2011 50th IEEE Conference on Decision and Control and
European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Model Selection in Stochastic Chemical Reaction Networks Using Flow Cytometry Data

Gabriele Lillacci and Mustafa Khammash

Abstract— **The model selection problem, that is picking the model that best explains an experimental data set from a list of candidates, arises frequently when studying unknown biological processes. Here, we propose a new method for model selection in stochastic chemical reaction networks using measurements from flow cytometry. A distinctive feature of our approach is its ability to perform statistically significant selection using a very small number of Monte Carlo simulations of the candidate stochastic models. After a comprehensive review of the theory associated with our procedure, we describe the model selection algorithm and we demonstrate it on an example drawn from molecular biology.**

I. INTRODUCTION

When studying an unknown biological pathway, investigators commonly face the problem of formulating and testing competing hypotheses on its components and the nature of their interactions. For example, it might be uncertain whether a certain reaction occurs with or without intermediate steps, or whether a certain species is positively or negatively regulated by another one. In this type of scenario, one could write several different models corresponding to the different hypotheses, collect experimental data from the biological process, and then devise a computational procedure to *select* the model that best explains the data.

This issue, which is known in the literature as the *model selection problem*, has been addressed for both deterministic and stochastic computational biology models with a variety of techniques. Classical approaches include the Akaike Information Criterion (AIC) and the subsequent developments stemming from it [1]. These methods require the knowledge of the likelihood function and the selection step is usually applied after some kind of estimation (usually maximum-likelihood estimation) of the model parameters.

Bayesian model selection has also been given a great deal of attention [2], [3], especially thanks to the recent advancements in the computational methods for its implementation [4]. The Bayesian approach has the attractive feature of evaluating the model structures by simultaneously taking into account all the possible value of the parameters. However, its application can limited by considerations of computational feasibility, therefore this type of analysis usually performs best in the case of small-dimensional problems.

This material is based upon work supported by the National Science Foundation under Grant NSF-CDI ECCS-0835847, NSF-ECCS-0802008, NIH R01GM049831, and the Institute for Collaborative Biotechnologies through Grant DAAD19-03-D-0004 from the U.S. Army Research office.

G. Lillacci and M. Khammash are with the Center for Control, Dynamical Systems and Computation, University of California at Santa Barbara, Santa Barbara, CA, USA. gabriele@engr.ucsb.edu, khammash@engr.ucsb.edu

In our previous work [5], we introduced a model selection criterion for deterministic systems based on the statistical comparison of moments of Gaussian random variables. Here, we extend this approach to stochastic chemical reaction networks [6]. Our proposed method achieves statistically significant model selection with a surprisingly small number of Monte Carlo simulations: it is, therefore, suitable for the comparison of very large models and/or of a very large number of models.

We specifically focus on model selection using data from flow cytometry. This increasingly popular experimental technique is based on measuring emission intensity from fluorescently labeled species (usually proteins) in individual cells. The cells pass one by one through a detection system, where the fluorochromes are excited by one or more lasers, and their emission at specific wavelengths is recorded.

The rest of the paper is organized as follows. In Section II, we state the model selection problem in more detail by introducing suitable notation. In Section III, we first review the theoretical foundation of our proposed method, and then we introduce the model selection algorithm. In Section IV, we demonstrate the method on an example drawn from molecular biology. Finally, in Section V, we summarize our findings and present some final remarks.

II. PROBLEM STATEMENT

Consider a biological process in which N chemical species interact through R reaction channels, and let Σ_q, $q = 1, \dots, Q$ be stochastic chemical reaction networks (SCRNs) which model the process. We assume that eventual unknown parameters in the SCRNs were estimated using an *independent* experiment (and, possibly, a dedicated computational procedure [7]).

The process is observed at K discrete time points t_k, $k = 1, \dots, K$. At each t_k, a flow cytometry scan is obtained. We assume that each scan yields $P \leq N$ signals Y_{kp}, $p = 1, \dots, P$ that are *proportional* to the abundance at time t_k of some chemical species of interest involved in the process (usually proteins).

When a cell passes through the flow cytometer's optical detection system, the instrument simultaneously records the fluorescence levels of all the P fluorochromes, i.e. a sample from the joint distribution of the P fluorescence levels at time t_k. Since each cell is only used once and then discarded, the samples can regarded as *independent*.

We are interested in working with the *marginal distributions* of the P fluorochromes. Hence, if we denote M_k the number of cells that are scanned at time t_k we can regard

the measurements as M_k independent samples of each of the marginals at time t_k. In other words, the measurements are *snapshots* of the probability density function (PDF) or of the cumulative distribution function (CDF) of the outputs at the time instants $t_1, \ldots, t_K$.

Given this framework, the model selection problem can be stated as follows: (i) discard any SCRN that is not *statistically consistent* with the experimental data; and (ii) pick the one that *explains the data the best* with respect to a suitable measure. The precise meaning of these statements will become clear in Section III, after we review the theory associated with our proposed method.

III. MODEL SELECTION ALGORITHM

In the context of our application, it is more convenient to work with CDFs rather than with PDFs for two main reasons: (i) CDFs can be approximated using nonparametric estimators; (ii) there is a well-developed theoretical machinery for the comparison of CDFs using such estimators. We begin our presentation by reviewing some of these notions. In the following, we will sometimes use the word *distribution* for CDF, and the word *density* for PDF.

A. The empirical cumulative distribution function

Let $X = \{x_m : m = 1, \ldots, M\}$ be a set of *independent and identically distributed* (i.i.d.) samples from an unknown continuous distribution F. We define the *empirical cumulative distribution function*, or *empirical distribution*, or ECDF, associated with the samples X the function:

$$\begin{aligned} \hat{F}_X(x) &= \frac{\text{number of samples } x_m \leq x}{\text{total number of samples}} \\ &= \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left(x_m \leq x\right), \end{aligned} \tag{1}$$

where $\mathbb{I}(A)$ denotes the indicator function of the event A. Intuitively, the ECDF is a staircase approximation of the true CDF that "jumps" $1/M$ corresponding to every occurrence of a data point.

The ECDF $\hat{F}$ has very nice properties as an estimator of the true CDF F. First of all, unlike histograms or kernel density estimates for the approximation of PDFs, it is a *nonparametric estimator*: no arbitrary choice (such as bin size in histograms, or kernel function in kernel density estimates) is required for the construction of $\hat{F}$. Furthermore, as the number of samples M goes to infinity, it converges to F *uniformly in* x. This result, known as *Glivenko-Cantelli theorem* [8], [9], guarantees that $\hat{F}$ is an *unbiased* and *consistent* estimator of F.

B. The Kolmogorov distance

The model selection method we are going to introduce is based on the comparison of CDFs. The key tool for this is a notion of *distance* between two CDFs. A commonly used one is the *Kolmogorov distance*, which is defined as follows. Given two CDFs F and G, with support over the same set $\Omega \in \mathbb{R}$, we write their Kolmogorov distance as:

$$D\left(F, G\right) = \sup_{x \in \Omega} \left|F(x) - G(x)\right| = \left\|F - G\right\|_\infty . \tag{2}$$

Let's now apply (2) to the case when one of the CDFs is an empirical distribution. More specifically, we will look at the case where F is a *continuous* CDF and $\hat{G}$ is an ECDF constructed using a set X of M i.i.d. samples from an unknown continuous CDF G. We have:

$$d_M\left(\hat{G}_X, F\right) = \sup_x \left|\hat{G}_X(x) - F(x)\right|, \tag{3}$$

where the set over which $\hat{G}$ and F have support is understood, and the subscript M reminds that the set X contains M samples. We realize that as the random samples in X change, $\hat{G}_X$ changes as well, and therefore $d_M\left(\hat{G}_X, F\right)$ defines a *random variable*.

Suppose we now want to compare F and G. We look for conditions that allow us to verify with any desired confidence level whether F and G are the same distribution or not. Under the null hypothesis $\mathcal{H}_0 : F = G$, the Glivenko-Cantelli theorem implies that, as $M \to \infty$, $d_M \to D = 0$ with probability 1. This asymptotic result, although very important in guaranteeing that ECDFs are well-behaved estimators, is not very useful for computational purposes, as it doesn't say anything about the properties of $\hat{G}$ for finite M. The inequality we present next fills this gap.

C. The DKW inequality

An important result, first proved by Dvoretzky, Kiefer and Wolfowitz in 1956 [10], and later strengthened by Massart in 1990 [11], gives a precise characterization of the rate of convergence of an ECDF to the corresponding exact CDF for any finite number of samples. This result is known as the *DKW inequality*.

If we look again at (3), it makes sense to ask what is the distribution of the random variable $d_M\left(\hat{G}_X, F\right)$, and how can we use it to test the null hypothesis $\mathcal{H}_0 : F = G$. Clearly, if $\mathcal{H}_0$ is true, then by (2) and the properties of norms we have $D = 0$. However, due to sampling, d_M will not be exactly 0 for finite M, but it will approach 0 as $M \to \infty$ (note that d_M is a distance, and therefore $d_M \geq 0$ by construction). If we now fix M and look at all the possible ECDFs $\hat{G}_X$ such that the set X contains exactly M samples we may ask what is the probability that d_M is greater than any number ϵ, with $0 < \epsilon < 1$.

The DKW inequality, bounds exactly this probability. If F is a *fully specified* continuous CDF, and $\mathcal{H}_0 : F = G$ is true, then for any set of i.i.d. samples X with $M < \infty$ samples and any $0 < \epsilon < 1$, the following statement is true:

$$P\left\{d_M\left(\hat{G}_X, F\right) > \epsilon\right\} \leq 2e^{-2M\epsilon^2}. \tag{4}$$

This result lends itself to several useful interpretations, depending on the meaning that is attributed to the different quantities in the inequality. The most natural way of looking at the left-hand side of (4) is the *probability of making a*

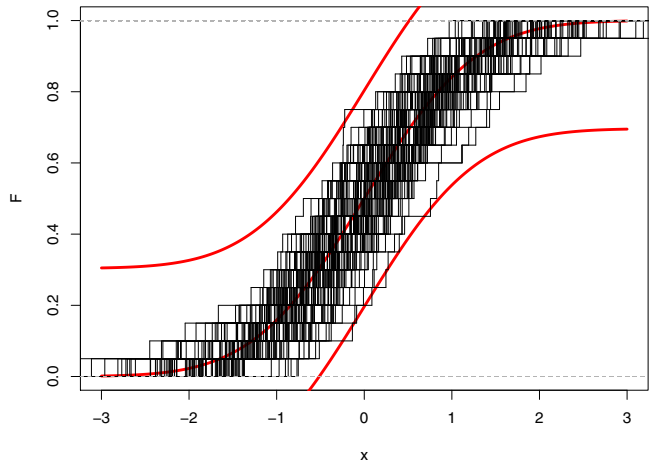

Fig. 1. **Illustration of the DKW bound.** The red lines show the exact CDF and the $\epsilon^{(c)}$-ball around it. The black dotted lines show 100 ECDFs generated using 20 random samples from the exact CDF. Only a very small number of them violates the bound.

mistake in using d_M to test $\mathcal{H}_0$, instead of D. Suppose we decide to accept the null hypothesis when d_M is less than or equal to a certain critical value $\epsilon^{(c)}$, and to reject it otherwise. Then, the DKW inequality (evaluated for $\epsilon = \epsilon^{(c)}$) bounds the probability of *rejecting the null hypothesis when it is true*, that is the *p-value* of the test (Type I statistical error).

We can use the inequality to make the p-value smaller than or equal to a desired *confidence level* α:

$$P\left\{d_M > \epsilon\right\} \leq 2e^{-2M\epsilon^2} \leq \alpha.$$

This can be solved for ϵ and used to set the critical value $\epsilon^{(c)}$. We find:

$$\epsilon \geq \sqrt{-\frac{1}{2M}\log\left(\frac{\alpha}{2}\right)} = \epsilon^{(c)}(\alpha, M). \tag{5}$$

The above expression gives the minimum value of d_M that allows one to conclude that F and G are different distributions with probability $1-\alpha$.

We can also fix α and ϵ and solve for M. We find:

$$M \geq \left\lceil -\frac{1}{2\epsilon^2}\log\left(\frac{\alpha}{2}\right)\right\rceil = M^{(c)}(\alpha, \epsilon), \tag{6}$$

where $\lceil x \rceil$ denotes the smallest integer not less than the real number x, i.e. the *ceiling* of x. This gives the minimum number of samples needed to claim that $d_M \leq \epsilon$ with probability $1-\alpha$.

We conclude this survey with yet another way of looking at (5). If F and G are the same distribution, then for any set X containing M i.i.d. samples, $\hat{G}_X$ is at most $\epsilon^{(c)}(\alpha, M)$ apart from F in the metric d_M, with probability $1-\alpha$. This defines an *infinity norm ball* of radius $\epsilon^{(c)}(\alpha, M)$ centered around F that contains most random ECDFs constructed using M samples from F. This is illustrated in Figure 1.

As a final remark, we note that all the bounds we derived using the DKW inequality, and the inequality itself, *do not depend on the functional form of the exact CDF*. This is one of the most powerful properties of (4), (5) and (6), and it is key to their application to the model selection problem.

D. Comparing experimental CDFs

When observing a stochastic chemical reaction network through flow cytometry, the exact CDF F is not available. What we can measure instead is a set of fluorescence levels Y that are *believed* to be i.i.d. samples from F. Given the set Y, it is straightforward to compute the ECDF associated with it using (1). We denote this empirical distribution $\tilde{F}_Y$. Note that the underlying exact CDF for $\tilde{F}_Y$ is F by assumption.

On the computational side, we have the *simulated* set of fluorescence levels produced by a SCRN. We denote this set X, and the ECDF associated with it $\hat{G}_X$. We denote S the number of simulations in X, and G the underlying exact CDF for $\hat{G}_X$.

The model selection problem can be, then, stated as follows: test the hypothesis $F = G$, i.e. the hypothesis that $\tilde{F}_Y$ and $\hat{G}_X$ have *the same* underlying exact CDF.

Let's now look at the Kolmogorov distance between $\tilde{F}_Y$ and $\hat{G}_X$, i.e. $d_{SM}\left(\hat{G}_X, \tilde{F}_Y\right)$. This quantity *can not* be bounded directly using the DWK inequality, since $\tilde{F}_Y$ and $\hat{G}_X$ are not continuous functions. However, the Kolmogorov distance is a norm, therefore we can apply the triangle inequality and write:

$$\begin{aligned} d_{SM}\left(\hat{G}_X, \tilde{F}_Y\right) &= \left\|\hat{G}_X - \tilde{F}_Y\right\|_\infty \\ &\leq \left\|\hat{G}_X - F\right\|_\infty + \left\|\tilde{F}_Y - F\right\|_\infty \\ &= d_S\left(\hat{G}_X, F\right) + d_M\left(\tilde{F}_Y, F\right). \end{aligned} \tag{7}$$

The left-hand side of (7) can be computed, since both $\tilde{F}_Y$ and $\hat{G}_X$ are known. The two quantities in the right-hand side can be bounded using the DKW inequality, *even if F is unknown*. Under the null hypothesis $\mathcal{H}_0 : F = G$, and for a fixed confidence level α, the following inequality must hold with probability at least $1-\alpha$:

$$\begin{aligned} d_{SM}\left(\hat{G}_X, \tilde{F}_Y\right) \leq \\ \sqrt{-\frac{1}{2S}\log\left(\frac{\alpha}{2}\right)} + \sqrt{-\frac{1}{2M}\log\left(\frac{\alpha}{2}\right)}. \end{aligned} \tag{8}$$

If (8) is violated, then with probability at least $1-\alpha$ the empirical distributions $\tilde{F}_Y$ and $\hat{G}_X$ do not have the same underlying exact CDF. If (8) does hold, then we either have $F = G$, or the number of simulations S is not large enough to conclude otherwise. In order to claim that $d_{SM}\left(\hat{G}_X, \tilde{F}_Y\right) \leq \epsilon$ with probability at least $1-\alpha$, the minimum number of simulations needed is given by:

$$S \geq \left\lceil \frac{-\log\left(\frac{\alpha}{2}\right)}{2\left(\epsilon - \sqrt{-\frac{1}{2M}\log\left(\frac{\alpha}{2}\right)}\right)^2} \right\rceil = S^{(c)}\left(\epsilon, \alpha, M\right). \tag{9}$$

This is illustrated in Figure 2. As expected, as the number of experimental samples M increases, the critical number of simulations $S^{(c)}\left(\epsilon, \alpha, M\right)$ decreases. The graph also shows

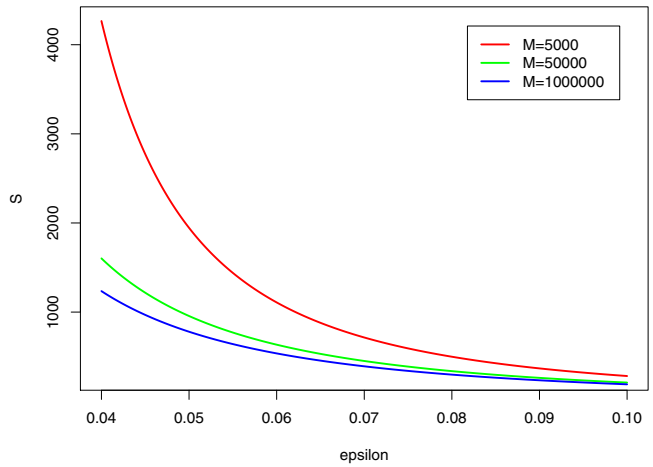

Fig. 2. **Simulations needed for significant comparison of ECDFs.** The three curves show the critical number of simulations $S^{(c)}$ in (9) as a function of ϵ, for $\alpha = 0.05$ and three different values of M.

a clear diminishing returns effect, where very little is gained in increasing M from 50000 (a typical number for flow cytometry) to 1000000 (limits of technical feasibility).

E. Simultaneous comparison of multiple CDFs

In practical flow cytometry experiments, we typically measure $P = 1-4$ species at $K = 10-20$ time points. Hence, the model selection problem involves the simultaneous comparison of KP pairs of CDFs. Note that the first time point is usually used as initial condition for the SCRN simulations, so in most cases the corresponding experimental and simulated ECDFs will match perfectly, up to sampling error.

The simultaneous comparison problem can be handled by combining information from multiple statistical tests on the same null hypothesis into an overall p-value, a procedure known in statistics as *meta-analysis*. In our application, we have $(K-1)P$ CDF equality tests, which provide evidence that either supports or refutes the global null hypothesis that a given SCRN adequately explains the data.

We first obtain an estimate of the p-value for the comparison of a single pair of CDFs. We fix a distance tolerance $\bar{\epsilon}$ and a confidence level α, and we perform $S^{(c)}(\bar{\epsilon}, \alpha, M)$ simulations of a candidate SCRN. With these choices the right-hand side of (8) evaluates precisely to $\bar{\epsilon}$. As established in Section III-D, if an experimental ECDF $\tilde{F}_{Y_{kp}}$ and the corresponding simulated ECDF $\hat{G}_{X_{kp}}$ have the same underlying exact CDF, then the inequality $d_{SM}\left(\hat{G}_{X_{kp}}, \tilde{F}_{Y_{kp}}\right) \leq \bar{\epsilon}$ must hold with probability at least $1-\alpha$. Consequently, the probability that d_{SM} is greater than $\bar{\epsilon}$ with the null hypothesis still being true, i.e. the p-value, is at most α.

Therefore, if for a given pair of CDFs the null hypothesis can be rejected on the basis of (8), we take the associated p-value to be α. On the other hand, if (8) holds, we have no reason to believe that $\hat{G}_{X_{kp}}$ and $\tilde{F}_{Y_{kp}}$ do not have the same underlying CDF, hence the associated p-value is set to 1.

The final step is to combine the individual p-values into an overall p-value for the global null hypothesis. One straightforward way of doing this is *Fisher's combined probability test* [12], which is based on the following statistic:

$$X^2 = -2\sum_{k=2}^{K}\sum_{p=1}^{P}\log\left(\alpha_{kp}\right). \tag{10}$$

We recall that in our application α_{kp} is either 1 or α. Under the null hypothesis, the X^2 statistic has the χ^2 distribution with $2(K-1)P$ degrees of freedom. Therefore, if we denote by $\chi^2_\nu(x)$ the χ^2 PDF with ν degrees of freedom, the p-value for Fisher's test is given by the *tail probability* of this PDF, and can be readily computed as follows:

$$\begin{aligned}\alpha^{(F)} &= \int_{X^2}^{\infty}\chi^2_{2(K-1)P}(\tau)d\tau \\ &= 1-\int_{0}^{X^2}\chi^2_{2(K-1)P}(\tau)d\tau.\end{aligned} \tag{11}$$

It must be noted that Fisher's p-value computed using (11) is exact only if the p-values that are combined into X^2 are independent. If they are correlated, (11) should be viewed as an approximation. As a result the confidence level that is fixed for the individual hypotheses (α in our case), may not be an adequate threshold for $\alpha^{(F)}$. In particular, it is well-known in the literature that if the correlation is positive (as it is usually the case), meta-analysis techniques have a tendency to overstate significance [13], resulting in a more likely rejection of the global null hypothesis.

Since the ranking of the candidate models is not affected (see below), an accurate investigation of this effect is beyond the scope of the present manuscript. However, we do remark that care should be taken when deciding whether or not a given SCRN is consistent with the experimental observations based on $\alpha^{(F)}$, especially if this quantity turns out to be close to the threshold.

F. Model selection algorithm

We are now ready to assemble all the notions we introduced so far into a practical algorithm, which we call *distribution comparison model selection (DCMS)*. The algorithm implements the two steps outlined in Section II as follows.

For each SCRN Σ_q, we compute the Fisher's p-value $\alpha_q^{(F)}$. Any SCRN for which $\alpha_q^{(F)} \ll \alpha$ is discarded as a model that is inconsistent with the experimental data (taking into account the remarks of Section III-E). Then, the candidate models are ranked according to two criteria: (i) the value of Fisher's statistic X^2, and (ii) the cumulative Kolmogorov distance of the simulated CDFs from the experimental CDFs, i.e. the following quantity:

$$\mathcal{D} = \sum_{k=2}^{K}\sum_{p=1}^{P} d_{SM}\left(\hat{G}_{X_{kp}}, \tilde{F}_{Y_{kp}}\right). \tag{12}$$

Among the candidates with the smallest value of X^2, the one with the smallest $\mathcal{D}$ is selected as the model that explains the data the best.

The individual steps of the DCMS method are further detailed in Algorithm 1.

Data: a set of experimental samples Y_{kp} and the corresponding experimental ECDFs $\hat{F}_{Y_{kp}}$, a set of SCRNs $\Sigma_1, \ldots, \Sigma_Q$
Result: the SCRN that best explains the data Y_{kp}

begin
- set confidence level α (e.g. 0.05);
- set distance tolerance $\bar{\epsilon}$ (e.g. 0.25);
- find the minimum number of samples M in Y_{kp};
- evaluate $S^{(c)}(\bar{\epsilon}, \alpha, M)$ using (9);
- **for** $q \leftarrow 1$ **to** Q **do**
  - perform $S^{(c)}$ simulations of Σ_q and obtain the simulated data set X_{kp};
  - **for** $k \leftarrow 2$ **to** K **do**
    - **for** $p \leftarrow 1$ **to** P **do**
      - evaluate $d_{SM}\left(\hat{G}_{X_{kp}}, \tilde{F}_{Y_{kp}}\right)$;
      - **if** $d_{SM} > \bar{\epsilon}$ **then** set $\alpha_{kp} = \alpha$;
      - **else** set $\alpha_{kp} = 1$;
    - **end**
  - **end**
  - evaluate X^2 and $\alpha_q^{(F)}$ using (10) and (11);
  - **if** $\alpha_q^{(F)} \ll \alpha$ **then**
    - discard Σ_q as inconsistent;
  - **end**
- **end**
- rank the SCRNs according to the value of Fisher's statistic X^2 and the cumulative Kolmogorov distance (12);

end

Algorithm 1: The DCMS method.

IV. EXAMPLE

We now demonstrate the DCMS algorithm on stochastic models of a popular synthetic gene regulation network known as the *repressilator*. This circuit, first implemented by Elowitz and Leibler in 2000 [14], consists of three genes connected in a feedback loop where each gene produces the repressor protein for the next promoter in the loop. The repressilator has been analyzed extensively, both through deterministic and stochastic models [15], and it is frequently used as an example system for numerical methods [16], [17].

The stochastic repressilator candidate models we consider here have up to $N = 21$ species and $R = 48$ reactions. For each gene, up to 7 species are considered: unoccupied promoter, promoter bound to 1–4 repressor proteins, mRNA and protein. Their interactions are described by up to 16 reactions: binding and dissociation of up to 4 repressor molecule proteins to/from the promoter (up to 8 reactions), transcription from the unoccupied and from the occupied promoter (up to 5 reactions), translation, degradation of the mRNA and of the protein.

The purpose of the present example is to study *cooperative binding* in the repressilator. In other words, we assume that each promoter can have 1–4 binding sites for the corresponding repressor protein and that each bound repressor molecule increases the promoter affinity for additional repressor molecules (positive cooperativity). This gives rise to $Q = 4^3 = 64$ candidate models. We proceed by first simulating a flow cytometry data set from the $(2, 2, 2)$ model (Σ_{22}), which we consider to be the true network. We, then, test all the other candidates against these data using the DCMS algorithm.

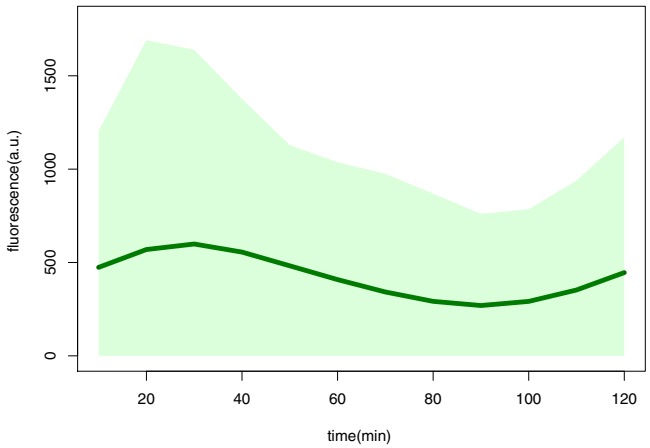

Fig. 3. **Simulated flow cytometry data set**. The bold curve shows the average fluorescence level of the second protein in the $(2, 2, 2)$ repressilator model Σ_{22} computed across $M = 50000$ simulations. The shaded area around the curve represents a one standard deviation interval around the mean. Note that fluorescence levels can not have negative values.

We assume that the network is observed at $K = 12$ time points (every 10 minutes for 2 hours), and that $P = 1$ species is measured (the second protein in the loop). The model selection data set was constructed using $M = 50000$ SSA simulations of the "true" network, with reaction rates derived from [14] and initial conditions adapted from [17] (Figure 3).

We fix a confidence level $\alpha = 0.05$ and a distance tolerance $\epsilon = 0.25$. With these choices, the critical number of simulations (9) evaluates to $S^{(c)} = 31$.

The output of the DCMS algorithm is detailed in Table I. For the sake of brevity, only the top 16 ranking models are listed. The true model Σ_{22} is correctly picked as the top-ranking candidate. We also note how all the other candidates have a value of $\alpha^{(F)}$ that is smaller than the chosen confidence level $\alpha = 0.05$. However, the second-ranking model Σ_{26} has $\alpha^{(F)} = 0.025$, which is relatively close to the threshold: therefore, this candidate should still be regarded as a plausible model for the process.

This analysis, which required a total of 1984 SSA simulations of the candidate models (not including the ones required for the generation of the simulated flow cytometry data), was carried out using a single-processor 2.6 GHz laptop running a custom-written C implementation of Gillespie's SSA direct method based on the GNU Scientific Library [18]. The total running time was about 46 minutes.

V. CONCLUSIONS

In the present manuscript, we described a new method for model selection in discretely observed stochastic chemical

TABLE I

MODEL SELECTION FOR THE STOCHASTIC REPRESSILATOR

Ranking	Model	Binding sites	X^2 statistic	Global p-value $\alpha^{(F)}$	Cumulative distance $\mathcal{D}$	DCMS result
1	Σ_{22}	$(2,2,2)$	0	1	1.8983	accept
2	Σ_{26}	$(2,3,2)$	41.9403	0.024899	5.5148	reject
3	Σ_{55}	$(3,2,4)$	53.9232	0.0010381	4.3885	reject
4	Σ_{60}	$(4,3,4)$	59.9146	0.00017219	4.8853	reject
5	Σ_{5}	$(1,2,1)$	59.9146	0.00017219	5.6034	reject
6	Σ_{11}	$(3,3,1)$	59.9146	0.00017219	6.0316	reject
7	Σ_{39}	$(3,2,3)$	59.9146	0.00017219	6.1064	reject
8	Σ_{54}	$(2,2,4)$	59.9146	0.00017219	6.3258	reject
9	Σ_{34}	$(2,1,3)$	59.9146	0.00017219	7.3595	reject
10	Σ_{40}	$(4,2,3)$	59.9146	0.00017219	8.1771	reject
11	Σ_{6}	$(2,2,1)$	59.9146	0.00017219	8.2341	reject
12	Σ_{44}	$(4,3,3)$	65.9061	2.5669e-05	5.1662	reject
13	Σ_{33}	$(1,1,3)$	65.9061	2.5669e-05	5.2885	reject
14	Σ_{12}	$(4,3,1)$	65.9061	2.5669e-05	5.3617	reject
15	Σ_{28}	$(4,3,2)$	65.9061	2.5669e-05	5.6541	reject
16	Σ_{10}	$(2,3,1)$	65.9061	2.5669e-05	6.1771	reject

The table shows the DCMS algorithm results for the top-ranking candidate models of the stochastic repressilator network. The "true" model Σ_{22} is correctly picked as the one that best explains the simulated flow cytometry data set. The second-ranking model Σ_{26} has a value of $\alpha^{(F)}$ that is less than $\alpha = 0.05$, but relatively close to the threshold, so it should still be regarded as a plausible model.

reaction networks using flow cytometry data, which we call *distribution comparison model selection* algorithm, or *DCMS*. Our proposed procedure is based on the comparison of empirical distributions from the candidate models and from the experimental data.

Using the notion of Kolmogorov distance and the Dvoretzky–Kiefer–Wolfowitz inequality, we derived a universal bound on the number of simulations needed for a statistically significant comparison of two ECDFs, which is independent of the functional form of the underlying exact CDFs. Furthermore, we found that for common data set sizes in flow cytometry experiments this number is surprisingly small, allowing one to screen large candidate models and/or a large number of candidate models in a very computationally efficient manner. In our example, we were able to compare 64 candidate models using only measurements from one species and 31 simulations for each SCRN.

Finally, we remark that the DCMS algorithm displays a compromise between speed and strictness: for fixed confidence level α and number of experimental samples M, a smaller Kolmogorov distance tolerance $\bar{\epsilon}$, corresponds to a larger critical number of simulations $S^{(c)}$. In other words, if one performs a very small number of simulations per SCRN, it is likely that more than one candidate will be accepted as statistically consistent by DCMS. Keeping this in mind, when screening a very large number of candidates, our proposed method can be easily applied in several stages, reducing the tolerance as the number of statistically consistent candidates decreases. In the stochastic repressilator scenario of Section IV, if we repeat the analysis using $\bar{\epsilon} = 0.6$ (8 simulations per SCRN), nineteen models are accepted as statistically consistent. If we then reapply DCMS with $\bar{\epsilon} = 0.25$ to these five models only, we obtain the same result of selecting Σ_{22} as the top-ranking candidate.

REFERENCES

[1] D. M. Bortz and P. W. Nelson, "Model selection and mixed-effects modeling of HIV infection dynamics," *B Math Biol*, vol. 68, no. 8, pp. 2005–2025, 2006.

[2] V. Vyshemirsky and M. A. Girolami, "Bayesian ranking of biochemical system models," *Bioinformatics*, vol. 24, pp. 833–9, Mar 2008.

[3] A. Milias-Argeitis, R. Porreca, S. Summers, and J. Lygeros, "Bayesian model selection for the yeast GATA-factor network: A comparison of computational approaches," in *49th IEEE Conference on Decision and Control (CDC2010)*, pp. 3379–3384, 2010.

[4] T. Toni and M. Stumpf, "Simulation-based model selection for dynamical systems in systems and population biology," *Bioinformatics*, vol. 26, no. 1, pp. 104–110, 2010.

[5] G. Lillacci and M. Khammash, "Parameter estimation and model selection in computational biology," *PLoS Comp Biol*, vol. 6, no. 3, p. e1000696, 2010.

[6] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *J Comput Phys*, vol. 22, no. 4, pp. 403–434, 1976.

[7] S. K. Poovathingal and R. Gunawan, "Global parameter estimation methods for stochastic biochemical systems," *BMC Bioinformatics*, vol. 11, p. 414, 2010.

[8] V. I. Glivenko, "Sulla determinazione empirica della legge di probabilità," *Giorn Ist Ital Attuari*, vol. 4, pp. 92–99, 1933.

[9] F. P. Cantelli, "Sulla determinazione empirica della legge di probabilità," *Giorn Ist Ital Attuari*, vol. 4, pp. 221–424, 1933.

[10] A. Dvoretzky, J. Kiefer, and J. Wolfowitz, "Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator," *Ann Math Statist*, vol. 27, no. 3, pp. 642–669, 1956.

[11] P. Massart, "The tight constant in the Dvoretzky–Kiefer–Wolfowitz inequality," *Ann Probab*, vol. 18, no. 3, pp. 1269–1283, 1990.

[12] R. A. Fisher, *Statistical Methods for Research Workers*. Oliver and Boyd, 1925.

[13] R. Delongchamp, T. Lee, and C. Velasco, "A method for computing the overall statistical significance of a treatment effect among a group of genes," *BMC Bioinformatics*, vol. 7 (Suppl 2), p. S11, 2006.

[14] M. B. Elowitz and S. Leibler, "A synthetic oscillatory network of transcriptional regulators," *Nature*, vol. 403, pp. 335–338, Jan 2000.

[15] A. Loinger and O. Biham, "Stochastic simulations of the repressilator circuit," *Phys Rev E*, vol. 76, p. 051917, Nov 2007.

[16] M. Quach, N. Brunel, and F. d'Alche Buc, "Estimating parameters and hidden variables in non-linear state-space models based on ODEs for biological networks inference," *Bioinformatics*, vol. 23, no. 23, pp. 3209–3216, 2007.

[17] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *J R Soc Interface*, vol. 6, no. 31, pp. 187–202, 2009.

[18] M. Galassi, J. Davies, J. Theiler, B. Gough, G. Jungman, P. Alken, M. Booth, and F. Rossi, *GNU Scientific Library reference manual*. Network Theory, third ed., 2009.