A Bayesian Approach to Model Checking Biological Systems*

Sumit K. Jha[1], Edmund M. Clarke[1], Christopher J. Langmead[1,2], Axel Legay[3], André Platzer[1], and Paolo Zuliani[1]

[1] Computer Science Department, Carnegie Mellon University, USA
[2] Lane Center for Computational Biology, Carnegie Mellon University, USA
[3] Institut d'Informatique INRIA, Rennes, France

Abstract. Recently, there has been considerable interest in the use of Model Checking for Systems Biology. Unfortunately, the state space of stochastic biological models is often too large for classical Model Checking techniques. For these models, a statistical approach to Model Checking has been shown to be an effective alternative. Extending our earlier work, we present the first algorithm for performing statistical Model Checking using Bayesian Sequential Hypothesis Testing. We show that our Bayesian approach outperforms current statistical Model Checking techniques, which rely on tests from Classical (aka Frequentist) statistics, by requiring fewer system simulations. Another advantage of our approach is the ability to incorporate prior Biological knowledge about the model being verified. We demonstrate our algorithm on a variety of models from the Systems Biology literature and show that it enables faster verification than state-of-the-art techniques, even when no prior knowledge is available.

1 Introduction

Computational models are increasingly used in the field of Systems Biology to examine the dynamics of biological processes (e.g., [8,10,20,30,34,37]). By 'computational', we mean discrete-variable and continuous or discrete-time models [4], where the components of the system interact and evolve by obeying a set of instructions or rules. In contrast to differential equation-based models, which are also widely used in Systems Biology, computational models can provide insights into the role of stochastic effects over discrete-populations of molecules or cells. Recently, there has been considerable interest in the application of Model

* This research was sponsored by the GSRC (University of California) under contract no. SA423679952, National Science Foundation under contracts no. CCF0429120, no. CNS0411152, and no. CCF0541245, Semiconductor Research Corporation under contract no. 2005TJ1366, Air Force (University of Vanderbilt) under contract no. 18727S3, International Collaboration for Advanced Security Technology of the National Science Council, Taiwan, under contract no. 1010717, the U.S. Department of Energy Career Award (DE-FG02-05ER25696), and a Pittsburgh Life-Sciences Greenhouse Young Pioneer Award.

P. Degano and R. Gorrieri (Eds.): CMSB 2009, LNBI 5688, pp. 218–234, 2009.

Checking [15] as a powerful tool for *formally* reasoning about the dynamic properties of such models (e.g., [1,6,9,11,14,18,24,38]). This paper presents a new Model Checking algorithm that is well-suited for verifying properties of very large stochastic models, such as those created and used in Systems Biology.

The stochastic nature of most computational models from Systems Biology gives rise to an instance of the *Probabilistic Model Checking* (PMC) problem [13,15,31]. Suppose $\mathcal{M}$ is a stochastic model over a set of states S, s_0 is a starting state, ϕ is a dynamic property expressed as a formula in temporal logic, and $\theta \in [0,1]$ is a probability threshold. The PMC problem is: given the 4-tuple $(\mathcal{M}, s_0, \phi, \theta)$, to decide algorithmically whether $\mathcal{M}, s_0 \models P_{\geq\theta}(\phi)$. In this paper, property ϕ is expressed in BLTL - Bounded Linear Temporal Logic [36,35,19]. Given these, PMC algorithms decide whether the model satisfies the property with at least probability θ.

Existing algorithms for solving the PMC problem fall into one of two categories. The first category comprises numerical methods (e.g. [2,3,12,16,31]) which can compute the probability with which the property holds with high precision. Numerical methods are generally only suitable for small systems ($\approx 10^6$ to 10^7 states). In a Biological System, the number of states can easily exceed this limit, which motivates the need for algorithms for solving the PMC problem in an approximate fashion. Approximate methods (e.g., [23,26,39,46]) work by sampling a set of *traces* from the model. Each trace is then evaluated to determine whether it satisfies the property. The number of satisfying traces is used to (approximately) decide whether $\mathcal{M}, s_0 \models P_{\geq\theta}(\phi)$.

Approximate PMC methods can be further divided into two sub-categories: (i) those that seek to *estimate* the probability that the property holds and then compare that estimate to θ (e.g., [26,39]), and (ii) those that reduce the PMC problem to a *hypothesis testing* problem (e.g., [46,47]). That is, deciding between two hypotheses — $H_0 : P_{\geq\theta}(\phi)$ versus $H_1 : P_{<\theta}(\phi)$. Hypothesis-testing based methods are more efficient than those based on estimation when θ (which is specified by the user) is significantly different than the true probability that the property holds (which is determined by $\mathcal{M}$ and s_0) [45].

Existing PMC methods based on hypothesis testing rely on *Classical* (aka *Frequentist*) statistical procedures, like Wald's Sequential Probability Ratio Test (SPRT) [42], to answer the decision problem. Our algorithm performs hypothesis testing, but uses *Bayesian* statistical procedures. This distinction is not trivial, as Bayesian and Classical statistics are two very different fields. We will show that in practice, our Bayesian approach requires fewer samples than Wald's SPRT. Finally, we note that because we adopt a Bayesian approach, our algorithm can incorporate prior knowledge, in the form of a probability distribution, $P(\theta)$, when available. This is relevant because in a Biological setting, it is often the case that prior knowledge is available.

The contributions of this paper are as follows:

- The first application of Bayesian Sequential Hypothesis Testing to statistical Model Checking,