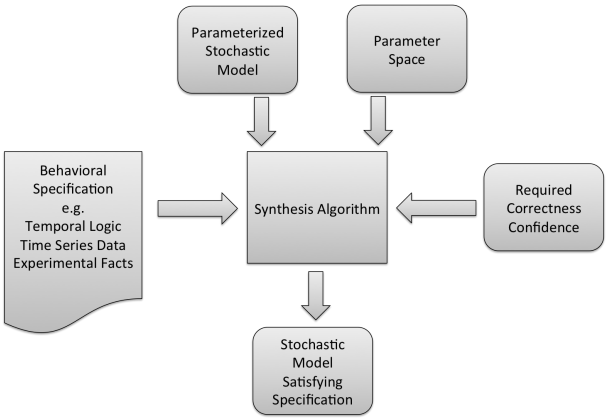

Figure 3. The Problem Definition: Given behavioral specifications about the system being modeled, the parameterized stochastic models, the parameter space and a correctness confidence, our algorithm seeks to discover the value of the parameters that enables the given model to satisfy the given probabilistic behavioral specification. The algorithm also ensures that the answer is correct with the given confidence.

bounded linear temporal logic specification. The model $\mathcal{M}$ is said to satisfy this PBLTL specification if at least ρ fraction of independently drawn random behaviors observed from the model satisfy the BLTL specification ϕ.

C. Sequential Probability Ratio Test (SPRT)

Given a stochastic model $\mathcal{M}$ and a property specification ϕ, the sequential probability ratio test (SPRT) decides which of the following hypotheses is true: (i) Null Hypothesis $H_0 : p \leq p_0$, (ii) Alternate Hypothesis $H_1 : \ p > p_0$. The SPRT procedure uses a sequential sampling procedure, 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.

In a sequential probability ratio test (SPRT), each observations can lead to three outcomes: (i) the null hypothesis may be rejected; (ii) the alternate hypothesis may be rejected; (iii) additional observations may be needed to reject either hypothesis and hence, the procedure continues. This process of generating additional samples continues until a decision is made to reject one of the hypotheses. The SPRT procedure is illustrated in Figure 2.

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. 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 an error of the second kind. A SPRT algorithm bounds the probability of making Type-I and Type-II during hypothesis testing. The errors are bounded by two integer constant a and b that parameterize the SPRT test $S(b, a)$.

IV. Our Approach

In this section, we present a new parameter discovery algorithm for stochastic models. Our approach brings together the sequential probability ratio test (SPRT) and the simulated annealing procedure. The algorithm uses fewer number of samples than a traditional approach based on statistical estimation and simulated annealing. The correctness proof of our algorithm is based on the fact that the expected number of samples required by the SPRT algorithm is related to the probability being estimated.

A. Algorithm

The parameter discovery problem is presented in Fig. 3. There are four inputs to our algorithm:

1) *Behavioral Specification*: The user provides a probabilistic specification about the behavior of the stochastic biological system. The specification may be provided as probabilistic bounded linear temporal logic (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) *Parameterized Stochastic Model*: The algorithm discovers the parameter values for parameterized 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) *Confidence*: The algorithm also accepts a probabilistic confidence as an input. Simulation based analysis of stochastic models requires establishing a confidence on the probability of a model satisfying a specification. This parameter (whose value should be close to 1) determines the confidence with which the algorithm must ascertain all probability estimates during SPRT based hypothesis testing.

Our parameter discovery algorithm builds on the classical simulated annealing approach for parameter discovery (See Figure 1). Simulated annealing is a stochastic optimization 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 optimization steps stochastically. Unlike deterministic optimization algorithms, it is iterative; the global minimum of a system is located by moving step by step through the function space, based on the value of an objective function. When caught up in a local minimum, the algorithm uses random walks, and probabilistically determines the next step in the parameter space. The probabilistic random walk depends on a parameter (often called "temperature") whose value is determined heuristically. This process continues until the algorithm converges to one of the global minima.

Algorithm 1 Simulated Annealing Algorithm

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

In optimization problems with stochastic models, an estimation of the objective function is made using statistical estimation and other approaches [1]. Several heuristic estimation methods have been used for computing the probability (objective function) of a model satisfying a specification during sequential annealing. However, these estimation based methods require a large number of samples and can not be readily used for parameter discovery. We propose the use of the sequential probability ratio test (SPRT) procedure to construct a computationally efficient objective function for sequential annealing. This approach helps us in comparing the various states of the system towards obtaining a global minima.

Algorithm 1 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 minima. The algorithm performs a biased random walk on the parameter space - it always accepts a better parameter value and sometimes accepts a worse parameter value. The probability of accepting a worse parameter value keeps becoming smaller with time. Thus, when the algorithm converges, it has obtained a global minima.

Algorithm 1 reaches an optimal value when $E(\omega^*)$ and $E(\omega)$ are estimated correctly. The estimation procedure is tedious and challenging for stochastic systems due to the presence of randomness and the consequent necessity of observing millions of model simulations before the model parameters can be adequately estimated. We observe that such a precise estimation of the probability values is not useful for most of the parameter space being explored. We suggest the use of the sequential probability ratio test for deciding if a parameter value is interesting *without* explicitly evaluating the probability value associated with the parameter. When used together with sequential annealing, SPRT enables an efficient synthesis of parameter values for stochastic models.

Algorithm 2 SPRT based Simulated Annealing Algorithm

Require: Parameter space Ω, Probabilistic Behavioral Specification $Pr_{\geq \rho}(\phi)$, Probability of the model satisfying the specification $\theta : \Omega \rightarrow [0, 1]$, *Temperature Cooling Schedule* $\mathrm{T} : \mathbb{N} \rightarrow (0, \infty)$, Starting Temperature t, Stopping Temperature t_0, Bounds on Type-I/II errors: α, β.
$H_0 : \mathcal{M}(\omega) \models Pr_{\leq \rho}(\phi)$
$H_1 : \mathcal{M}(\omega) \not\models Pr_{\leq \rho}(\phi)$
$a = \frac{1-\beta}{\alpha}$
$b = \frac{\beta}{1-\alpha}$
ω = Pick a random point in Ω.
$N(\omega) = 0$
while $t \geq t_0$ **do**
  for all neighbors of ω **do**
    Select a neighbor $\omega^{'}$ randomly
    $N(\omega^{'})$ = Number of samples required to perform SPRT(a, b) for testing H_1 against H_0.
    if $SPRT(a, b) \implies H_0$ is rejected **then**
      print ω as recommended model parameter
      return
    end if
    if $N(\omega^{'}) \leq N(\omega)$ **then**
      $\omega \leftarrow \omega^{'}$ //Make a move
    end if
    if $N(\omega^{'}) < N(\omega)$ **then**
      $\omega \leftarrow \omega^{'}$ with probability $e^{-(N(\omega^{'}) - N(\omega))/T(t)}$
    end if
  end for
end while
Ensure: Algorithm synthesizes ω such that $\mathcal{M}(\omega) \models Pr_{\geq \rho}(\phi)$

Instead of using an estimate of the probability with which the parameterized model satisfies the specification, Algorithm 2 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 parameterized model satisfies a behavioral specification. Let p be the probability with which the model $\mathcal{M}$ satisfies the specification ϕ, i.e., $\mathcal{M} \models Pr_{=p}(\phi)$. The binomial SPRT algorithm decides which of the following two hypotheses should be rejected: (i) Null Hypothesis $H_0 : p \leq p_0$ (ii) Alternate Hypothesis $H_1 : p > p_1$. Note that $0 \leq p_0 \leq p_1 < 1$.

Theorem 1: If SPRT(b,a) accepts the null hypothesis, the average number of samples observed by the SPRT algorithm increases as p increases.

A proof of the theorem is presented in [9]. We note that Algorithm 1 does not actually use the exact value of the probability but merely needs to compare the probability for different choices of parameter values. As such, another easily computable metric that preserves the ordering relationship among parameter values would be sufficient. Thus, Algorithm 2 replaces the computation of the exact probability with the computation of the number of samples required by the hypothesis testing procedure.

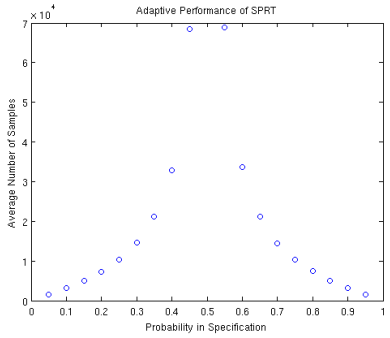

Figure 4. Number of samples needed by the SPRT algorithm: $p_{system} = 0.5$

V. Experimental Results

In this section, we study the performance of our algorithm using a parallel implementation. We used a 12 core 16GB machine with an NVIDIA CUDA card to generate multiple simulations in parallel. We also perform an experimental comparative study of the sequential probability ratio test and fixed sample size statistical estimation. Our results show that an SPRT based simulated annealing parameter synthesis algorithm is a more scalable approach to the discovery of parameters for stochastic biological and biomedical models against behavioral specifications.

A. Comparison of SPRT with Statistical Estimation

The key focus of our algorithm is to replace the statistical estimation of the probability of a parameterized model with a sequential probability ratio test (SPRT) based algorithm. Using our theoretical result that the average number of samples used by the SPRT algorithm is related to the probability of a specification being true on a parameterized stochastic model, we suggest a parameter synthesis algorithm that avoids the use of statistical estimation altogether.

We study the number of samples required by a statistical estimation algorithm to compute the probability with which the model satisfies a behavioral specification. We have used the z-test to determine the number of samples required by statistical estimation. 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 $np(1-p)$. Thus, the z-test is an appropriate choice for a parametric statistical estimation procedure. We find that a confidence interval of 0.001 and a confidence level of 99% requires over 1 million simulations.

The sequential probability ratio test (SPRT) is an *adaptive* procedure where the number of samples is a function of both the true probability with which the system satisfies the specification (say, p_{system}) and the probability with which the behavioral specification requires the system to satisfy the specification (say, p_{spec}). In Figure 4, we demonstrate the effect of varying p_{spec} when p_{system} 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 behavior is true for the stochastic model and the probability with which the specification requires the behavior to be true. We note that this is indeed the usual case for a simulated annealing based parameter discovery algorithm.

B. Benchmark Example: Studying the influence of pancreas on glucose insulin metabolism

The synthesis of parameters for biochemical and biomedical models is a primary focus of our research. We developed a massively parallel CUDA based implementation of a detailed glucose-insulin model that can be used to perform in silico validation of artificial pancreas. We simulated a population of patients whose glucose intake was varied, and modeled as a normal distribution. We note that our approach does not require us to *a priori* fix the size of the in-silico 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 in literature before.

Three parameters in the glucose insulin metabolism model determine the influence of pancreas on the glucose-insulin dynamics:

- Pancreatic responsivity to glucose rate of change,
- Delay between glucose signal and insulin secretion, and

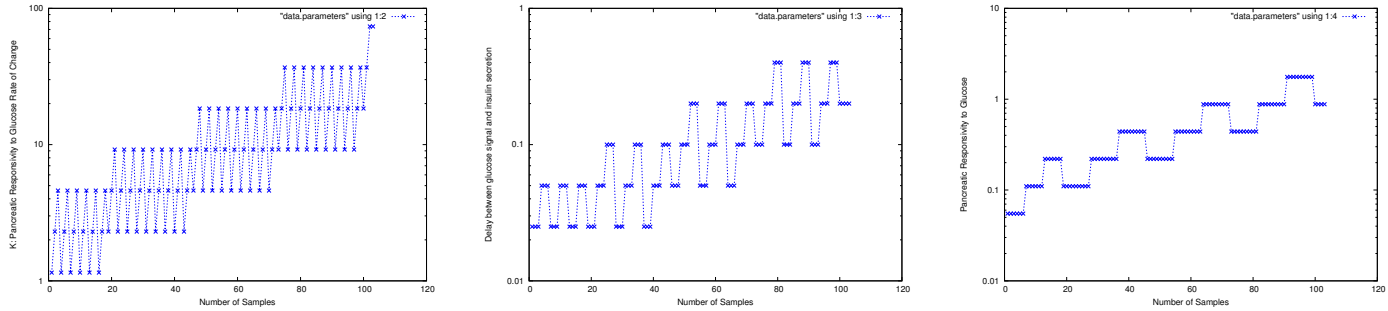

Figure 5. Results of the synthesis algorithm

- Pancreatic response to glucose

We synthesized the 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 result of our synthesis algorithm is presented in Figure 5. Thus, we show that the model is capable of reproducing pancreata-induced diabetes by a re-parameterization of three parameters related to the pancreata in the model. We notice that the statistical estimation approach needs an *uniform* number of samples throughout the synthesis algorithm, while an approach based on SPRT is adaptive and the number of samples is large only when the parameterized model reaches close to a correctly parameterized model.

VI. Conclusions

We have proposed a new sequential probability ratio test (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 estimation of the probability of a given stochastic model satisfying a given behavioral specification, we show that such an estimate is computationally expensive to obtain. Further, we argue that the computation of such an estimate is not needed. We present theoretical results to show that the number of samples obtained during the sequential probability ratio test procedure for verifying whether a model satisfies a given probabilistic behavioral specification is a good surrogate for the actual estimate of the probability itself. We also argue that the SPRT algorithm requires much fewer samples than the statistical estimation algorithm. We finally present 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 simple re-parameterization of a stochastic model may not require us to obtain new samples. Instead, change of measure arguments 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 behavior and periodic oscillations. Further, the curse of dimensionality in parameter discovery is well known, and needs to be addressed before this research can be deployed in a readily usable parameter discovery tool.

References

[1] P. S. Maybeck, T. In, A. Form, O. B. Means, O. Mechanical, I. Photocopy, and M. P. S, "Stochastics models, estimation, and control: Introduction," 1979.

[2] A. Kalyanaraman, W. R. Cannon, B. Latt, and D. J. Baxter, "MapReduce implementation of a hybrid spectral Library-Database search method for large-scale peptide identification," *Bioinformatics*, Sep. 2011.

[3] S. K. Jha, "d-ira: A distributed reachability algorithm for analysis of linear hybrid automata," in *HSCC*, 2008, pp. 618–621.

[4] S. K. Jha, R. G. Dutta, C. J. Langmead, S. Jha, and E. Sassano, "Synthesis of insulin pump controllers from safety specifications using bayesian model validation," in *Proceedings of the Tenth AsiaPacific Bioinformatics Conference*, 2012.

[5] O. Nikolova, J. Zola, and S. Aluru, "A parallel algorithm for learning bayesian networks," *Computer Engineering*, pp. 2–6, 2007.

[6] S. K. Jha, C. J. Langmead, S. Mohalik, and S. Ramesh, "When to stop verification?: Statistical trade-off between expected loss and simulation cost," in *DATE*. IEEE, 2011, pp. 1309–1314.

[7] S. K. Jha and C. J. Langmead, "Exploring behaviors of sde models of biological systems using change of measures," in *ICCABS*, I. I. Mandoiu, S. Miyano, T. M. Przytycka, and S. Rajasekaran, Eds. IEEE Computer Society, 2011, pp. 111–116.

[8] A. Pnueli, "The temporal logic of programs," in *Proceedings of the 18th IEEE Symposium on Foundations of Computer Science*, 1977, pp. 46–67.

[9] F. Hussain, R. G. Dutta, S. K. Jha, C. J. Langmead, and S. Jha, "Parameter discovery for stochastic biological models against temporal behavioral specifications using an sprt based metric for simulated annealing," University of Central Florida, Tech. Rep., 2012.