Bayesian model selection for the yeast GATA-factor network: a comparison of computational approaches

Milias-Argeitis, Andreas ; Porreca, Riccardo ; Summers, Sean ; Lygeros, John

Abstract: A common situation in System Biology is to use several alternative models of a given biochemical system, each with a different structure reflecting different biological hypotheses. These models then have to be ranked according to their ability to reproduce experimental data. In this paper, we use Bayesian model selection to test four alternative models of the yeast GATA-factor genetic network. We employ three different computational methods to calculate the necessary probabilities and evaluate their performance for medium-scale biochemical systems.

DOI: https://doi.org/10.1109/CDC.2010.5717307

49th IEEE Conference on Decision and Control
December 15-17, 2010
Hilton Atlanta Hotel, Atlanta, GA, USA

Bayesian model selection for the yeast GATA-factor network: a comparison of computational approaches

Andreas Milias-Argeitis, Riccardo Porreca, Sean Summers, and John Lygeros

***Abstract*— A common situation in System Biology is to use several alternative models of a given biochemical system, each with a different structure reflecting different biological hypotheses. These models then have to be ranked according to their ability to reproduce experimental data. In this paper, we use Bayesian model selection to test four alternative models of the yeast GATA-factor genetic network. We employ three different computational methods to calculate the necessary probabilities and evaluate their performance for medium-scale biochemical systems.**

I. Introduction

Systems Biology is largely based on the development of quantitative models for biochemical systems to provide insights into their behavior. This task, however, is often complicated due to several factors, such as the scarcity of experimental data, conflicting biological hypotheses and ignorance of mechanistic details of many biological processes. Using a set of alternative mathematical models to describe different hypotheses about a complex biochemical network is a common practice in this case [1] and gives rise to the problem of model selection, i.e. the determination of the most plausible model (or set of models) given an experimental dataset.

Bayesian model comparison is a good candidate for this difficult task, since it calculates the marginal posterior probabilities of the models given experimental data in a way consistent with prior knowledge, embodied in the prior distributions over models and parameters. Moreover, it naturally prevents overfitting by penalizing excessively complex models, as will become clear in the sequel.

While theoretically attractive, Bayesian model selection faces challenging problems when it comes to practical implementation. In many cases, calculating the required marginal posterior probabilities becomes computationally prohibitive, as it involves high-dimensional integration. Several methods have been proposed to tackle this problem and in this paper we consider three of them: Annealed Importance Sampling [2], a method based on Markov Chain Monte Carlo (MCMC) which we will refer to as the Chib method [3], and Approximate Bayesian Computation [4].

Using simulated datasets from four alternative mathematical models of the GATA-factor genetic network in yeast (*Saccharomyces cerevisiae*), we compare the performances of the chosen computational methods. The use of artificial data makes it possible to compare the accuracy, scalability and computational bottlenecks of the methods in a very controlled manner, contrary to the use of a real dataset. Consequently, we are able to accurately assess the strengths and weaknesses of each method when applied to a medium-scale biochemical system, rather than toy models, and gain the necessary computational experience to approach a forthcoming real experimental dataset.

This work was supported in part by the SystemsX.ch research consortium under the project YeastX.

Andreas Milias-Argeitis, Riccardo Porreca, Sean Summers, and John Lygeros are with the Automatic Control Laboratory, Department of Information Technology and Electrical Engineering, ETH Zurich, Switzerland milias,rporreca,summers,lygeros@control.ee.ethz.ch

II. Modeling the GATA network

Yeast cells can sense and utilize a great variety of nitrogen sources in their environment. They can also discriminate and selectively utilize good nitrogen sources (e.g. glutamine) in preference to poor ones (e.g. proline), a process called *nitrogen catabolite repression* (NCR). Through NCR yeast cells are able to repress the expression of genes coding for enzymes and permeases required for importing and degrading poor nitrogen sources when a preferred nitrogen source is present [5]. NCR-sensitive gene expression is coordinated by four GATA-type transcription factors (TFs): two transcriptional activators (Gln3 and Gat1) and two repressors (Dal80 and Gzf3), that control NCR genes combinatorially. In the presence of a good nitrogen source, Gln3 and Gat1 are both sequestered in the cytoplasm. Depletion of the preferred source causes the activation of the TOR signaling pathway, which eventually results in the nuclear translocation of Gln3 and Gat1 ([6] and references therein).

In this paper we focus on the network of interactions among the GATA factors, displayed in Figure 1. In this biological model, solid lines represent interactions confirmed by several literature sources, while dashed lines indicate interactions that biologists hypothesize to exist, but cannot be unambiguously inferred yet. In order to study the dynamics of the GATA network, we have created an ODE-based mathematical model of the regulatory interactions among GATA factors, as well as their documented protein-protein interactions (such as homo- and heterodimerizations) that - to the best of our knowledge - incorporates all available biological hypotheses on the GATA network. From this full model it is possible to derive submodels reflecting any desirable combination of hypotheses.

The full model comprises 14 states, representing 4 mRNAs, 4 monomeric nuclear TFs, 2 cytoplasmic TFs and 4 different dimer combinations.

For each gene, we consider a two-step kinetic model of mRNA and protein production (of the form $d/dt[mRNA] = k_t Y - k_{dm}[mRNA]$, $d/dt[Protein] = k_s[mRNA] -$

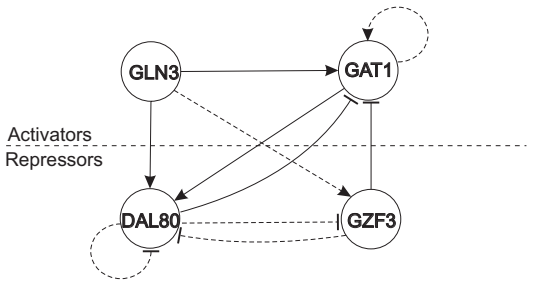

Fig. 1. The GATA-factor genetic network. Sharp arrows denote activation, while blunt arrows denote repression.

$k_{dp}[Protein]$), where Y, the promoter occupancy (a.k.a. *regulation function*), is a nonlinear function of the regulators of that gene. Dimer formation dynamics are described by mass-action kinetics and nuclear import/export follows first-order linear dynamics. The full model contains 36 parameters plus 14 initial conditions, for several of which approximate values have been found in the literature [7]–[9]. Henceforth, we shall not distinguish parameters from initial conditions. The model equations are reported in [10].

III. Bayesian Model Selection

Consider a set of competing biological hypotheses $\{\mathcal{H}_k\}_{k=1}^K$, each represented by a mathematical model $\mathcal{M}_k$, to be compared given experimental data D. For each model $\mathcal{M}_k$ dataset D is assumed to have a density $P(D|\mathcal{M}_k, \boldsymbol{\theta}_k)$, where $\boldsymbol{\theta}_k$ is the parameter vector of model $\mathcal{M}_k$. Our prior knowledge about parameter values and plausibility of the models is encoded by the prior distributions $P(\boldsymbol{\theta}_k|\mathcal{M}_k)$ and $P(\mathcal{M}_k)$ respectively. The first key object in Bayesian model selection is the marginal density of D under $\mathcal{M}_k$,

$$P(D|\mathcal{M}_k) = \int P(D|\mathcal{M}_k, \boldsymbol{\theta}_k) P(\boldsymbol{\theta}_k|\mathcal{M}_k)\, d\boldsymbol{\theta}_k, \quad (1)$$

also called the *evidence* for $\mathcal{M}_k$. The second is the *Bayes factor* [11] B_{ij} of $\mathcal{M}_i$ to $\mathcal{M}_j$, given by

$$B_{ij} = P(D|\mathcal{M}_i)/P(D|\mathcal{M}_j), \quad (2)$$

which can be interpreted as the "weight of evidence" provided by the data in favor of model i as opposed to model j. A Bayes factor greater than 10 is commonly interpreted as strong evidence in favor of model i [11].

The Bayes factor is used to update our initial beliefs in models i and j and obtain the posteriors ratio $P(\mathcal{M}_i|D)/P(\mathcal{M}_j|D) = B_{ij}P(\mathcal{M}_i)/P(\mathcal{M}_j)$. The model posterior probability $P(\mathcal{M}_k|D)$ is the fundamental quantity in Bayesian model selection. Having the Bayes factors, we can compute it as

$$P(\mathcal{M}_k|D) = \left[\sum_{j=1}^{K} \frac{P(\mathcal{M}_j)}{P(\mathcal{M}_k)} B_{jk}\right]^{-1}. \quad (3)$$

$P(\mathcal{M}_k|D)$ can be interpreted as a measure of the "plausibility" of model k after all information provided by D has been considered. Thus, model selection is accomplished by choosing the most probable model (or models, in case the data is not discriminative enough) [12].

This approach to model selection naturally strikes a balance between data misfit and model complexity. The key to this balance comes from (1): a simple model $\mathcal{M}_s$ has a limited predictive ability, which means that its evidence $P(D|\mathcal{M}_s)$ is concentrated in a small region of the space of all possible datasets D. On the contrary, a more complex model $\mathcal{M}_c$ is able to generate a larger variety of datasets, which however implies that its evidence is spread over a larger region of the data space. If the data we observe fall within the high-density region of $P(D|\mathcal{M}_s)$, the simpler model will end up as the most probable model assuming that both models have equal priors (a more detailed discussion can be found in [13, Ch.28]).

As a simple example, consider two nested models $\mathcal{M}_i \subset \mathcal{M}_j$ (meaning that $\boldsymbol{\theta}_j = [\boldsymbol{\theta}_i \ \boldsymbol{\phi}_0]$). If $P(\boldsymbol{\theta}_j|\mathcal{M}_j) \propto P(\boldsymbol{\theta}_i|\mathcal{M}_i)P(\phi_0|\mathcal{M}_j)$ then, although $\sup_{\boldsymbol{\theta}_i} P(D|\mathcal{M}_i, \boldsymbol{\theta}_i) \leq \sup_{\boldsymbol{\theta}_j} P(D|\mathcal{M}_j, \boldsymbol{\theta}_j)$ a fortiori, the more complex model will not be favored by (1) unless it can greatly improve upon the predictive ability of the simpler model.

Thus, we see that the principle of Occam's razor is automatically incorporated into Bayesian model selection. Moreover, this holds true even when we compare nonnested models, something very hard to achieve within a frequentist model selection framework.

IV. Computational methods

From a computational viewpoint, our efforts concentrate on the calculation of (3) for each model $\mathcal{M}_k$. However, except for very special cases, the integral of (1) is analytically intractable, while Monte Carlo integration (such as Importance Sampling) quickly becomes inefficient as the parameter space grows. In the following, we shall describe two methods that can provide reliable estimates of the model evidence for medium-size problems and one that avoids the calculation of (1) altogether.

A. Annealed Importance Sampling (AIS)

The possibility of using Importance Sampling for (1) with $P(\boldsymbol{\theta}|\mathcal{M})$ as the proposal and $P(\boldsymbol{\theta}|D, \mathcal{M},) \propto P(D|\mathcal{M}, \boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{M})$ as the (unnormalized) target distribution leads in most cases to very high variance of the importance weights, since the priors are usually diffuse while the posterior is much more concentrated. The method of Annealed Importance Sampling (AIS) tries to circumvent this problem by defining a sequence of bridging distributions f_β according to a "cooling schedule": $f_{\beta_i}(\boldsymbol{\theta}) \propto P(D|\mathcal{M}, \boldsymbol{\theta})^{\beta_i} P(\boldsymbol{\theta}|\mathcal{M})$, for $0 = \beta_0 < \beta_1 < \cdots < \beta_N = 1$. After drawing a population of particles $\{\boldsymbol{\theta}_{(0)}^j\}_{j=1}^M$ from f_{β_0} with weights $\{w_0^j\}_{j=1}^M$ all equal to 1, the basic scheme follows Algorithm 1.

Neal proves [2] that $1/M(\sum_{j=1}^M w_N^j) \xrightarrow{M\to\infty} Z_N/Z_0 = \int f_{\beta_N}(\boldsymbol{\theta})\, d\boldsymbol{\theta} / \int f_{\beta_0}(\boldsymbol{\theta})\, d\boldsymbol{\theta}$, which is the integral of interest. $K_n(\cdot|\boldsymbol{\theta}_{(n-1)}^j)$ is a Markov kernel that leaves $f_{\beta_n}(\boldsymbol{\theta})$ invariant. A simple choice is to let $K_n(\cdot|\boldsymbol{\theta}_{(n-1)}^j)$ consist of a few

Algorithm 1 Annealed Importance Sampling

1: **for** $j = 1$ to M **do**
2: **for** $n = 1$ to N **do**
3: Generate $\boldsymbol{\theta}^j_{(n)}$ from $K_n(\cdot|\boldsymbol{\theta}^j_{(n-1)})$
 Update $w^j_n = w^j_{n-1} \dfrac{f_{\beta_n}(\boldsymbol{\theta}^j_{(n-1)})}{f_{\beta_{n-1}}(\boldsymbol{\theta}^j_{(n-1)})}$
4: **end for**
5: **end for**

(10-20) Metropolis-Hastings (M-H) updates of $\boldsymbol{\theta}^j_{(n-1)}$ with $f_{\beta_n}(\boldsymbol{\theta})$ as the target density. It is not necessary that the K_n's mix fast, although this is desirable for reducing the weight variance [2]. However, if the particles are too few or the annealing steps too big, the importance weights may still end up having a large variance, resulting in a small effective sample size (ESS), and correspondingly a bad estimate.

In fact, in the high-dimensional cases in which we applied AIS, the above version of the algorithm resulted in very small ESS even with many ($\sim$ 200) annealing steps and 15000 particles. This prompted us to examine a variation of the algorithm, in which we monitor the ESS at each step n and resample the particles from the current approximation of f_{β_n} if $ESS < 0.5M$. The correct way to do this can be found within the general framework of Sequential Monte Carlo methods, of which AIS is a particular example [14, Algorithm 3.1.1].

B. The Chib method

According to the approach of [3], it is possible to exploit the fact that $P(D|\mathcal{M})$ is the normalizing constant of the posterior density $P(\boldsymbol{\theta}|D,\mathcal{M})$: $P(D|\mathcal{M}) = P(D|\mathcal{M},\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{M})/P(\boldsymbol{\theta}|D,\mathcal{M})$.

Notice that this relation holds for any value of $\boldsymbol{\theta}$ and that both likelihood $P(D|\mathcal{M},\boldsymbol{\theta})$ and prior distribution $P(\boldsymbol{\theta}|\mathcal{M})$ are assumed to be known. Given a suitable point $\boldsymbol{\theta} = \boldsymbol{\theta}^*$, the calculation of $P(D|\mathcal{M})$ thus reduces to computing an estimate of the posterior density $P(\boldsymbol{\theta}^*|D,\mathcal{M})$ at $\boldsymbol{\theta}^*$. As pointed out in [3], taking $\boldsymbol{\theta}^*$ as a high density point under the posterior can improve the estimation efficiency. The estimation of $P(\boldsymbol{\theta}^*|D,\mathcal{M})$ is obtained from samples of the posterior density generated with the Metropolis-Hastings (M-H) algorithm. Let $q(\boldsymbol{\theta},\boldsymbol{\theta}'|D,\mathcal{M})$ be the proposal density of the M-H algorithm for the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$ and $\alpha(\boldsymbol{\theta},\boldsymbol{\theta}'|D,\mathcal{M})$ denote the probability of moving from $\boldsymbol{\theta}$ to the proposed $\boldsymbol{\theta}'$. It is then possible to express the posterior density at $\boldsymbol{\theta}^*$ as $P(\boldsymbol{\theta}^*|D,\mathcal{M}) = \mathbf{E}_{\text{post}}\left[\alpha(\boldsymbol{\theta},\boldsymbol{\theta}^*|D,\mathcal{M})q(\boldsymbol{\theta},\boldsymbol{\theta}^*|D,\mathcal{M})\right] / \mathbf{E}_q\left[\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}|D,\mathcal{M})\right]$, where $\mathbf{E}_{\text{post}}$ and $\mathbf{E}_q$ denote expectation with respect to the distributions $P(\boldsymbol{\theta}|D,\mathcal{M})$ and $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|D,\mathcal{M})$, respectively.

The expectations above are estimated using N samples $\{\boldsymbol{\theta}^{(i)}\}$ from $P(\boldsymbol{\theta}|D,\mathcal{M})$ (provided by the M-H algorithm) and M samples $\{\boldsymbol{\theta}^{(j)}\}$ from $q(\boldsymbol{\theta}^*,\boldsymbol{\theta}|D,\mathcal{M})$, thus obtaining the estimate $\hat{P}(\boldsymbol{\theta}^*|D,\mathcal{M}) = (M/N)\left(\sum_{i=1}^{N}(\alpha(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}^*|D,\mathcal{M})q(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}^*|D,\mathcal{M})\right) / \left(\sum_{j=1}^{M}(\alpha(\boldsymbol{\theta}^*,\boldsymbol{\theta}^{(j)}|D,\mathcal{M})\right)$.

In order to address high-dimensional problems, the method has been extended to sample $\boldsymbol{\theta}$ in B blocks $\theta_1,\ldots,\theta_B$. In this case, $P(\boldsymbol{\theta}^*|D,\mathcal{M})$ is obtained as $P(\boldsymbol{\theta}^*|D,\mathcal{M}) = \prod_{b=1}^{B} P(\theta^*_b|D,\mathcal{M}_k,\theta^*_1,\ldots,\theta^*_{b-1})$, where each term of the product is computed in way a similar to the single block case presented above (see [3] for further details). Note that that B MCMC runs are needed for this implementation, each performed by fixing $\theta_1,\ldots,\theta_{b-1}$ to the corresponding entries of $\boldsymbol{\theta}^*$.

C. Approximate Bayesian Computation (ABC)

In the parameter estimation framework, "likelihood-free" ABC methods provide an attractive alternative to traditional Bayesian approaches by sampling directly from the posterior $P(\boldsymbol{\theta}|D) \propto f(D|\boldsymbol{\theta})P(\boldsymbol{\theta})$ via simulation. That is, given the prior distribution $P(\boldsymbol{\theta})$, the approximate Bayesian computation algorithm jointly simulates $\boldsymbol{\theta}' \sim P(\boldsymbol{\theta})$ and $D' \sim f(D|\boldsymbol{\theta}')$, and accepts the sampled $\boldsymbol{\theta}'$ if and only if the simulated dataset $D' = D$, where $D \sim f(D|\boldsymbol{\theta})$ is the observed dataset. This algorithm is exact in that the accepted parameter values $\boldsymbol{\theta}'$ are distributed according to the posterior $P(\boldsymbol{\theta}|D)$. However, in most cases the restriction that $D' = D$ is computationally prohibitive and is replaced by a tolerance condition $d(D,D') \leq \epsilon$, where d is a distance function that quantifies the discrepancy between the datasets and ϵ is a tolerance value. The output is then distributed proportional to the density $P(\boldsymbol{\theta})P_{\boldsymbol{\theta}}\{d(D,D') \leq \epsilon\}$. It is supposed that if the distance function is appropriately chosen and ϵ is small enough, $P(\boldsymbol{\theta})P_{\boldsymbol{\theta}}\{d(D,D') \leq \epsilon\}$ is a good approximation of the true posterior $P(\boldsymbol{\theta}|D)$. Successful algorithms for ABC methods include those based on rejection sampling [15], MCMC [16], and SMC-type samplers [14], [17].

Recently, ABC has been extended from the parameter estimation framework to that of model selection [4], [18]. Here, the focus shifts from $P(\boldsymbol{\theta}|D)$ to $P(\mathcal{M}|D)$. Currently, we focus on the joint space-based approach which first aims to approximate the joint posterior distribution $P(\mathcal{M},\boldsymbol{\theta}|D) \propto f(D|\mathcal{M},\boldsymbol{\theta})P(\mathcal{M},\boldsymbol{\theta})$ via ABC simulation, where the prior over the joint model-parameter space is given as $P(\mathcal{M},\boldsymbol{\theta}) = P(\mathcal{M})P(\boldsymbol{\theta}|\mathcal{M})$. Then, the approximate marginal posterior over the model space is obtained by marginalizing over the parameter space. In [4], the ABC rejection model selection algorithm presented in [18] was extended to the Sequential Monte Carlo (SMC) framework [14], [17]. In the ABC SMC framework, the particles are sampled from the prior distribution, and propagated through a sequence of intermediate distributions $P(\mathcal{M},\boldsymbol{\theta})P_{(\mathcal{M},\boldsymbol{\theta})}\{d(D,D') \leq \epsilon_i\}$ until they represent a sample from the target distribution $P(\mathcal{M},\boldsymbol{\theta})P_{(\mathcal{M},\boldsymbol{\theta})}\{d(D,D') \leq \epsilon_K\}$. The ABC SMC algorithm [4] relies on a predetermined cooling schedule $\epsilon_1 > \epsilon_2 > \ldots > \epsilon_K \geq 0$ and importance sampling in order to gradually evolve towards the target joint posterior distribution.

In this paper, we implement a variation of the ABC SMC model selection algorithm of [4], in which we adaptively select the cooling schedule $\{\epsilon_1,\ \epsilon_2,\ldots,\ \epsilon_K\}$ according the

algorithm proposed in [19]. A pseudocode version of the algorithm can be found in [10].

V. Computational study

A. Setup

1) Alternative models, data generation: We consider four alternative models of the GATA-factor network and two sets of free parameters for each model, $\mathcal{S}_1$ and $\mathcal{S}_2$ ($\mathcal{S}_1 \subset \mathcal{S}_2$), keeping the remaining parameters fixed. The free parameter sets that we chose are much larger than those usually found in toy model examples in the literature, making the computational procedures more challenging than previously described. Note that these sets were determined for the full model, which means that some of the submodels end up with a reduced number of free parameters. Table I summarizes the characteristics of the models considered ($|\cdot|$ denotes set cardinality):

TABLE I
ALTERNATIVE MODELS

Model	Characteristics	$\lvert\mathcal{S}_1\rvert$	$\lvert\mathcal{S}_2\rvert$
$\mathcal{M}_1$	full model	10	15
$\mathcal{M}_2$	no repression on Dal80	10	13
$\mathcal{M}_3$	no regulation of Gzf3	9	13
$\mathcal{M}_4$	no Gat1 self-activation, no activator cooperativity	9	14

$\mathcal{S}_1$ consists of production and degradation rates for Dal80 mRNA and protein (all of which remain largely unspecified in the literature), the initial condition of Dal80 protein, as well as parameters of the regulation functions of Gat1 and Gzf3. $\mathcal{S}_2$ augments $\mathcal{S}_1$ by adding more regulation function parameters of Dal80 and Gzf3 (more details in [10]).

The number of free parameters, while about a third of the total, is enough to enable one model to fit relatively well data generated by another. More specifically, $\mathcal{M}_1$ and $\mathcal{M}_4$ are the most structurally similar, while $\mathcal{M}_3$ differs significantly from the others. To test the aforementioned computational methods, we simulate a shift from glutamine to proline at time $t = 0$, generate a noisy dataset with each of the models (shown in Figure 2) and then try to identify the correct model each time.

Our datasets consist of noisy measurements of three mRNAs (Gat1, Dal80 and Gzf3) made at times $T = \{20, 40, 80, 120, 160, 200, 250\}$ minutes after the shift. The sparse sampling of three out of fifteen states and the addition of noise make the datasets similar to what one can expect from real experiments.

All parameter priors are set to uniform on closed intervals of width 1-2 orders of magnitude, containing the values used to generate the data. The measurement noise is assumed to be additive i.i.d. Gaussian with independent components, giving rise to a likelihood function of the form $P(D|\mathcal{M}, \boldsymbol{\theta}) = \prod_{t\in T} \mathcal{N}(D_t; x(t), \Sigma)$, where D_t is the measurement vector at time t, $\mathcal{N}(\cdot; \mu, \Sigma)$ the multivariate Gaussian density with mean μ and covariance Σ (diagonal in our case) and x contains the three measured states predicted under model $\mathcal{M}$

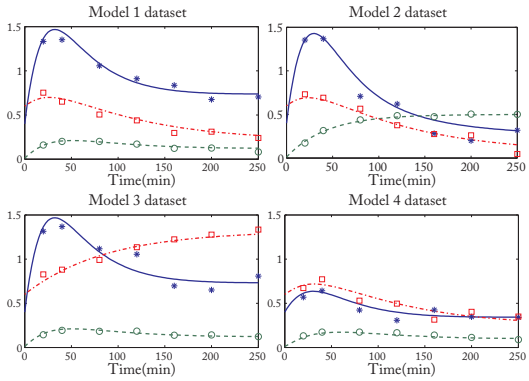

Fig. 2. Data used for model selection. Full lines+asterisk markers: Gat1 mRNA trajectory+measurements. Dashed lines+circle markers: Dal80 mRNA trajectory+measurements. Dash-dot lines+square markers: Gzf3 trajectory+measurements

with parameter vector $\boldsymbol{\theta}$. Finally, all model prior probabilities are set to 0.25.

2) Computational method settings:

a) AIS: For $\mathcal{S}_1$ we propagate 2500 particles through 100 annealing steps, spaced uniformly in $[0, 1]$. Each K_n consists of 15 M-H updates using Gaussian proposals centered on the current state with appropriately tuned bandwidth to ensure adequate mixing. For $\mathcal{S}_2$ we propagate 5000 particles through 208 annealing steps, with 10 steps spaced uniformly in $[0, 0.01]$ and another 198 spaced uniformly in $[0.015, 1]$. The K_n's, similarly defined as above, consist of 20 M-H updates.

b) ABC: We use $N = 1000$ particles for $\mathcal{S}_1$ and $N = 2000$ particles for $\mathcal{S}_2$. For all experiments the initial tolerance value is $\epsilon_1 = 25$ and the target tolerance is $\epsilon_K = 6.8$. The distance function is $d(D, D') = \sqrt{\sum_{t\in T}\sum_{j=1}^{3}\left(\frac{D_t(j)-D'_t(j)}{\sigma(j)}\right)^2}$, where $\sigma(j)$ is the standard deviation of the measurement noise associated with the $j-$th measurement in the data vector. The model perturbation kernel at population k is $KM_k(\mathcal{M}|\mathcal{M}^*) = \frac{1}{2}$ if $\mathcal{M} = \mathcal{M}^*$, otherwise $KM_k(\mathcal{M}|\mathcal{M}^*) = \frac{1}{6}$. The parameter perturbation kernel at population k is $KP_k(\theta|\theta^*) = \mathcal{N}(\theta^*, \Sigma^2)$, where Σ is a diagonal matrix of standard deviations corresponding to around 10 percent of the parameter interval.

c) The Chib method: The method is applied using 2 and 3 blocks for $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively, with 5 parameters per block (reduced if needed, to account for the fixed parameters in the submodels). The M-H algorithm employs Gaussian proposals, with standard deviations corresponding to 3% of each parameter interval. A full MCMC sampling is first run for 10^5 steps, using only the last $2.5 \cdot 10^4$ samples for the subsequent computations to avoid convergence problems. The point $\boldsymbol{\theta}^*_k$ is chosen as the sample characterized by the highest value of the posterior density. Each of the reduced MCMC runs is then performed, for increasing values of b, by fixing blocks $\theta_{k,1}, \ldots, \theta_{k,b}$ to $\theta^*_{k,1}, \ldots, \theta^*_{k,b}$ and sampling the remaining blocks for $2.5 \cdot 10^4$ steps. All samples obtained are used to compute the marginal posterior density value $P(\boldsymbol{\theta}^*_k|D, \mathcal{M}_k)$.

B. Results

Given the uniform prior over the models, the posterior probability ratios are equal to the Bayes factors, according to (3). In Tables II and III we report the logarithms of the posterior ratios $P_{ij} = \log_{10}(P(\mathcal{M}_i|D)) - \log_{10}(P(\mathcal{M}_j|D))$, calculated with the three methods for the two parameter sets. Each row ($i = 1, \dots 4$) contains results obtained by considering $\mathcal{M}_i$ as the correct model. Every triplet of numbers indicates the P_{ij} value estimated with the three methods in the following order: AIS, Chib method, ABC.

TABLE II
POSTERIOR RATIO LOGARITHMS FOR $\mathcal{S}_1$

P_{i1}	P_{i2}	P_{i3}	P_{i4}
—	9.6, 9.6, 0.7	21.9, 313, ∞	1.5, 1.1, 0.5
5, 6.5, 3	—	37.2, ∞, ∞	9, 9.1, 4
32.7, 71, ∞	28.6, 70.1, ∞	—	32.5, 53.8, ∞
7.3, 8.6, 3.2	15.2, 16.2, 4	23, 271.6, ∞	—

TABLE III
POSTERIOR RATIO LOGARITHMS FOR $\mathcal{S}_2$

P_{i1}	P_{i2}	P_{i3}	P_{i4}
—	9.6, 9.8, 0.8	20.8, 312.1, ∞	2, 1.6, 0.9
1.3, 2.7, ∞	—	57.2, ∞ ,∞	4.9, 5.2, 1.1
3.2, 5.8, 1.8	7.7, 8.4, 1.9	—	5.2, 6.4, 2.3
6.5, 7.1, 1.2	14.6, 14.5, 1.4	19.6, 270.2, ∞	—

From Tables II and III we observe that all three methods are able to correctly identify the model that generated the data in each case. The estimates provided by AIS and the Chib method are quite close when the P_{ij}'s are smaller than about 20. For bigger P_{ij} values (which are, anyway, more than enough to safely discriminate between two models) the estimates diverge significantly. This is no surprise, as the wrong models often have to be very finely tuned to fit the data and this necessarily leads to extremely small high-density regions in the parameter posteriors, which the samplers can easily miss. In any case, calculating the model posteriors to a high precision would be useless in most cases arising in Systems Biology. If P_{ij}'s turn out even below 0.5, no safe conclusions should be drawn, since the noise in the measurements and the sparse sampling can easily tip the balance in favor of one model against another. On the other hand, a P_{ij} greater than 1, is enough to decisively support one of the models and cannot be attributed to noise realization.

Bayesian model selection shows its strength in the case of relatively small P_{ij} (e.g. P_{14}), where the maximum likelihoods are comparable and sometimes the wrong model has a higher maximum than the correct. By taking into account the overall ability of a model to fit the data (through (1)) instead of comparing the best fits, which could change significantly given different measurement noise realizations, the Bayesian approach can also prevent more complex models ($\mathcal{M}_1$) from being selected when data is generated by a simpler one.

It is clear that the ABC SMC algorithm produces estimates that are not quantitatively similar to those of the other methods, due to several practical algorithmic factors. In the current case we will focus on the effect of the tolerance schedule and the minimal tolerance value. As stated in Section IV-C, the ABC SMC approach exactly approximates the true posterior as $\epsilon \to 0$. Yet, in practice this restriction is computationally infeasible, and as a result the terminal tolerance must be taken as $\epsilon > 0$. It is theorized that the epsilon-approximate posterior and the true posterior are almost equal for very low ϵ. Yet, how small ϵ must be for this to hold is both problem dependent and generally intractable. Thus, it is not surprising that the results from the ABC SMC algorithm differ significantly from those of AIS and the Chib method given that they inherently approximate different posteriors.

We also see that the ABC SMC algorithm often results in infinite factors as opposed to large factor values. Again, this is not surprising given that the ABC SMC algorithm is effectively sampling from the joint model and parameter space. If the likelihood for a model is extremely small for some ϵ, it is feasible (and expected) that it would simply be lost (sampled out) and attributed a marginal posterior equal to zero. Note that this event is far from desirable, especially for large ϵ, since in the worst case a model with a low marginal probability for high ϵ may be lost even though it has a high probability for low ϵ. Losing viable hypotheses during the early stages of the tolerance schedule should be avoided, yet cannot be predicted, and therefore represents a significant risk within the ABC SMC framework.

C. Discussion

The methods employed to calculate the model marginal probabilities in this paper are far more computationally demanding than those used most frequently in Bayesian model selection, such as the Prior Arithmetic Mean Estimator and the Posterior Harmonic Mean Estimator. A recent survey [20], however, has shown that these popular, straightforward estimators can fail spectacularly even in the case of nonlinear biochemical models much smaller than the ones considered here. The reason is that model nonlinearities, combined with a small set of measured variables, can lead to very complex, multimodal parameter posteriors, which we also observed in this study. Despite these difficulties, the methods employed here could consistently identify the correct model.

The first one, based on Annealed Importance Sampling, has already been reliably used for model selection [20], albeit for much smaller parameter spaces. In this work, we demonstrate that it still works in 10 and 15-dimensional spaces with appropriate tuning of the annealing schedule and the Markov kernels. An attractive feature of the method is that the Markov kernels do not need to accurately sample their invariant distributions, as explained in [2], thus avoiding a common MCMC problem. However, it is also the slowest method in terms of computation time, with runs lasting about 3 hours for $\mathcal{S}_1$ and 8 hours for $\mathcal{S}_2$ (using a computer with quad-core AMD processor at 2.6 GHz running Matlab-generated code) . The problem of low effective sample size, which invariably occurred both in 10 and 15 dimensions, was treated by using the resampling scheme reported in [14]. This

improved substantially the estimate variability (observed by running the algorithm several times on the models that had the smallest P_{ij} values), although the estimate means obtained before and after the modification did not change significantly. A proper statistical analysis remains to be done to quantify precisely the effects of resampling.

The second alternative, the Chib method for marginal likelihood estimation, is computationally much more efficient compared to AIS and relatively straightforward to tune and implement, due to its MCMC origins. However, convergence of MCMC gets increasingly difficult as the parameter space grows and the blocking approach, while usually alleviating the problem, might result in biased results if the blocks are not selected to be "maximally uncorrelated". In some of our tests the variance of the P_{ij} estimates was around 1-1.5, while other P_{ij}'s were estimated very consistently. Thus, the results shown on the tables above are averaged over several (5-10) runs of the algorithm for each dataset and model. Multiple runs are recommended, especially when P_{ij} values are relatively small, to ensure that the method will pick the correct model. Even with repeated runs, this method is a very time-efficient alternative to AIS, with a 15-parameter run taking around 1 hour as opposed to 8 (using the same computer as above).

The third method, based on Approximate Bayesian Computation (ABC), has been advertised as simple to implement and intuitive, and therefore well suited for a large audience, e.g. experimental biologists. However, as the parameter and model space grow to realistic sizes, the importance of the user-defined distance function, perturbation kernels, tolerance schedule, particle sample size, and method of computing the importance weights become increasingly important. For small sample sizes, particle degeneracy can lead to a skewed representation of the marginal posterior in the model space. This can lead to incorrect model selection, and in the worse case, can cause the correct model to lose all probability before the target tolerance has been reached.

VI. Conclusions

The work reported in this article served as a primary feasibility study for Bayesian model selection applied to the GATA-factor network, in view of the forthcoming mRNA and protein time-course datasets from dynamic shifts of nitrogen sources, provided to us by our project collaborators.

Our numerical results show that the Annealed Importance Sampling and the Chib method can estimate the marginal likelihood for each of the models satisfactorily, even with as many as 15 free parameters. Naturally, as the problem size grows, the computational demands of these methods can no longer be ignored and more efficient parallel implementations have to be considered.

On the other hand, the ABC method in its current form does not seem equally capable of accomplishing the model selection task as the number of free parameters grows. Overall, it seems that successful implementation of ABC with many alternative models over large parameter spaces requires very careful tuning of the method to avoid particle degeneracy and other problems. In contrast, the other two methods operate on one model at a time, instead of sampling the joint space of models and parameters, which seems a safer option.

In any case, however, Bayesian model selection may not be able pick a unique model unambiguously. Still, the results obtained can be useful, since they may exclude certain models from the selection procedure and guide the experimental efforts towards discriminating among the remaining high-probability models.

References

[1] L. Kuepfer, M. Peter, U. Sauer, and J. Stelling, "Ensemble modeling for analysis of cell signaling dynamics," *Nature Biotechnology*, vol. 25, pp. 1001–1006, 2007.

[2] R. M. Neal, "Annealed importance sampling," *Statistics and Computing*, vol. 11, no. 2, pp. 125–139, 2001.

[3] S. Chib and I. Jeliazkov, "Marginal likelihood from the Metropolis-Hastings output," *Journal of the American Statistical Association*, vol. 96, no. 453, pp. 270–281, 2001.

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

[5] B. Magasanik and C. A. Kaiser, "Nitrogen regulation in saccharomyces cerevisiae," *Gene*, vol. 290, no. 1-2, pp. 1 – 18, 2002.

[6] I. Georis, A. Feller, F. Vierendeels, and E. Dubois, "The Yeast GATA Factor Gat1 Occupies a Central Position in Nitrogen Catabolite Repression-Sensitive Gene Activation," *Mol. Cell. Biol.*, vol. 29, no. 13, pp. 3803–3815, 2009.

[7] S. Ghaemmaghami, W. Huh, K. Bower, R. Howson, A. Belle, N. Dephoure, E. O'Shea, and J. Weissman, "Global analysis of protein expression in yeast," *Nature*, vol. 425, no. 6959, pp. 737–741, 2003.

[8] A. Belle, A. Tanay, L. Bitincka, R. Shamir, and E. K. O'Shea, "Quantification of protein half-lives in the budding yeast proteome," *Proceedings of the National Academy of Sciences*, vol. 103, no. 35, pp. 13 004–13 009, 2006.

[9] F. Holstege, E. Jennings, J. Wyrick, T. Lee, C. Hengartner, M. Green, T. Golub, E. Lander, and R. Young, "Dissecting the regulatory circuitry of a eukaryotic genome," *Cell*, vol. 95, pp. 717–728, 1998.

[10] A. Milias-Argeitis, R. Porreca, S. Summers, and J. Lygeros, "Bayesian model selection for the yeast GATA-factor network: a comparison of computational approaches," Tech. Report, ETH Zurich http://control.ee.ethz.ch/index.cgi?page=publications&action=details&id=3635.

[11] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[12] H. Chipman, E. I. George, and R. E. McCulloch, "The practical implementation of Bayesian model selection," in *Model selection*, ser. IMS Lecture Notes Monogr. Ser. Inst. Math. Statist., 2001, vol. 38, pp. 65–134.

[13] D. MacKay, *Information theory, inference, and learning algorithms*. Cambridge Univ Press, 2003.

[14] P. Del Moral, A. Doucet, and A. Jasra, "Sequential Monte Carlo samplers," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 3, pp. 411–436, 2006.

[15] M. A. Beaumont, W. Zhang, and D. J. Balding, "Approximate Bayesian Computation in Population Genetics," *Genetics*, vol. 162, no. 4, pp. 2025–2035, 2002.

[16] P. Marjoram, J. Molitor, V. Plagnol, and S. Tavar, "Markov chain Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15 324–15 328, 2003.

[17] S. A. Sisson, Y. Fan, and M. M. Tanaka, "Sequential Monte Carlo without likelihoods," *Proceedings of the National Academy of Sciences*, vol. 104, no. 6, pp. 1760–1765, 2007.

[18] A. Grelaud, C. P. Robert, and J.-M. Marin, "Abc methods for model choice in gibbs random fields," *Comptes Rendus Mathematique*, vol. 347, no. 3-4, pp. 205 – 210, 2009.

[19] P. D. Moral, A. Doucet, and A. Jasra, "An adaptive sequential monte carlo method for approximate bayesian computation," Imperial College, Tech. Rep., 2009.

[20] V. Vyshemirsky and M. A. Girolami, "Bayesian ranking of biochemical system models," *Bioinformatics*, vol. 24, no. 6, pp. 833–839, 2008.