Computing Biological Model Parameters by Parallel Statistical Model Checking *

T. Mancini, E. Tronci, I. Salvo, F. Mari, A. Massini, and I. Melatti
Computer Science Department, Sapienza University of Rome, Italy

Abstract. Biological models typically depend on many *parameters*. Assigning suitable values to such parameters enables *model individualisation*. In our clinical setting, this means finding a model for a given patient. Parameter values cannot be assigned arbitrarily, since *inter-dependency* constraints among them are not modelled and ignoring such constraints leads to biologically meaningless model behaviours. Classical parameter identification or estimation techniques are typically not applicable due to scarcity of clinical measurements and the huge size of parameter space. Recently, we have proposed a statistical algorithm that finds (almost) all biologically meaningful parameter values. Unfortunately, such algorithm is computationally extremely intensive, taking up to months of sequential computation. In this paper we propose a parallel algorithm designed as to be effectively executed on an arbitrary large cluster of multi-core heterogenous machines.

1 Introduction

Systems biology models aim at providing quantitative information about time evolution of biological species. One of the main goals of systems biology in a health-care context is to *individualise* models in order to compute patient-specific predictions (see, e.g., [24]) for the time evolution of species (e.g., hormones).

Depending on the system at hand, many modelling approaches are currently investigated. For example, see [22,21] for an overview on discrete as well as continuous modelling approaches, and [49] for a survey on stochastic modelling approaches. In biological networks modelled with a system of *Ordinary Differential Equations (ODEs)* depending on a set of parameters (as in, e.g., [35,50,39]) model individualisation can be done by assigning suitable values to the model parameters. Such biological models depend on many (easily hundreds of) parameters, whose values cannot be chosen arbitrarily because of *inter-dependency* constraints among them (see, e.g., [26]) that, usually, are not explicitly known and thus are not modelled. If model parameter values are chosen ignoring such constraints, then the resulting model behaviour is biologically meaningless.

Model identification (see, e.g., [27]) techniques are typically used to compute values for model parameters so that a suitable error function measuring mismatch between model predictions and experimental data is minimised (*parameter estimation*). If such a value exists and is unique, the model is said *identifiable*. In a clinical setting, for each patient, only a small number of measurements is available, since they can be costly, invasive and time-consuming. Therefore, although in principle model identification techniques could be used to compute

* This work has been partially supported by the EC FP7 project PAEON (Model Driven Computation of Treatments for Infertility Related Endocrinological Diseases, 600773).

patient-specific model parameters, in practice, because of the large amount of measurements needed (see, e.g., [9]), they are typically used to compute a *default parameter value* that averages among the behaviours of many patients (as, e.g. in [39]). *Parameter estimation* approaches cannot be used either, since with such a few data they would not take into due consideration inter-dependencies among model parameters [26].

Motivations To overcome scarcity of measurements, we proposed a two-phase approach [47]. First, an *off-line* phase that accounts for parameter inter-dependencies [26] greatly narrows down the search space to a set S of parameter values yielding *biologically meaningful* model behaviours. Second, an *on-line* phase computes a patient-specific model by selecting in S those parameter values that minimise mismatch with respect to patient measurements. This enables fast patient-specific predictions for the time evolution of each species of interest.

In general, to decide if time evolution of species concentration is biologically meaningful takes a domain expert. To build a general purpose tool that can automatically search through millions of model parameter values, in [47] we proposed a criterion which regards as *Biologically Admissible (BA)* those parameter values entailing time evolutions with a second order statistics *close enough* to that of the model default parameter values.

The computation of the set S of BA values for the model parameters requires to explore the set of all possible values for the parameter vector, that is typically huge. Since an exhaustive exploration would be unfeasible, in [47] a Statistical Model Checking (SMC) based approach is proposed. Nonetheless, the exploration of the parameter space may take *months* of sequential computation.

Main contributions Our main contribution is a SMC parallel algorithm and its distributed multi-core implementation to compute the set of all BA parameter values. We propose a master-slave architecture where a single master process (*Orchestrator*) implements the SMC algorithm and delegates to a high number of slaves (*BA Verifier*s) the numerical integration of the ODE system defining the model. As a consequence, our parallel algorithm will benefit from the availability of many heterogeneous computational units (e.g., a data-centre in the cloud).

The SMC algorithm proposed in [47] would require too much synchronisation in a parallel context. Here, we define a new random sampling process at the basis of our SMC algorithm, which enables massive parallelisation of BA Verifiers.

We developed our distributed multi-core tool in the C language using Message Passing Interface (MPI) [42]. We evaluate effectiveness of our approach by using it on the *GynCycle* model in [39], an ODE model which predicts blood concentration of several species during female menstrual cycle. We show that our implementation achieves high efficiency even when using dozens of computational cores (e.g., efficiency is 74% when using 80 cores).

Related work The input to our algorithm consists of a system model along with the *default value* for its parameters. The *GynCycle* model considered in our case study has been presented in [39] and the default (inter-patient) values for its parameters have been computed in [11] using model identification (often referred to as *parameter identification* in our setting) techniques [27].

In recent years, parallel and distributed computing has received attention in order to cope with the complexity of biological systems. See [5] for a survey of parallel methods to solve ODEs, parallel model checking, and parallel simulations in biological applications.

Statistical Model Checking mainly addresses system verification of stochastic systems with respect to probabilistic temporal properties or continuous stochastic properties (see, e.g., [41]). Several parallel and distributed approaches to SMC have been introduced (see, e.g., [3,40]), some of them motivated by the complexity of biological models [4]. Here, we focus on deterministic biological systems modelled with ODEs, and we apply SMC techniques (along the lines of [18]) to infer statistical completeness of our set of BA parameters.

Parallel approaches close to ours are those in [7,44,8], where the problem of computing all (discretised) model parameter values meeting given LTL properties has been investigated. We extend such works in two directions. First, the above mentioned papers focus on piecewise affine ODE systems, whereas we can handle any (possibly non-linear) ODE system. Second, they aim at computing a maximal set of parameters satisfying a given LTL property. Thus, when the model changes, a new LTL property has to be provided by domain experts. Our approach infers such a system property by the default value for the model parameters, thus decreasing the amount of input needed from domain experts.

A key feature of parameter identification approaches is their ability to give information about parameter *identifiability* (see, e.g., [9] and citations thereof). Gradient-based methods, as, e.g., the classical one in [25], provide a local optimum solution to the parameter estimation problem. Global methods, such as [28], provide a global optimum solution whereas heuristics approaches as evolutionary algorithms (see, e.g., [6,45]), provide near-global optimal solutions. All such approaches do not provide information about parameter identifiability. When observations are scarce, parameters usually become non-identifiable. Studying the correlation among system parameters can reduce the number of data needed for identifiability, see for example [37,26]. Our goal here is to support model individualisation from clinical measurements. This means that we need to compute model parameters from a few (say, 3) observations about a small subset (4 in our case study) of the species occurring in the model (33 in our case). Because of scarcity of measurements, neither model identification approaches nor parameter estimation approaches can be used in our setting.

Model checking based parameter estimation approaches have been investigated for example in [20,12,38]. Such approaches differ from ours, since they do not address the problem of automatically restricting the search space. Model checking techniques have been widely used in systems biology, to verify time behaviours. Examples are in [23,19,14,16,35]. Such approaches focus on verifying a given property for the model trajectories, whereas our main problem here is to compute *all* biologically plausible values for the model parameters.

We note that computing the set of *all* model parameter values that satisfy a given property is closely related to that of computing *all* control strategies satisfying a given property. In a discrete time setting this problem has been addressed, for piecewise affine systems and safety properties, in [33,2,1,34].

2 Background

Unless otherwise stated, all forthcoming definitions are based on [47,43]. Throughout the paper, we denote with $[n]$ the set $\{1, 2, \ldots, n\}$ of the first n natural numbers and with $\mathbb{R}^+$, $\mathbb{R}^{\geq 0}$ and $\mathbb{R}$ the sets of, respectively, positive, non-negative and all real numbers. We also denote with $(\mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0})^*$ the set of pairs $(a, b) \in \mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0}$ such that $a \geq b$.

2.1 Parametric Dynamical Systems

We model biological systems using dynamical systems. Usually, a dynamical system comes equipped with a function space that models both *controllable* (e.g., treatments) and *uncontrollable* inputs (*disturbances*). Here, we do not address treatments or disturbances and accordingly we omit inputs from Def. 1.

Definition 1 (Parametric Dynamical System). *A* Parametric Dynamical System *(or, simply, a Dynamical System) $\mathcal{S}$ is a tuple $(\mathcal{X}, \mathcal{Y}, \Lambda, \varphi, \psi)$, where:*

- $\mathcal{X} = X_1 \times \ldots \times X_n$ *is a non-empty set of* states *(*state space *of $\mathcal{S}$);*
- $\mathcal{Y} = Y_1 \times \ldots \times Y_p$ *is a non-empty set of* outputs *(*output value space*);*
- Λ *is a non-empty set of* parameters *(*parameter value space*);*
- $\psi : \mathbb{R}^{\geq 0} \times \mathcal{X} \to \mathcal{Y}$ *is the* observation function *of $\mathcal{S}$;*
- $\varphi : (\mathbb{R}^{\geq 0} \times \mathbb{R}^{\geq 0})^* \times \mathcal{X} \times \Lambda \to \mathcal{X}$ *is the* transition map *of $\mathcal{S}$. Intuitively, $\varphi(t_2, t_1, x, \lambda)$ is the state reached by the system (with parameter values λ) at time t_2 starting from the state $x \in \mathcal{X}$ at time t_1.*

Remark 1. To simplify notation, unless otherwise stated, we assume that the set of parameters Λ has the form $\mathcal{X} \times \Gamma$ (where Γ is a non-empty set). Therefore, a parameter $\lambda = (x_0, \gamma) \in \Lambda$ embodies information about the initial state x_0 of a system trajectory. Such a system trajectory is a function of time $x(\lambda)(t)$, which, for each $t \in \mathbb{R}^{\geq 0}$, evaluates to $\varphi(t, 0, x_0, \gamma)$. In the following, abusing notation, we write $x(\lambda, t)$ instead of $x(\lambda)(t)$. Analogously, we write $x_i(\lambda, t)$ $[y_i(\lambda, t)]$ for the time evolution $x_i(\lambda)(t)$ $[y_i(\lambda)(t)]$ of the i^{th} state [output] component with parameters γ starting in x_0 from time 0.

Example 1. Dynamical systems whose dynamics is described by a system of Ordinary Differential Equations (ODEs) depending on parameters are currently of great interest as a mathematical model for biological networks (see, e.g., [15,39]). In this paper, we will use as a case study the *GynCycle* model presented in [39]. It is a ODE model for the feedback mechanisms between Gonadotropin-Releasing Hormone (GnRH), Follicle-Stimulating Hormone (FSH), Luteinizing Hormone (LH), development of follicles and corpus luteum, and the production of Estradiol (E2), Progesterone (P4), Inhibin A (IhA), and Inhibin B (IhB) during the female menstrual cycle. The model aims at predicting blood concentrations of LH, FSH, E2, and P4 during different stages of the menstrual cycle. The model is intended as a tool to help in preparing and monitoring clinical trials with new drugs that affect GnRH receptors (*quantitative and systems pharmacology*).

In our *black-box* approach, the system transition map models our call to a solver (namely, *Limex* [13]) computing a solution to the ODEs defining our dynamical system. This is along the lines of simulation based system level formal verification as in [29,31,30,32,46,10].

2.2 Biological admissibility

In general, given a value λ for the (vector of) model parameters, it takes a domain expert to decide if a time evolution $x(\lambda, t)$ is *biologically meaningful*. Indeed, many parameter values lead to time evolutions for the model species that are not compatible with the laws of biology. Our goal is to build a general purpose tool that automatically filters out biologically meaningless parameter values. Following [47], we provide a formal criterion for biological admissibility, by asking that the time evolution of $x(\lambda, t)$ is *similar enough* to that of $x(\lambda_0, t)$, that is the one entailed by the model default parameter value λ_0. To this end, we introduce three measures of how similar two trajectories are.

Given a function f from $\mathbb{R}$ to $\mathbb{R}$ and $\alpha, \tau \in \mathbb{R}$, we denote with $f^{\alpha,\tau}$ the function defined by $f^{\alpha,\tau}(t) = f(\alpha(t+\tau))$ for all t. Here, α and τ are used to model, respectively, a stretch and a shift of f. Given two functions f and g from $\mathbb{R}$ to $\mathbb{R}$, the *cross-correlation* (see, e.g., [48]) $\langle f, g\rangle(\xi)$ between f and g is a function of ξ (where $\xi \in \mathbb{R}$ is the *time lag*) defined as: $\langle f, g\rangle(\xi) = \int_{-\infty}^{+\infty} f(t)g(t+\xi)\mathrm{d}t$. We consider the *normalised zero-lag cross-correlation* function $\rho_{f,g}$, defined as $\rho_{f,g} = \frac{\langle f,g\rangle(0)}{\|f\|\|g\|}$, where, for any f, $\|f\|$ is the L^2 norm of f, i.e., $\sqrt{\langle f, f\rangle(0)}$. The higher $\rho_{f,g}$ the more *similar* f and g (e.g., f and g have the same peaks). In particular, $\rho_{f,g}$ is 1 if f is equal to g up to an amplification factor.

Let $\mathcal{S}$ be dynamical system with n state variables and a default parameter value λ_0. Given a parameter value λ and a finite horizon $h \in \mathbb{R}^{\geq 0}$, let $x_i(\lambda_0, t)$ and $x_i(\lambda, t)$ be the time evolutions of species x_i (for each $i \in [n]$) under parameters λ_0 and λ respectively. Being time evolutions, both $x_i(\lambda_0, t)$ and $x_i(\lambda, t)$ are defined for $0 \leq t \leq h$. Anyway, to easily match the above general definition of cross-correlation, we define such functions on the whole set of real numbers, as being 0 for any $t < 0$ or $t > h$. In order to model biological admissibility, we define the following three functions (i ranges over $[n]$, $\alpha, \tau \in \mathbb{R}$):

$$\rho_{\lambda_0,\lambda,i}(\alpha,\tau) = \rho_{x_i(\lambda_0),x_i^{\alpha,\tau}(\lambda)} \qquad \mu_{\lambda_0,\lambda,i}(\alpha,\tau) = \left|\frac{\int_0^h (x_i(\lambda_0,t) - x_i^{\alpha,\tau}(\lambda,t))\mathrm{d}t}{\int_0^h x_i(\lambda_0,t)\mathrm{d}t}\right|$$

$$\chi_{\lambda_0,\lambda,i}(\alpha) \;=\; \left|(\|x_i(\lambda_0)\|^2 - \|x_i^{\alpha,\tau}(\lambda)\|^2)\right| \;/\; \|x_i(\lambda_0)\|^2$$

The *normalised zero-lag cross-correlation* $\rho_{\lambda_0,\lambda,i}(\alpha,\tau)$ measures the similarity of the trajectories $x_i(\lambda_0, t)$ and $x_i(\lambda, t)$ as for qualitative aspects (for example, if they have the same peaks), when $x_i(\lambda, t)$ is subject to stretch α and time-shift τ. The *normalised average differences* $\mu_{\lambda_0,\lambda,i}(\alpha,\tau)$ and the *normalised squared norm differences* $\chi_{\lambda_0,\lambda,i}(\alpha,\tau)$ are two measures of the average distance between $x_i(\lambda_0, t)$ and $x_i(\lambda, t)$, when $x_i(\lambda, t)$ is subject to stretch α and time-shift τ.

In Def. 2, we use these functions to formalise the notion of Biologically Admissible (BA) parameter λ with respect to a default parameter λ_0. Intuitively, λ is BA if the three measures above are all above or below certain thresholds.

Definition 2 (Biologically Admissible parameter). *Let λ_0, $\lambda \in \mathcal{X} \times \Lambda$ be two parameters. Let $\mathbb{A} \subseteq \mathbb{R}^+$, $\mathbb{B} \subseteq \mathbb{R}$ be two sets of real numbers such that $1 \in \mathbb{A}$ and $0 \in \mathbb{B}$. Given a tuple $\Theta = (\theta_1, \theta_2, \theta_3)$ of positive real numbers, we say that λ is Θ-*biologically admissible *with respect to λ_0, notation $\mathrm{adm}_{\mathbb{A},\mathbb{B}}(\lambda_0, \lambda, \Theta)$, if there exist $\alpha \in \mathbb{A}$ and $\tau \in \mathbb{B}$ such that, for all $i \in [n]$: $(\rho_{\lambda_0,\lambda,i}(\alpha,\tau) \geq \theta_1) \wedge (\mu_{\lambda_0,\lambda,i}(\alpha,\tau) \leq \theta_2) \wedge (\chi_{\lambda_0,\lambda,i}(\alpha,\tau) \leq \theta_3)$.* □

3 Computation of Admissible Parameters

Our goal is to compute the set S of (with high confidence) all Biologically Admissible (BA) parameter values with respect to a default parameter value λ_0 validated by the model designer as biologically meaningful.

Since small differences in values are meaningless from a biological point of view, we consider a (grid-shaped) *discretised parameter space* $\hat{\Lambda}$ that is a finite subset of the set of possible parameter values Λ. An exhaustive search on $\hat{\Lambda}$ would be unfeasible, due to the large number of parameters to identify (75 in our case study) that makes $\hat{\Lambda}$ huge (10^{75} elements if we consider 10 possible values for each parameter). To overcome such an obstruction, we follow an approach inspired by Statistical Model Checking (SMC) [18,17]. Statistical Hypothesis Testing is used in [47] to compute, with high statistical confidence, the set S of all BA values with respect to a default value λ_0 for the model parameters.

Given arbitrary values in $(0, 1)$ for ε (probability threshold) and δ (confidence threshold), the SMC algorithm in [47] computes the set S of BA parameters by randomly sampling the discretised parameter space $\hat{\Lambda}$ and adding to S those parameter values $\lambda \in \hat{\Lambda}$ which are shown (by simulation) to be BA. The algorithm terminates when set S remains unchanged after $N = \lceil \ln \delta / \ln(1 - \varepsilon) \rceil$ attempts. At this point, following [18], in [47] it is proved, that, with statistical confidence $1 - \delta$, the probability that the sampling process will extract a BA parameter vector value not already in S is less than ε.

Unfortunately, the SMC algorithm proposed in [47] *cannot* be extended to work in a parallel context efficiently, as too much synchronisation would be required. Here, we define a new random sampling process at the basis of our SMC approach, which enables massive parallelisation of the computation of S. Our new algorithm has been explicitly designed as to be easily deployed on a cluster of heterogeneous multi-core machines connected by a network.

3.1 Algorithm Outline

An overall high-level view of our algorithm deployed on multiple machines connected by a network is shown in Fig. 1. The parallel algorithm that we present here consists of one *Orchestrator* and many *BA Verifiers*.

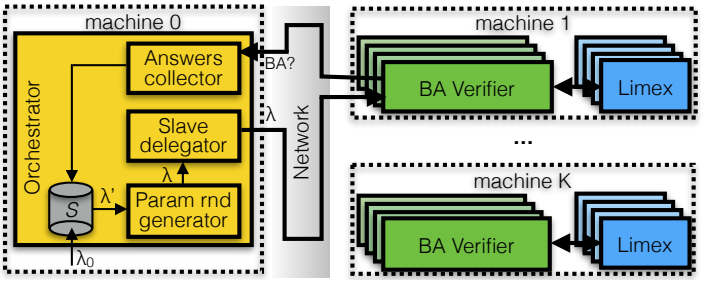

Fig. 1: Parallel algorithm Architecture.

Orchestrator The orchestrator initialises the set S of BA parameter values to the singleton set $\{\lambda_0\}$. Then, at each iteration, it randomly chooses N parameter

values $\lambda_1, \ldots, \lambda_N \in \hat{\Lambda}$ independently (where $N = \lceil \ln \delta / \ln(1-\varepsilon) \rceil$) and delegates the verification of each of them to an idle BA Verifier. After having collected all the N answers, the Orchestrator adds to the set S those parameter values returned as BA. If S changes (i.e., at least one of the N randomly generated parameters is BA and not already in S), a new round of this process starts, otherwise the set S computed so far is returned as the final result.

The sampling space $\hat{\Lambda}$ is given by the set of discretised values for the model parameters. Our sampling strategy (Sect. 3.2) guarantees that any parameter value $\lambda \in \hat{\Lambda}$ can be extracted with non-zero probability, as required by [18,47]. To speed up our procedure, we give a higher probability to parameter values that differ from some parameters already in S for a small number of components.

Note that, at each iteration, the Orchestrator adds up to N parameters to S. Thus, increasing the number of parallel BA Verifiers helps a faster growth of the set of BA parameter values S.

BA Verifiers Each BA Verifier repeatedly takes a parameter λ as input from the Orchestrator, checks if it is Biologically Admissible, and sends back the answer (consisting of the result and the parameter value) to the Orchestrator. To check whether parameter λ is admissible, the BA Verifier in charge runs its *own* instance of the *Limex* solver to compute the time evolutions of all species under parameter λ and checks whether the normalised zero-lag cross-correlation, the normalised average differences, and the normalised squared norm differences for all species are above or below the given thresholds $\Theta = (\theta_1, \theta_2, \theta_3)$, as prescribed by Def. 2.

We observe that the computation distributed to the BA Verifiers is the heaviest part, since it entails to numerically solve the system of differential equations (for a given discretisation of the time output period $[0, h]$ into a finite set T of time-points) and to compute the functions defined in Sect. 2.2 by numerical integration. In order to speed up their computation, BA Verifiers invoke the *Limex* solver just once for each parameter value λ: given the requested finite output time set T and the sets $\mathbb{A}$ and $\mathbb{B}$ for the allowed stretch and time-shift factors, they simulate the system $\mathcal{S}$ computing the trajectory $(t, x(\lambda, t))$ for all time points in a set $T_{\mathbb{A},\mathbb{B}}$ defined as $T \cup \{t' \mid t' = \alpha(t+\tau), t \in T, \alpha \in \mathbb{A}, \tau \in \mathbb{B}\}$. The set $T_{\mathbb{A},\mathbb{B}}$ contains all time instants in which species values are to be known in order to evaluate whether parameter λ satisfies Def. 2.

3.2 Parameter Probability Space

The probability distribution over the discretised parameter space $\hat{\Lambda}$ used by the Orchestrator to generate new parameter values to examine is parametric to the set S of BA parameter values found so far. To speed up the finding of new BA parameter values (with respect to, e.g., uniform sampling), parameter values that are close to those in S are most likely to be chosen.

Given a set S, we extract the N values $\lambda_1, \ldots, \lambda_N$ to examine at each iteration of the Orchestrator independently as follows. For all $i \in [1, N]$: 1) We randomly choose $\lambda'_i \in S$ considering a uniform probability distribution over S. 2) We randomly choose the maximum number h_i of components in which λ_i will differ from λ'_i. In this case, the set $[n]$ is considered distributed as a power-law of the

form $\mathbf{Pr}[h] = a{h_i}^{-b}$, with $b > 1$ and a being a normalisation constant. This implies that, with high probability, λ_i will differ from λ'_i in a small number of components. 3) We randomly choose a subset H_i of h_i different components in $[n]$, assuming a uniform distribution over the set of subsets of cardinality h_i. 4) Finally, the parameter value λ_i is such that for all $j \in H_i$ $\lambda_{i,j}$ is choosen in $\hat{\Lambda}_j$ uniformly at random and $\lambda_{i,j} = \lambda'_{i,j}$ for all $j \in [n] \setminus H_i$.

This sampling technique defines a probability space $(\hat{\Lambda}, \mathcal{P}(\hat{\Lambda}), \mathbf{Pr}^S)$ parametric with respect to a set $S \subseteq \hat{\Lambda}$. By multiplying the (conditional) probabilities of steps 1)–4) above, we have: $\mathbf{Pr}^S[\lambda] = \frac{1}{|S|} \sum_{\lambda' \in S} a\, |d(\lambda, \lambda')|^{-b} \binom{n}{|d(\lambda,\lambda')|}^{-1} \prod_{i \in d(\lambda,\lambda')} \frac{1}{|\hat{\Lambda}_i|}$, where $d(\lambda, \lambda')$ is the set of the components on which λ and λ' differ. Note that $\mathbf{Pr}^S[\lambda]$ is non-zero for all λ.

3.3 Algorithm Correctness

The guarantee that, upon termination, with high statistical confidence, all BA parameter values are in S depends only on the fact that the sampling process consecutively fails N times to find a BA parameter value outside S, and not on how the set S has been populated in the previous iterations of the algorithm.

Stemming from the above considerations, we show the following theorem, stating the correctness of our parallel algorithm.

Theorem 1. *Given a dynamical system $\mathcal{S}$ as in Def. 1, a finite subset $\hat{\Lambda}$ of Λ, a value $\lambda_0 \in \hat{\Lambda}$, a tuple Θ of biological admissibility thresholds, two real numbers ε and δ in $(0, 1)$, and two finite sets of real numbers $\mathbb{A}$ and $\mathbb{B}$ (with $1 \in \mathbb{A}$ and $0 \in \mathbb{B}$), our parallel algorithm is such that:*

1. it terminates;

2. upon termination, it computes a set $S \subseteq \hat{\Lambda}$ of Θ-Biologically Admissible parameter values;

3. with confidence $1 - \delta$: $\mathbf{Pr}^S[\{\lambda \in \hat{\Lambda} \setminus S \mid adm_{\mathbb{A},\mathbb{B}}(\lambda_0, \lambda, \Theta)\}] < \varepsilon$. □

4 Experimental Results

The computational effectiveness of our distributed multi-core implementation has been evaluated on the *GynCycle* model [39]. Such a model has 114 parameters, 75 of which are patient-specific (at least for our purposes), and consists of 41 differential equations defining the time evolution of 33 species.

We implemented our tool in the C programming language using Message Passing Interface (MPI) [42] to enable the communication between the Orchestrator and BA Verifiers spread on multiple machines connected by a network.

4.1 Experimental setting

Experiments have been carried out on a cluster of 7 Linux heterogeneous machines: 1 machine equipped with 2 × Intel(R) Xeon(R), 2.83 GHz and 8GB of RAM (category A), 2 machines equipped with 2 × Intel(R) Xeon(R), 2.66 GHz and 8GB of RAM (cat. B), and 4 machines equipped with 2 × Intel(R) Xeon(R), 2.27 GHz and 16GB of RAM (cat. C). We used a maximum number of 81 CPU

cores (7 out of the 8 available cores for machines of categories A and B and 15 out of the 16 available cores for machines of cat. C). The single Orchestrator process was always run on a core of the machine in cat. A.

We set both ε and δ to 10^{-3}. The stretch factor α (see Def. 2 in Sect. 2.2) ranges in the set $\mathbb{A} = \{0.90, 0.95, 1.00, 1.05, 1.10\}$, while the set $\mathbb{B}$ for the shift factor τ (see Def. 2 in Sect. 2.2) consists of all values from -3 to 3 days multiple of 6 hours. The discretisation $\hat{\Lambda}$ of Λ has been obtained by uniformly discretising the range of each parameter into 5 values. We set Limex to compute time evolutions for all species over $h = 90$ days, returning values with a time step of 15 minutes. Integrals for cross-correlation and norms have been computed numerically with a time step of 15 minutes.

In [47], suitable values for the biological admissibility thresholds $\theta_1, \theta_2, \theta_3$ have been considered, in order to largely cover the set of model meaningful biological behaviours. Here we are interested in evaluating the *speedup* and the *efficiency* of our distributed multi-core algorithm. To this end, in order to execute multiple experiments in reasonable time, we set the biological admissibility thresholds $\theta_1, \theta_2, \theta_3$ to, respectively, 0.99, 0.01, 0.01. Such values are way overly restrictive from a biological point of view, and allow us to compute only a *tiny fraction* (only 8 parameter values) of the set of the BA parameters shown in [47] (which consists of several thousands of Biologically Admissible (BA) parameter values). Anyway, the overall number of random parameter values generated and examined in our case (27620) is sufficiently large to let us correctly evaluate the computational performance of our algorithm.

4.2 Experimental results

Table 1 shows the overall computation time (column "*time*") when varying the number of BA Verifiers (col. "*# proc.*") used in parallel by our algorithm. Each BA Verifier runs on a *different* core of a machine in our cluster. To make the different values comparable (given the stochastic nature of our algorithm and the heterogeneity of our cluster machines), we started all runs using the *same* random seed and used the *same* proportion of machines of each category in all runs (col. "*# cores*"). To neutralise biases due to the heterogeneity of our cluster machines, we determined the computation time of our algorithm when using a *single* BA Verifier (sequential time) by carrying out three runs allocating the (single) BA Verifier on a core of a machine of each category. Such computation times are listed in Table 2. From such data we have computed the completion time in the first line of Table 1 by averaging the three sequential execution times, using the proportion of the number of cores for each machine category as weights.

Column "*speedup*" in Table 1 shows the speedup achieved by our algorithm. For each number v of parallel BA Verifiers, the speedup is the ratio t_v/t_1, where t_v and t_1 are the computation times shown in Table 1 when using, respectively, v and 1 BA Verifiers. Column "*eff.*" shows the efficiency of our algorithm and is computed, as typically done in the evaluation of parallel algorithms, by dividing the speedup by the number of the parallel BA Verifiers used.

From Table 1 we can see that our distributed multi-core implementation scales well with the number of used parallel BA Verifier instances. The observed lack of efficiency, mostly due to network delays, is typical in a cluster setting. We

#proc.	# cores			time	speedup	eff.
	A	B	C	(h:m:s)		
1	–	–	–	*238:16:55*	1×	100%
26	2	4	20	9:16:57	25.67×	98.73%
52	4	9	39	5:16:25	45.18×	86.88%
80	6	14	60	4:1:12	59.27×	74.09%

Table 1: Computation times

machine cat. for the sequential alg.	time (h:m:s)
A	194:47:45
B	206:19:15
C	250:5:18

Table 2: Sequential time

note that high-performance parallel simulation typically has efficiency values in the range 40%-80% (e.g., see [36]). Accordingly, an efficiency of 74% (last row of Table 1) is to be considered state-of-the-art.

5 Conclusions

We presented a parallel algorithm which efficiently computes the set of Biologically Admissible (BA) parameters for an ODE-based biological model. In our approach, this is a crucial step to enable fast computation of patient-specific predictions from clinical trials. The main ingredient of our parallel algorithm is a novel random sampling process which allows the parallel execution of an arbitrarily high number of processes to check their biological admissibility (which is the most computationally demanding part). Such processes are independent and communicate only with an orchestrator. Our results show that our distributed multi-core implementation scales well with the number of available cores.

References

1. V. Alimguzhin, F. Mari, I. Melatti, I. Salvo, and E. Tronci. A map-reduce parallel approach to automatic synthesis of control software. In *Proc. of SPIN*, volume 7976 of *LNCS*, pages 43–60, 2013.
2. V. Alimguzhin, F. Mari, I. Melatti, I. Salvo, and E. Tronci. On-the-fly control software synthesis. In *Proc. of SPIN*, volume 7976 of *LNCS*, pages 61–80, 2013.
3. M. AlTurki and J. Meseguer. Pvesta: A parallel statistical model checking and quantitative analysis tool. In *Proc. of CALCO, LNCS 6859*, pages 386–392, 2011.
4. P. Ballarini, M. Forlin, T. Mazza, and D. Prandi. Efficient parallel statistical model checking of biochemical networks. In *Proc. of PDMC, EPCTS 14, p. 47–61*, 2009.
5. P. Ballarini, R. Guido, T. Mazza, and D. Prandi. Taming the complexity of biological pathways through parallel computing. *Briefings in Bioinformatics*, 10(3):278–288, 2009.
6. E. Balsa-Canto, M. Peifer, J. R. Banga, J. Timmer, and C. Fleck. Hybrid optimization method with general switching strategy for parameter estimation. *BMC Systems Biology*, 2:26, 2008.
7. J. Barnat, L. Brim, I. Černá, S. Dražan, and D. Šafránek. Parallel model checking large-scale genetic regulatory networks with DiVinE. *ENTCS*, 194(3):35–50, 2008.
8. J. Barnat, L. Brim, D. Šafránek, and M. Vejnár. Parameter scanning by parallel model checking with applications in systems biology. In *Proc. of HiBi/PDMC*, pages 95–104. IEEE, 2010.
9. O-T. Chis, J. R. Banga, and E. Balsa-Canto. Structural identifiability of systems biology models: A critical comparison of methods. *PLoS ONE*, 6(11), 2011.
10. G. Della Penna, B. Intrigila, E. Tronci, and M. Venturini Zilli. Synchronized regular expressions. *Acta Inf.*, 39(1):31–70, 2003.

11. T. Dierkes, S. Röblitz, M. Wade, and P. Deuflhard. Parameter identification in large kinetic networks with BioPARKIN. *CoRR*, abs, 2013.
12. R. Donaldson and D. Gilbert. A model checking approach to the parameter estimation of biochemical pathways. In *Proc. of 6th CMSB 2008, LNCS 5307*, 2008.
13. R. Ehrig, U. Nowak, L. Oeverdieck, and P. Deuflhard. Advanced extrapolation methods for large scale differential algebraic problems. In *High Performance Scient. and Eng. Comp.*, LNCSE, 1999.
14. H. Gong, P. Zuliani, A. Komuravelli, J. R. Faeder, and E. M. Clarke. Analysis and verification of the hmgb1 signaling pathway. *BMC Bioinform.*, 11(S-7):S10, 2010.
15. H. Gong, P. Zuliani, A. Komuravelli, J. R. Faeder, and E. M. Clarke. Computational modeling and verification of signaling pathways in cancer. In *Proc. of 4th ANB*, volume 6479, pages 117–135, 2010.
16. H. Gong, P. Zuliani, Q. Wang, and E. M. Clarke. Formal analysis for logical models of pancreatic cancer. In *Proc. of 50th CDC*, pages 4855–4860. IEEE, 2011.
17. R. Grosu and S. A. Smolka. Quantitative model checking. In *Preliminary Proc. of ISoLA*, pages 165–174, 2004.
18. R. Grosu and S. A. Smolka. Monte carlo model checking. In *Proc. of TACAS*, pages 271–286, 2005.
19. J. Heath, M. Z. Kwiatkowska, G. Norman, D. Parker, and O. Tymchyshyn. Probabilistic model checking of complex biological pathways. *Theor. Comput. Sci.*, 391(3):239–257, 2008.
20. 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. In *Proc. of 2nd ICCABS*, pages 1–6. IEEE, 2012.
21. B. Ingalls and P. Iglesias. *Control Theory and Systems Biology*. MIT Press, 2009.
22. H. De Jong. Modeling and simulation of genetic regulatory systems: A literature review. *Journal of Computational Biology*, 9:67–103, 2002.
23. M. Kwiatkowska, G. Norman, and D. Parker. Using probabilistic model checking in systems biology. *ACM SIGMETRICS Perf. Eval. Rev.*, 35(4):14–21, 2008.
24. C. J. Langmead. Generalized queries and bayesian statistical model checking in dynamic bayesian networks: Application to personalized medicine. In *Proc. of CSB*, pages 201–212, 2009.
25. K. Levenberg. A method for the solution of certain non-linear problems in least squares. *The Quarterly of Applied Math*, 2:164–168, 1944.
26. Pu Li and Quoc D. Vu. Identification of parameter correlations for parameter estimation in dynamic biological models. *BMC Systems Biology*, 7(1):91+, 2013.
27. Lennart Ljung. *System Identification (2Nd Ed.): Theory for the User*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 1999.
28. S. Stahl M. Brusco. *Branch-and-Bound Applications in Combinatorial Data Analysis*. Statistics and Computing. Springer, 2005.
29. T. Mancini, F. Mari, A. Massini, I. Melatti, F. Merli, and E. Tronci. System level formal verification via model checking driven simulation. In *Proc. 25th CAV*, volume 8044 of *LNCS*, pages 296–312, 2013.
30. T. Mancini, F. Mari, A. Massini, I. Melatti, and E. Tronci. Anytime system level verification via random exhaustive hardware in the loop simulation. In *Proc. of DSD*, pages 236–245, 2014.
31. T. Mancini, F. Mari, A. Massini, I. Melatti, and E. Tronci. System level formal verification via distributed multi-core hardware in the loop simulation. In *Proc. of PDP*, 2014.
32. T. Mancini, F. Mari, A. Massini, I. Melatti, and E. Tronci. SyLVaaS: System level formal verification as a service. In *Proc. of PDP*. IEEE, 2015.

33. F. Mari, I. Melatti, I. Salvo, and E. Tronci. Synthesis of quantized feedback control software for discrete time linear hybrid systems. In *Proc. of 23rd CAV*, volume 6174 of *LNCS*, pages 180–195, 2010.
34. F. Mari, I. Melatti, I. Salvo, and E. Tronci. Model based synthesis of control software from system level formal specifications. *ACM TOSEM*, 23(1):1–42, 2014.
35. N. Miskov-Zivanov, P. Zuliani, E. M. Clarke, and J. R. Faeder. Studies of biological networks with statistical model checking: Application to immune system cells. In *Proc. of BCB*, pages 728–729. ACM, 2007.
36. J. C. Phillips, Y. Sun, N. Jain, E. J. Bohm, and L. V. Kalé. Mapping to irregular torus topologies and other techniques for petascale biomolecular simulation. In *Proc. of SC14*, pages 81–91. IEEE, 2014.
37. A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, 25(15):1923–1929, 2009.
38. A. Rizk, G. Batt, F. Fages, and S. Soliman. On a continuous degree of satisfaction of temporal logic formulae with applications to systems biology. In *Proc. of 6th CMSB*, pages 251–268, 2008.
39. S. Röblitz, C. Stötzel, P. Deuflhard, H. M. Jones, D.-O. Azulay, P. van der Graaf, and S. W. Martin. A mathematical model of the human menstrual cycle for the administration of GnRH analogues. *Journ. of Theor. Biology*, 321:8–27, 2013.
40. S. Sebastio and A. Vandin. Multivesta: statistical model checking for discrete event simulators. In *Proc. of ValueTools*, pages 310–315, 2013.
41. K. Sen, M. Viswanathan, and G. Agha. On statistical model checking of stochastic systems. In *Proc. of CAV*, volume 3576 of *LNCS*, pages 266–280, 2005.
42. M. Snir, S. Otto, S. Huss-Lederman, D. Walker, and J. Dongarra. *MPI-The Complete Reference, Vol. 1: The MPI Core*. MIT Press, 2nd edition, 1998.
43. Eduardo D. Sontag. *Mathematical Control Theory: Deterministic Finite Dimensional Systems. (2nd Edition)*. Springer, New York, 1998.
44. A. Streck, A. Krejci, L. Brim, J. Barnat, D. Safranek, M. Vejnar, and T. Vejpustek. On parameter synthesis by parallel model checking. *IEEE/ACM Trans. on Comput. Biology and Bioinf.*, 9(3):693–705, 2012.
45. J. Sun, J. M. Garibaldi, and C. Hodgman. Parameter estimation using metaheuristics in systems biology: A comprehensive review. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 9(1):185–202, 2012.
46. E. Tronci, T. Mancini, F. Mari, I. Melatti, I. Salvo, M. Prodanovic, J. K. Gruber, B. Hayes, and L. Elmegaard. Demand-aware price policy synthesis and verification services for smart grids. In *SmartGridComm*, pages 236–245. IEEE, 2014.
47. E. Tronci, T. Mancini, I. Salvo, S. Sinisi, F. Mari, I. Melatti, A. Massini, F. Davì, T. Dierkes, R. Ehrig, S. Röblitz, B. Leeners, T. H. C. Krüger, M. Egli, and F. Ille. Patient-specific models from inter-patient biological models and clinical records. In *Proc. of FMCAD*, pages 207–214, 2014.
48. S. V. Vaseghi. *Advanced Digital Signal Processing and Noise Reductio*. John Wiley & Sons, 2006.
49. D. J. Wilkinson. *Stochastic Modelling for Systems Biology*. Chapman & Hall, 2006.
50. P. Zuliani, A. Platzer, and E. M. Clarke. Bayesian statistical model checking with application to Stateflow/Simulink verification. *Formal Methods in System Design*, 43(2):338–367, 2013.