Statistical Applications in Genetics and Molecular Biology

Volume 10, Issue 1

2011

Article 49

Bayesian Learning from Marginal Data in Bionetwork Models

Fernando V. Bonassi, *Duke University*

Lingchong You, *Duke University*

Mike West, *Duke University*

Recommended Citation:

Bonassi, Fernando V.; You, Lingchong; and West, Mike (2011) "Bayesian Learning from Marginal Data in Bionetwork Models," *Statistical Applications in Genetics and Molecular Biology*: Vol. 10: Iss. 1, Article 49.

DOI: 10.2202/1544-6115.1684

Available at: <http://www.bepress.com/sagmb/vol10/iss1/art49>

©2011 De Gruyter. All rights reserved.

Bayesian Learning from Marginal Data in Bionetwork Models

Fernando V. Bonassi, Lingchong You, and Mike West

Abstract

In studies of dynamic molecular networks in systems biology, experiments are increasingly exploiting technologies such as flow cytometry to generate data on marginal distributions of a few network nodes at snapshots in time. For example, levels of intracellular expression of a few genes, or cell surface protein markers, can be assayed at a series of interim time points and assumed steady-states under experimentally stimulated growth conditions in small cellular systems. Such marginal data on a small number of cellular markers will typically carry very limited information on the parameters and structure of dynamic network models, though experiments will typically be designed to expose variation in cellular phenotypes that are inherently related to some aspects of model parametrization and structure. Our work addresses statistical questions of how to integrate such data with dynamic stochastic models in order to properly quantify the information—or lack of information—it carries relative to models assumed. We present a Bayesian computational strategy coupled with a novel approach to summarizing and numerically characterizing biological phenotypes that are represented in terms of the resulting sample distributions of cellular markers. We build on Bayesian simulation methods and mixture modeling to define the approach to linking mechanistic mathematical models of network dynamics to snapshot data, using a toggle switch example integrating simulated and real data as context.

KEYWORDS: approximate Bayesian computation (ABC), biological signatures, dynamic stochastic network models, flow cytometry data, posterior simulation, synthetic gene circuit, systems biology, toggle switch model

Author Notes: We are grateful to the editors and two anonymous referees for constructive comments on the original version of this paper, and to Yu Tanouchi of Duke University for experimental assistance in generating the *E. coli* toggle switch data. This work was supported in part by the U.S. National Institutes of Health under grants P50-GM081883 and RC1-AI086032, and by the National Science Foundation under grant DMS-1106516. Any opinions, findings and conclusions or recommendations expressed in this work are those of the authors and do not necessarily reflect the views of the NIH and/or NSF. Computer code, with examples to replicate the applications in the paper, is available to interested readers.

1 Introduction

Much interest exists in the area of dynamic cellular network modeling in systems biology. In coming years, advances in technology that will enable access to data reflecting gene and protein expression on multiple genes in time course studies will aid advances in our understanding of dynamic cellular networks at basic biochemical levels as well as in relation to the resulting cellular phenotypes. The roles of stochastic modeling and statistical methods to integrate models with data are already to the fore in current studies, although progress is currently very limited due to sparse, incomplete data, and this will be true for a number of years. What is currently accessible in many studies, based on increasingly widely exploited technologies such as flow cytometry, is data on *marginal distributions* of a few network nodes at *snapshots* in time; for example, levels of intracellular expression of a few genes, or cell surface protein markers, can be assayed at a series of interim time points and assumed steady-states under experimentally stimulated growth conditions in small cellular systems. In several areas, including cancer biology and immunology, for example, interest in such experiments is escalating and generating such snapshot data coupled with phenotypic characteristics of cells that may be linked to genetic or environmental perturbations under study. Such marginal data on a small number of cellular markers will typically carry very limited information on the parameters and structure of dynamic network models, though experiments will typically be designed to expose variation in cellular phenotypes that are inherently related to some aspects of model parametrization and structure.

This raises statistical questions of how to integrate such data with dynamic stochastic models in order to properly quantify the information— or lack of information— it carries relative to models assumed. This paper focuses on these issues, presenting a Bayesian analysis and computational strategy coupled with a novel approach to summarizing and numerically characterizing biological phenotypes that are represented in terms of the resulting sample distributions of cellular markers. We build on Bayesian simulation methods— variants of ABC (approximate Bayesian computation)— and mixture modeling to define the approach to linking mechanistic mathematical models of network dynamics to snapshot data, using a toggle switch example integrating simulated and real data as context. The general approach is applicable to contexts where multiple independent samples of a process are observed at a particular snapshot in time with, by extension, potential to apply also to non-steady state scenarios and the analysis of data at multiple time points.

2 Models and Data

2.1 Dynamic Network Models

Consider a number of genes or proteins as nodes of a dynamic network and represent by $x_{c,t}$ the state-vector of mRNA levels of the set of genes within cell c at time t . We use single-cell, discrete-time stochastic models of the form

$$x_{c,t+h} = x_{c,t} + hF(x_{c,t}, \Theta)x_{c,t} + h\omega(x_{c,t}, \Theta)\xi_{c,t}, \quad t = 1, 1+h, \dots, \quad (1)$$

where h is a small time step, $F(\cdot, \Theta)$ is a non-linear state evolution function depending on some parameters in the global parameter set Θ , $\omega(\cdot, \Theta)$ is a possibly state-dependent noise scale factor for cell c , and the $\xi_{c,t}$ are independent stochastic noise terms perturbing the states, reflecting intrinsic noise within the cell as well as model misspecification. Variation in parameters can reflect genetic or environmental differences or perturbations at the cell population level; thus inference on parameters can link to observation phenotypes including aspects of the distributions of measured expression on some or all of the genes at some specific times. We note also that variations in parameters can be regarded as proxies for changes in the response of the network to perturbations, such as experimental growth conditions. Further, network model structure is also part of the parametrization; a non-zero activation rate in a transcriptional feedback component of the model has transcriptional strength measured by a parameter, while setting that parameter to zero deletes that network link and so represents a change in structure.

Use of such models in systems biology typically heavily utilizes forward simulation to assess the impact of different parameters—often informally and subjectively—on aspects of the resulting temporal trajectories as well as on marginal distributions of subsets of states across cells at chosen time points (e.g. You et al., 2003, Tan et al., 2009, Hallen et al., 2011). Generating detailed experimental time course data is still in its infancy; although aided by advances in single-cell imaging and data extraction (e.g. Wang et al., 2008, 2010), significant progress in experimental biology and biochemical technologies are needed to advance to the point where data will be available on more than very few genes and proteins at time resolutions that are quite limited modulo the interest in model identification and estimation.

Forward simulation and statistical model fitting will utilize a discrete-time formulation of dynamic models of some form, whether derived from fundamental biochemical dynamics or idealized continuous-time stochastic differential equations (Wilkinson, 2006). We adopt this natural discrete time formulation directly, while noting that our Bayesian analysis approach will apply

whatever modeling strategy is taken; all that is required for our approach is the ability to forward simulate the model— in a computationally efficient and, ideally, parallelized manner— across multiple cells under multiple values of the model parameters Θ and initial conditions.

2.2 Population Phenotypes and Marginal Data

Partial data observed at one or more time points can take a number of forms. The example in Yao et al. (2011) provides one context in which experimental data is very sparse indeed, albeit using deterministic rather than stochastic network models. There the outcome recorded was the occurrence of a bistable response in levels of expression of a single gene, in the context of studies of cell cycle regulation and control in the Rb-E2F network. Here bistability of the progression of cells into the cell cycle is evidenced in terms of bimodal distributions of expression of the E2F1 gene in later times once the cells progress into steady-state growth conditions (Yao et al., 2008). Using a parametrized network model involving several genes and proteins, Yao et al. (2011) defined priors over model parameters and included subnetworks defined by setting one or more parameter to zero. Very large scale simulation of these priors produced synthetic time trajectories of expression of network node genes, and the consistency with bistability of E2F1 was recorded to provide assessment of the prior predictive probability of the bistable outcome. In this case, the observed population-level phenotype is a simple indicator $S = 0/1$ of bistability. The focus is on assessing consistency of structure, rather than on updating/refining priors to posteriors over ranges of parameters, but gives an example and simple setting that partly underlies the broader development here.

Here we address similar issues in the context of data from a (or, in obvious extensions of the work, several) marginal distribution at a particular time point. We assume that noisy, marginal data is generated by observing some or all of the network nodes at a particular time point $t = T$, such as at an assumed steady-state in the growth response of a network following experimental stimulation. This defines data as a vector of observations y_c on each cell c where

$$y_c = f(x_{c,T}) + s(x_{c,T}, \Theta)\eta_c, \quad (2)$$

with observational noise η_c , possibly state and parameter dependent noise scale factors $s(x, \Theta)$ and some function $f(x)$ mapping the unobserved state vector at time T to observations. In experiments generating direct measures of expression of a subset of the genes, $f(\cdot)$ simply selects those genes; for example, in our toggle switch study, y_c is a scalar, a noisy measurement of

the first element of a 2–vector $x_{c,T}$. The data is a sample of such observations on typically thousands or tens of thousands of cells; i.e, a sample distribution $Y = \{y_c, c = 1 : C\}$ for some large sample size C and with measurements as well as state variables independent across cells.

2.3 Extended Toggle Switch Model Example

The toggle-switch is a bivariate, doubly repressive dynamic model known to reflect potentially bistable behavior in components of many bionetworks (Gardner et al., 2000, Hallen et al., 2011). The model has a 2–dimensional state vector $x_{c,t} = (u_{c,t}, v_{c,t})'$ representing cell c -specific expression of two genes, u, v , coupled in non-linear dynamic behavior represented by the equations below. Our experimental data sets from a bacterial model incorporating a synthetic toggle switch circuit provides observations on many cells but on only one of the two genes (u) at a fixed, assumed state-state time point $t = T$. The forms of dependence and noise characteristics represent decay of mRNA, reciprocal repression and increased levels of observation noise at lower levels. An extension of the canonical toggle-switch model to incorporate realistic levels of measurement noise– with increased variation at very low levels corresponding to basal level synthesis of the protein– is adopted.

Specifically, for each cell $c = 1 : C$ consider potential measurements $y_c = (y_{c,u}, y_{c,v})'$ where

$$\begin{aligned} y_{c,u} &= u_{c,T} + \mu + \mu\sigma\eta_{c,u}/u_{c,T}^\gamma, \\ y_{c,v} &= v_{c,T} + \mu + \mu\sigma\eta_{c,v}/v_{c,T}^\gamma, \end{aligned} \quad (3)$$

where for each $t \leq T$, the state $x_t = (u_{c,t}, v_{c,t})'$ evolves according to model equations

$$\begin{aligned} u_{c,t+h} &= u_{c,t} + h\alpha_u/(1 + v_{c,t}^{\beta_u}) - h(\kappa_u + \delta_u u_{c,t}) + h\tau_u\xi_{c,u,t}, \\ v_{c,t+h} &= v_{c,t} + h\alpha_v/(1 + u_{c,t}^{\beta_v}) - h(\kappa_v + \delta_v v_{c,t}) + h\tau_v\xi_{c,v,t}, \end{aligned} \quad (4)$$

based on some initial state and where the noise processes η . and ξ . are independent and mutually independent sequences. Practical considerations require that the states and observations be positive, so we truncate the conditional normal distributions for state evolution and data to ensure positivity. In theoretical manipulations of the model, this introduces technical complications, whereas from the viewpoint of forward model simulation– a cornerstone of the statistical computations we develop– this is essentially trivial.

Of the several model parameters, μ represents a low, basal level and (σ, γ) define increased spread of responses at lower levels. The (α, β) parameters define logistic-like repressive responses and play the central roles in

determining the toggle switch behavior, while the (κ, δ) parameters related to mRNA (or protein) decay. For differing values across the (α, β) parameter space, the model is capable of generating unimodal or bimodal steady-state distributions in (u_T, v_T) , and in the corresponding observed data $y_{c,u}$ on the u margin alone. Although very simplified versions of this model are open to theoretical delineation of parameter regions consistent with bistability, and hence bimodality of the steady-state distributions (Gardner et al., 2000), this is not true in our more general model context; the theory is intractable and this kind of question has to be explored by simulation.

Some examples of simulated data sets appear in Figure 1. These adopt a time discretization of $h = 1$ over an interval of $T = 300$ steps to steady-state, initialized at $x_0 = (10, 10)'$. The $p = 13$ model parameters of interest are

$$\Theta = \{\mu, \sigma, \gamma; \alpha_u, \beta_u, \kappa_u, \delta_u, \tau_u; \alpha_v, \beta_v, \kappa_v, \delta_v, \tau_v\}. \quad (5)$$

The first three relate to noise characterization while the others represent control and differentiate bistability in terms of their roles in leading to varying levels of the steady-states (u_T, v_T) as well as the modality of the steady-state bivariate distribution.

Figure 1: Contoured steady-state data distributions of noisy outcomes $(y_u, y_v)'$ of $(u_T, v_T)'$ based on simulation of the model in equations (3) and (4). The three distributions are based on $\kappa_g = 1$, $\delta_g = 0.03$ and $\tau_g = 0.5$ for $g = u, v$ with, from left to right, the remaining parameters $\{\mu, \sigma, \gamma, \alpha_u, \beta_u, \alpha_v, \beta_v\}$ given by $(258.5, 0.49, 0.28, 38.1, 4.0, 40.8, 4.9)$, $(430.4, 0.44, 0.07, 26.9, 2.8, 13.4, 1.1)$ and $(473.5, 0.34, 0.30, 11.3, 2.7, 25.7, 4.1)$. Samples of size 1,000 were simulated and contoured for presentation.

Figure 2: Histograms of measured outcomes $y_{c,u}$ for several thousand cells in 4 of 10 experimental data sets involving a bacterial toggle switch network.

Figure 2 shows experimental data measuring steady-state levels of a single gene from a series of 10 experiments on bacterial cells. This data involves *E. coli* JM2.3000 cells that harbour an engineered synthetic toggle switch network based on the pTAK131 plasmid (Gardner et al., 2000). Cell cultures were grown under IPTG perturbations that induce a range of “snapshot” distributional phenotypes consistent with differing regions of parameter space in the dynamic model. The data in Figure 2 represent the corresponding marginal distribution of $y_{c,u}$ in our notation; these 4 experiments show both unimodal and bimodal distributions at the chosen time point. We assume these to be approximately steady-state although whether they are truly in equilibrium is a side-issue from the viewpoint of the methodology development here; the time point selected for assays is the same across experiments. In the context of the model of equations (3) and (4), these data clearly contain some information relevant to constraining/updating priors over the model parameters to reflect the different qualitative outcomes observed.

3 Bayesian Analysis and Computation

3.1 Strategy

Consider a specified model of the form of equations (1) and (2) with Θ representing all unknown parameters, subject to a specified prior with density $p(\Theta)$. Experimental data is the form of a random sample $Y_{obs} \equiv Y = \{y_c, c = 1 : C\}$ at some specified snapshot time T , and we will typically be interested in approximating posterior inferences for each of several data sets in experimental comparisons.

Direct development of standard Bayesian computation via simulation methods is not an option as the likelihood function $p(Y_{obs}|\Theta)$ is defined implicitly and so cannot be evaluated; standard approaches of importance sampling or Markov chain Monte Carlo Metropolis Hastings typically rely on likelihood evaluations. Direct rejection-based ABC methods, ABC-MCMC or sequential Monte Carlo variants (Marjoram et al., 2003, Sisson et al., 2007, Blum and François, 2010, Nunes and Balding, 2010) are useful in such contexts where there is an ability to simulate a model but not evaluate likelihood functions. They are of increasing interest in biological system applications, in particular (e.g. Toni and Stumpf, 2010, Wilkinson, 2011). Our computational strategy utilizes and extends ABC, involving the two defining ingredients: (i) forward simulation of the dynamic model, i.e., an ability to easily and cheaply simulate a large sample Y at time T from the implicit $p(Y|\Theta)$ at any chosen Θ ; and (ii) the use of a summary measure of discrepancy between any synthetic data set and any observed data Y with the latter represented in terms of a reduced dimensional summary statistic $S = S(Y)$.

We address some of the issues in using traditional ABC with statistical and biologically motivated steps, as follows.

- Define a set of global *reference distributions* with density functions $f_r(y_c)$, ($r = 1 : R$), representing a range of biologically plausible cell population response distributions. For any real or simulated data set Y , measure consistency with $f_r(\cdot)$ via an appropriate statistical metric $s_r = s_r(Y)$. Treat the resulting *reference signature* vector $S \equiv S(Y) = (s_1, \dots, s_R)'$ as a summary statistic to be used in an ABC analysis: with $S_{obs} = S(Y_{obs})$, measure closeness via a distance measure $\delta(S, S_{obs})$.
- Use large-scale prior:model simulation to generate

$$\{ \Theta_i, Y_i = (y_{i,c}, c = 1 : C) \}, \quad i = 1 : P, \quad (6)$$

where P is large and each $y_{i,c}$ is generated by forward model simulation under parameter Θ_i . The resulting large sample from the prior predictive distribution of $\delta(\cdot, S_{obs})$ is used to calibrate ABC in defining what “close” means in restricting attention to the corresponding region of Θ space.

- Use flexible multivariate mixture modeling to approximate, and hence emulate, joint distributions of (Θ, S) in such regions and induce the smoothed, nonlinear regressions $p(\Theta|S)$ so implied. This defines direct posterior approximations $p(\Theta|S_{obs})$ that aim to improve upon traditional ABC methods via theoretical interpolation; this also defines potential Metropolis proposals for further adaptive refinement.

We stress that reduction of the observed sample data Y to summary statistics $S = S(Y)$ is necessary in order to advance the Bayesian computations and reflects a modification of the inference goal; we now aim to approximate the implied posterior $p(\Theta|S)$ at $S = S_{obs}$. We now detail the above steps and the key ideas underlying them.

3.2 Reference Phenotypic Signatures

3.2.1 Representative synthetic data sets

The key idea here is to specify an encompassing set of reference distributions that vary in forms to reflect real phenotypic variation when viewing the marginal distribution itself as the phenotype. The $f_r(y_c)$ may be based on real data, or model simulations at carefully selected parameters. The goal is to represent the full range of biologically plausible cell population response distributions, spanning the space of outcomes that show practical differences and that are consistent with the model and prior specification. We may use past data sets if available. As a general strategy, selecting references based on the prior predictive model simulations is a natural strategy we have adopted and that can be automated.

Figure 3 gives an example from the toggle switch study, where y_c is univariate and so the $f_r(y_c)$ are univariate densities. The figure displays histograms of a selection of data sets simulated from the prior: model combination and that represent the range of biologically meaningful variations in this phenotype; the distributions include a range of unimodal and bimodal forms with variation in location and spread around modes. This set of $R = 100$ histograms representing diversity of steady-state cellular phenotypes is selected via the following automated procedure.

Figure 3: Histograms of $R = 100$ reference steady-state distributions for $y_{c,u}$ selected from prior predictive simulation samples from the toggle switch model. These represent and span a range of phenotypes with unimodal and bimodal outcomes having varying location and shape characteristics.

First, create a binned representation of each of the simulated data sets Y_i of equation (6) on a specific range of outcome values. Do this using a histogram with a relatively large number of bins; in cases of more than one dimensional data, vectorize the bin counts from multidimensional histograms. For each i this defines a vector z_i of some specified length K that form the rows of a resulting $P \times K$ array Z of histogram frequencies. Next, perform the singular value decomposition $Z = ADF$ where A has orthogonal columns, the diagonal matrix D has non-zero diagonal entries ordered $d_1 \geq d_2 \geq \dots$ and F is orthonormal. The rows of F are the singular factors underlying the overall variation among outcome distributions and the columns of A define loadings of each of the distributions on these factors. Then, select a smaller number R of representative references using indices $i_1, \dots, i_R$ that define the R percentiles of the values in the first column of A . Variation along this column spans the range of dominant aspects of variation exhibited in the full set of synthetic data sets, so this will provide a smaller set of “representatives” capturing this phenotypic variability.

We use this approach in the toggle example below using an initial prior: model simulation of size $P = 25,000$ and $R = 100$ references. There and in the discussion of Section 5 we discuss some of the practicalities includ-

ing questions of robustness to the choice of R and variants of this approach that have been extensively explored in this context; these evaluations led to adoption of this specific, simply automated strategy.

3.2.2 Smoothing reference samples

Given a specified set of reference distributions in of data histograms, whether from real data or simulation samples, create continuous densities $f_r(\cdot)$ for signature evaluation. Here we do this using normal mixtures models which applies to general multivariate y_c as well as the specific univariate example in the toggle switch study. We utilize very computationally efficient code for flexible, semi-parametric Bayesian mixtures (Suchard et al., 2010, Cron and West, 2011) to do this, as in other applications in flow cytometry data analysis (Chan et al., 2008).

3.2.3 Reference signature definition and evaluation

Given the reference densities and any test sample $Y = \{y_c, c = 1 : C\}$, define the *reference signature* vector $S(Y) = (s_1(Y), \dots, s_R(Y))'$ to measure consistency of Y with each of the references and represent the data in this R -dimensional phenotypic space. The formal statistical reference signature adopted is the per-cell log likelihood

$$s_r \equiv s_r(Y) = C^{-1} \sum_{c=1:C} \log(f_r(y_c)). \quad (7)$$

This places the data set Y in the space of reference biological “models” in a standard statistical sense; for any sample size C , this is a scalar multiple of the log likelihood of a “model” defined by reference density $f_r(\cdot)$ and viewing the test sample Y as data—a very natural statistical measure of concordance between the test data and the reference distribution. At a theoretical level is also of interest to note a limiting interpretation. Denote by $F(\cdot)$ the true underlying distribution function of the y_c . Then, as $C \rightarrow \infty$ the weak law of large numbers implies that s_r converges in probability to $\int \log(f_r(y))dF(y) = -K_r(F) + c_F$ where c_F depends on F but not the reference r and $K_r(F)$ is the Kullback-Leibler divergence (or relative entropy) of reference distribution r from F , a natural theoretical measure of concordance of the test distribution with the reference.

A useful technical modification enables further reduction in dimension and complexity of the reference signature vectors, with computational as well as statistical advantages, as follows. Compute the signature vector $S(Y_i)$ on

each of the P synthetic data sets to form the $P \times R$ matrix T with rows $S(Y_i)'$. Perform a singular value decomposition of T and assess variability in signatures across these data sets explained by the resulting principal components. In cases of redundancy among the R reference distributions, we can then choose a smaller number $r \leq R$ of principal components to use as transformed signatures. That is, for any future data set Y , compute $S(Y)$ and then map to a $HS(Y)$ where H is the $r \times R$ matrix defining the map from signatures to the first r dominant principal components in the prior simulation data sets. This yields benefits in terms of reducing computations in the mixture model fitting and ABC analysis (next section) and in reducing statistical redundancies arising from the initial selection of R references. This modification is used in the toggle switch example below; there the first $r = 11$ principal components of the $25,000 \times 100$ signature matrix T explain over 99.5% of the total variation; so, the analysis uses the resulting 11×100 matrix H to map initial signature vectors to this reduced dimension.

For simplicity of notation, we maintain the use of $S = S(Y)$ for reference signatures of dimension R whether they are based on the full set of references or projected and reduced in dimension this way.

As mentioned and stressed above, we are now in a context of aiming to develop approximate computations representing $p(\Theta|S_{obs})$ having mapped from Y to the reduced dimensional, scientifically motivated space of signatures $S = S(Y)$. In general, the aim and idea is that S captures all salient features of Y so as to form an approximate sufficient statistic in the formal sense, although the current methodological development does not assume nor directly address that theoretically; rather, we base the development on the perspective that S itself is the data to process for inferences on model parameters.

3.3 Large-Scale Simulation and ABC Strategies

Traditional ABC rejection draws $\{\Theta, Y\}$ from the prior: model and accepts Θ as an approximate draw from $p(\cdot|Y_{obs})$ if $\delta(S(Y), S_{obs}) < \epsilon$ for some distance $\delta(\cdot, \cdot)$ and small $\epsilon > 0$ (Marjoram et al., 2003). In our context we take a somewhat different view; since prior: model simulations are computationally cheap we simulate large samples as in equation (6) to more adequately explore the prior predictive distribution implied for Y . This has two important features: first, and critically, it provides large samples from the prior predictive distribution of $\delta(\cdot, S_{obs})$ that allow us to address a main challenge of calibrating ABC and assigning practically relevant values of ϵ ; second, it leads to constrained prior sample parameters whose sampled outcomes are “close”

to the observed signature, providing opportunity for more effective adaptation towards approximate posterior sampling. We comment further on these two aspects. The analysis depends on a chosen distance measure which we take as Euclidean distance between the R -dimensional signature vectors.

To elaborate, note first that the reference signatures on the observed data, $S(Y_{obs})$, and on each of the large prior sample of synthetic data sets, $S(Y_i), i = 1 : P$, define a sample of size P from the prior predictive distribution of $\delta(S, S_{obs})$. We can use this to calibrate “closeness” of synthetic to observed data sets; if ϵ is taken as the lower 5% quantile, thresholding leads to a constrained sample $\{\Theta_i, Y_i\}$ of size $P/20$ corresponding to the closest 5% synthetic data sets. We can use this feature in a number of ways. One strategy is to use a relatively relaxed threshold, such as 5 or 10% quantiles, and then use ABC-MCMC like moves applied to each of the parameters defined by thresholding. This requires an MCMC kernel to move parameters; one that is attractive is to fit a multivariate normal (or T) mixture to the thresholded parameters and use that as an independence Metropolis sampling proposal distribution. This has been found useful in this context, and builds on the use of such mixture strategies in pre-MCMC adaptive importance sampling (West, 1992, 1993b,a). Importantly, we are able to fit flexible semi-parametric Bayesian mixture models in moderate dimensions (e.g., up to 20 – 30 dimensions) using computationally efficient code (Suchard et al., 2010, Cron and West, 2011).

A second approach is to simply use the thresholded parameter samples as a direct posterior approximation. This makes sense when the thresholded number is relatively large. Then, some form of local smoothing of this reduced set of samples is typically desirable and, again, we find mixture model-based smoothing attractive. In particular, following recent related ideas of local smoothing for interpolation of ABC-thresholded parameters (e.g. Beaumont et al., 2002, Blum and François, 2010), the use of nonlinear regression smoothing based on fitting Bayesian mixture models (Müller et al., 1996) is recommended. The idea here is simple and this innovation represents a distinction of the current approach relative to traditional ABC and these recent interpolation methods, introducing smoothing for interpolation in thresholded regions of the $\{\Theta, S\}$ space near the observed $S = S_{obs}$. As the ABC-thresholded subsample of the $\{\Theta_i, Y_i\}$ is a random sample from a distribution concentrated in the region of the true posterior, fit a flexible multivariate mixture in (Θ, Y) jointly, say $g(\Theta, Y)$. Use this to induce, theoretically, the implied conditional mixture $g(\Theta|Y)$ and evaluate that at $Y = Y_{obs}$. With mixtures of multivariate normal or T distributions, this leads to analytic conditional forms that are normal or T, and whose conditional means define nonlinear kernel-like regression of Θ on Y in the constrained region over which Y is “close” to Y_{obs} . The

conditional mixture $g(\Theta|Y)$ can be easily evaluated and simulated. This can then be used either as a direct posterior approximation, so generating a local interpolation to overlay the initial use of rejection-based ABC, or as an independent Metropolis proposal density for further adaptation via ABC-MCMC.

4 Toggle Switch Model Analyses

4.1 Initial Discussion and Examples

Our example uses the model of Section 2.3 as already partly illustrated. Now Θ is the set of $p = 13$ parameters in equation (5) and the prior is a product of univariate uniforms with lower and upper bounds carefully chosen so that prior predictions reflect a wide and meaningful range of phenotypic outcomes. We stress the importance of evaluation of prior and contextual information to advise on the specification of prior distributions as critical elements of the overall model formulation. We regard and interpret priors as subjective, scientifically informed representations of belief throughout. Specification naturally involves a combination of subjective evaluation of literature-based information combined with experimentation with model simulations based on ranges of possible priors, the latter being particularly useful in ruling out regions judged implausible based on synthetic data they generate. We also note again that the data in this case study is a large steady-state sample on just one of the two genes, $y_{c,u}$.

The definition of reference distributions and signature evaluation follows Section 3.2, with some details as follows. A first set of 25,000 prior model simulations was created with $C = 1,000$ cells in each steady-state sample. Each of these defines a histogram on 50 equally spaced bins and these binned data sets were used to create 100 initial reference distributions as described in Section 3.2.1. Variations of the number of bins and the number of references were examined extensively to inform these choices. A reasonably large number of bins in the histograms is needed to capture phenotypic variation, and the resulting analysis summaries are very similar with repeat analyses using in the range of 40-50 bins. Choosing more than 100 reference distributions quickly leads to increased redundancies that add computationally but do not change or improve the characterization of the range of phenotypic variation the references aim to reflect; reducing below 100 begins to quickly lose the ability to reflect key differences. Subjective comparison of the selected reference histograms with the real bacterial data sets provided further assurance of the ability of this set to represent key and realistic features of toggle switch

derived data as well as to encompass the theoretically known modalities of outcomes on the y_u scale; see Figure 3.

For each of the 100 reference distributions, we fitted a Gaussian mixture with up to 5 components using the Bayesian EM method (Suchard et al., 2010, Cron and West, 2011); this defines the densities $f_r(\cdot)$ as discussed in Section 3.2.2. These then yield initial 100-dimensional signatures on all of the prior simulated data sets. The strategy detailed in Section 3.2.3 was then applied, indicating that the first $r = 11$ principal components of the $25,000 \times 100$ signature matrix explain over 99.5% of the total variation in the raw 100-dimensional signatures across the 25,000 samples. Hence we use the resulting linear transformation to map to 11-dimensional effective signature vectors S .

Following this first stage definition of references and signatures, we increased the prior: model simulations to a full set of $P = 200,000$ samples each with $C = 1,000$ cells in the steady-state sample. This expanded prior sample size gives us greater coverage of the outcome space for the ABC and mixture modeling analysis; in particular, the much larger prior sample ensures that we have relatively large samples remaining after ABC thresholding. The reduced dimensional signatures were then computed on this extended prior sample of 200,000 synthetic steady-state data sets.

From analysis of one of the 10 bacterial data sets, Figure 4 shows aspects of the prior predictive distribution via scatter plots and histograms of elements of the $\{\Theta_i, Y_i\}$. Also shown are summary conditionals based on ABC-thresholding to those samples satisfying $\delta(S(Y_i), S_{obs}) < \epsilon$ where ϵ is the lower 5% quantile of the prior predictive distribution of the distance. This shows first adaptation of parameters into a region close to that supported by the true posterior, and the utility of large predictive simulation samples. Figure 5 shows the effect of model-based smoothing of the ABC-thresholded sample using a mixture model fitted in the 24 dimensions to the 5%-closest among the ABC sample; see also Figure 6. We use an encompassing multivariate Gaussian mixture based on flexible nonparametric Bayesian methods; $g(\Theta, S)$ uses parameters taken as modal estimates from a Bayesian EM algorithm applied in a Dirichlet process mixture (Suchard et al., 2010, Cron and West, 2011). We fit the normal mixture to real-valued transformations of the parameters so that the resulting emulator of the posterior is properly constrained to the finite parameter region over which we defined uniform priors. The implied conditional mixture $g(\Theta|S_{obs})$ is then easy to compute and simulate; samples from this in Figure 5 show the adaptation to define an approximate posterior sample using this model-based interpolation.

Figure 4: Analysis of one of the bacterial toggle switch data sets. Frames 4(a) and 4(c) give scatter plots of prior predictive samples in selected dimensions of the 24-dimensional distribution over 13 parameters and the reduced 11 summaries of phenotypic signatures. A subsample of the $P = 200,000$ prior samples is shown in black. Following a 5% ABC-thresholding, the red scatter represents the resulting conditional distribution, with corresponding histograms for the two model parameters in frames 4(b) and 4(d).

Figure 5: Frames 5(a) and 5(c) display scatter plots corresponding to those in Figure 4; now, however, the points in grey represent the 5% ABC-thresholded conditional distribution (in red in Figure 4) while those in blue are samples from the mixture interpolation and posterior emulator $g(\Theta|S_{obs})$ approximating the true posterior. Frames 5(b) and 5(d) show the corresponding approximate marginal posteriors for the two model parameters, with the mixture adapted posterior approximation showing clearly increased precision and resolution.

Figure 6: Mixture based smoothing for interpolation. The frame shows a 3D scatter plot of samples from the 5% ABC-thresholded region in the margin for 2 reference signatures and 1 parameter. The point clouds indicate the suitability of analytic smoothing using mixtures of multivariate normals. The mixture model smoother is used in the full 24 dimensions and is fitted to the ABC-thresholded sample of size 10,000; the points here are just a random subsample for visualization of the main idea.

4.2 Simulation Examples using Synthetic Data Sets

Figures 7 and 8 display summaries from two analyses: each uses one of the simulated data sets Y_i as a synthetic steady-state sample Y_{obs} , the point being to demonstrate the adaptive Bayesian analysis in terms of parameter inference in a context where the “true” parameters are known. The two synthetic data sets were chosen so that one is unimodal and the other bimodal, the latter reflecting toggle parameters in a region consistent with bistability.

The initial prior: model simulation defines a parameter: signature sample of size 200,000; recall that the parameters have independent uniform priors. This prior sample is then thresholded using the ABC approach to identify the 10,000 samples whose corresponding Y_i are “closest” to the chosen data set Y_{obs} . These 5% parameter regions imply marginal conditional densities represented by the unshaded histograms in the two figures. Fitting a multivariate mixture to these 10,000 samples leads to the conditional mixture $g(\Theta|Y_{obs})$; directly sampling this mixture yields approximate posterior draws whose margins are

Figure 7: Summaries of analysis of a unimodal synthetic data set (upper left) showing posterior margins for the 13 toggle switch model parameters. On each, the true known parameter value is marked as a triangle. Unshaded histograms are conditionals defined by 5% ABC-thresholding of prior samples; black histograms are of Monte Carlo draws from the mixture posterior emulator fitted to the ABC samples.

Figure 8: Summaries of analysis of a bimodal synthetic data set consistent with toggle model parameters defining bistability. Details of the figure are as in Figure 7.

represented by the black histograms in the figure. Relative to the uniform priors, note the ABC thresholding step clearly begins to identify parameter regions around the known true values, while the mixture adaptation step then modifies this to an approximate posterior that concentrates a good deal more, in or around the true values, for some of the parameters. Repeat evaluations using multiple synthetic draws confirm that resulting approximate 90 and 95% posterior credible intervals yield coverage close to nominal.

With the unimodal data set, Figure 7 shows more substantial learning on baseline location μ , scale σ and Hill coefficients β_u, β_v than for other parameters. Parameters for which the data are relatively uninformative will have posteriors close to the uniform priors, as evidenced particularly for the δ, κ and τ parameters in this example. In the case of the bimodal data set of Figure 8, there is substantial learning evident for several parameters, indicating that the information content of steady-state data in bistable cases indeed has the potential to identify meaningful constraints on model parameters in this formal statistical sense.

4.3 Bacterial Data Sets

Further evaluation of the analysis using the set of 10 experimental data sets highlights several additional features. In addition to driving the new adaptive ABC-like computational approach, the novel reference signature methodology provides some insights into the general questions of model adequacy and fit. Figure 9 displays 2 of the real data sets, together with representations of the synthetic data sets that are closest in terms of the metric $\delta(S_i, S_{obs})$ defined by the reference signatures and used in the adaptive simulation-based analysis. It is clear that, among the 200,000 prior:model simulations, there are synthetic outcomes very similar to these real data sets and the same is true for the remaining 8 bacterial data sets.

The analysis also exemplifies the ability of steady-state marginal data to inform on at least some aspects of model parametrization, with the mixture adaptive ABC analysis refining the initial uniform priors to posterior emulators that probabilistically locate the each of the experimental data sets on relevant, refined regions of model parameter space. Figure 10 provides illustration of one aspect of this, showing interval summaries of the approximate marginal posterior for some of the toggle switch model parameters plotted in 3D, color coded for each of the 10 data sets. Recall again that the priors are uniform and span the scales used in this figure. Learning about some of the parameters is evident. In particular, 8 of the cases are grouped together

Figure 9: The upper frames 9(a) show the histogram and empirical cdf (red curve) for the outcomes Y_{obs} in 1 of the 10 bacterial toggle switch data sets. Overlaid in the upper center frame is the empirical cdf (black curve) of that synthetic steady-state data set Y_i closest to Y_{obs} in the reference signature-based metric $\delta(S_i, S_{obs})$. The upper right hand frame shows a set of such curves representing the closest 30 synthetic data sets. The lower frames 9(b) show similar displays for a second bacterial data set, this one consistent with bistability.

Figure 10: Interval summaries (one standard deviation interval around the mean) of the approximate marginal posterior for some of the toggle switch model parameters, displayed in 3D to show some aspects of adaptation from initial priors that are uniform across the scales shown in each parameter dimension. The 2 cases (in green and blue, respectively) that sit apart from the rest are from the analyses of the 2 of the 10 experimental data sets that have unimodal outcomes so show no evidence of bistability. The remaining 8 cases are all bimodal and hence favor parameters in somewhat different regions consistent with bistability of the toggle switch model.

though showing some differences; these corresponding to the 8 bacterial data sets that have bimodal outcomes, so the parameters reflect bistability in the toggle model. The other 2 cases generate unimodal phenotypes; these lead to approximate posteriors that locate these cases in somewhat different regions of the parameter space, while also reflecting more limited parameter learning.

5 Further Discussion

Evaluation of formal models of dynamic cellular networks based on marginal data sets is becoming topical and important in systems and synthetic biology with an increasing ability to generate large data sets using flow cytometry methods, among others. Adaptive Bayesian simulation methods using variants of ABC have emerged in a number of areas of biology in recent years, and our work builds on these developments with a focus on marginal data integration. The two main contributions are *(i)* the definition and use of biologically inspired, relevant and interpretable summary measures of marginal outcomes, in terms of reference signatures, and *(ii)* the use of large-scale prior:model simulations to calibrate ABC methods, and novel mixture model-based adaptation and interpolation to refine ABC-based analysis. Our analyses show how the approach can be used to integrate marginal data with dynamic stochastic models in order to properly quantify the information– or sometimes the lack of information– it carries relative to models assumed.

Readers familiar with statistical work in genomics and other high-throughput molecular areas will recognize that the concept and use of the term “signature” is consistent with its development in other areas of biomolecular statistics. A signature defined as a biologically descriptive and predictive set of numerical summaries of a higher-dimensional phenotype is well-established (e.g. Huang et al., 2003, Lucas et al., 2009, Chen et al., 2010) and inspired the ideas proposed here. The construction of signatures using a biologically relevant spectrum of reference distributions, coupled with the obvious formal statistical placement of a new data set on that spectrum via equation (7), is specific to the marginal data context though may well be more broadly relevant in other areas in future.

The Bayesian computational strategy uses very large prior:model simulations to generate parameters and synthetic data sets to threshold using ABC methods and then interpolate using flexible mixture models. The more typical sequential use of ABC rejection and ABC-MCMC methods can be rather poor in terms of rejection rates and efficacy in adapting to observed data. Engineering and applied mathematics communities engaged in systems biology

research embrace aggressive computation and massive prior: model simulation and our work links to that tradition. Generating very large prior: model samples provides opportunities to explore and understand the diversity of model predictions and aspects of consistency with expected or observed biological phenotypes at a general level, as well as with any one specific observed data set. This also provides a key advance through the opportunity to explore the prior predictive distribution of ABC metrics, enabling calibration and understanding of the topology of phenotypic space defined by prior: model simulations thresholded according to a quantile of the assumed distance.

On theoretical issues related to our approach and to other variants of ABC generally, there is interest in the questions of how well a resulting posterior $p(\Theta|S_{obs})$ describes $p(\Theta|Y_{obs})$ when there is no explicit theoretical guarantee that $S = S(Y)$ resembles a sufficient statistic. Developing formal theoretical characterizations of the extent to which a chosen reference signature set defines an approximate sufficient data summary seems challenging, although the general question is emerging as a research agenda among researchers in the ABC and sequential Monte Carlo. Our approach and new strategy for defining data summaries may promote new directions here. From the current methodological perspective as earlier described, the focus is rather to create references so that S captures all practically salient features of Y and then base inferences directly on S_{obs} as the “data” itself.

In the toggle switch example we discussed some key aspects of specification of the reference distributions and signature evaluation of Section 3.2, particularly on details of reference selection of Section 3.2.1. We note here that there are obvious generalizations of that strategy that may be explored and we are currently exploring refinements in the toggle example and with other models. In particular, the SVD-based analysis selects references on a scale defined by the first principal component underlying variation in simulated data sets. It is natural to ask if the overall analysis is practically modified by adding in consideration of higher-order principal components. We have explored this in a number of ways and are satisfied that the overall results of the toggle model example are robust, partly as the range of forms of data distributions is well understood in this context. However, in more complex models with multidimensional outputs and more parameters, refinements of this sort in the definition of the set of reference distributions may be needed.

Turning to computation, we stress again that forward simulations are inherently massively parallelizable and so it makes little sense to ignore these opportunities for deeper and detailed understanding of model implications *a priori*; our overall analysis strategy embraces this view. With rapidly advancing software and systems tools for utilizing multi-core and GPU processing

as well as clusters and cloud-based facilities, we are in a position to generate incredibly large simulations easily and cheaply; in a context where we aim to then analyze several or many real data sets against one prior: model specification, this is particularly appealing, as large prior samples provide the opportunity to “get in the ballpark” of any one data set, and then follow-up with more refined, adaptive analyses.

Fitting mixture models to ABC-thresholded simulations provides a number of opportunities noted in the paper, in addition to the direct use of fitted mixtures to define direct posterior emulators as demonstrated in our examples. This component of the work represents a key distinction relative to standard ABC, using mixture-model based smoothing for interpolation in thresholded regions of the $\{ parameter, signature \}$ space near the observed data in order to deduce the theoretical approximation to the target posterior. Of a number of potential uses and directions for methodological refinements, developing the use of such mixture emulators as Metropolis proposals for adaptive ABC-MCMC seems promising, building as it does on both the past success of adaptive importance sampling using flexible mixture approximations and coupling into the recent work on ABC-MCMC methods generally. We are also interested in potential extensions of the approach that use priors with point masses at zero as a general strategy for introducing dynamic model selection/evaluation; clearly the use of normal mixture approximations will need modification to include such priors. Additional current questions include extensions of the analysis to integrate additional data into models, such as based on one or more cellular samples of “interim” marginal distributions at a selection of time points not restricted to the steady-state.

References

- Beaumont, M. A., W. Zhang, and D. J. Balding (2002): “Approximate Bayesian computation in population genetics,” *Genetics*, 162, 2025.
- Blum, M. G. B. and O. François (2010): “Non-linear regression models for Approximate Bayesian Computation,” *Statistics and Computing*, 20, 63–73.
- Chan, C., F. Feng, J. Ottinger, D. Foster, M. West, and T. B. Kepler (2008): “Statistical mixture modelling for cell subtype identification in flow cytometry,” *Cytometry, A*, 73, 693–701.
- Chen, J. L., D. Merl, M. West, and J. T. A. Chi (2010): “Lactic acidosis triggers starvation response with paradoxical induction of TXNIP through MondoA,” *PLoS Genetics*, 6, e1001093.

- Cron, A. J. and M. West (2011): “Efficient classification-based relabeling in mixture models,” *The American Statistician*, 65, 16–20.
- Gardner, T. S., C. R. Cantor, and J. J. Collins (2000): “Construction of a genetic toggle switch in *Escherichia coli*,” *Nature*, 403, 339–342.
- Hallen, M., B. Li, Y. Tanouchi, C. M. Tan, M. West, and L. You (2011): “Computation of steady-state probability distributions in stochastic models of cellular networks,” *PLoS Computational Biology*, 7, e1002209.
- Huang, E. S., M. West, and J. R. Nevins (2003): “Gene expression phenotypes of oncogenic pathways,” *Cell Cycle*, 2, 415–417.
- Lucas, J. E., C. M. Carvalho, and M. West (2009): “A Bayesian analysis strategy for cross-study translation of gene expression biomarkers,” *Statistical Applications in Genetics and Molecular Biology*, 8, Article 11.
- Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003): “Markov chain Monte Carlo without likelihoods,” *Proceedings of the National Academy of Sciences USA*, 100, 15324–15328.
- Müller, P., A. Erkanli, and M. West (1996): “Bayesian curve fitting using multivariate normal mixtures,” *Biometrika*, 83, 67–79.
- Nunes, M. A. and D. J. Balding (2010): “On optimal selection of summary statistics for Approximate Bayesian Computation,” *Statistical Applications in Genetics and Molecular Biology*, 9, Article 34.
- Sisson, S. A., Y. Fan, and M. M. Tanaka (2007): “Sequential Monte Carlo without likelihoods,” *Proceedings of the National Academy of Sciences USA*, 104, 1760–1765.
- Suchard, M. A., Q. Wang, C. Chan, J. Frelinger, A. J. Cron, and M. West (2010): “Understanding GPU programming for statistical computation: Studies in massively parallel massive mixtures,” *Journal of Computational and Graphical Statistics*, 19, 419–438.
- Tan, C. M., P. Marguet, and L. You (2009): “Emergent bistability by a growth-modulating positive feedback circuit,” *Nature Chemical Biology*, 5, 842–848.
- Toni, T. and M. P. H. Stumpf (2010): “Simulation-based model selection for dynamical systems in systems and population biology,” *Bioinformatics*, 26, 104–110.
- Wang, Q., J. B. Niemi, C. M. Tan, L. You, and M. West (2010): “Image segmentation and dynamic lineage analysis in single-cell fluorescent microscopy,” *Cytometry A*, 77, 101–110.
- Wang, Q., L. You, and M. West (2008): “Celltracer: Software for automated image segmentation and lineage mapping for single-cell studies,” URL: www.stat.duke.edu/research/software/west/celltracer/.

- West, M. (1992): “Modelling with mixtures (with discussion),” in J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, eds., *Bayesian Statistics 4*, Oxford University Press, 503–524.
- West, M. (1993a): “Approximating posterior distributions by mixtures,” *Journal of the Royal Statistical Society (Ser. B)*, 54, 553–568.
- West, M. (1993b): “Mixture models, Monte Carlo, Bayesian updating and dynamic models,” *Computing Science and Statistics*, 24, 325–333.
- Wilkinson, D. J. (2006): *Stochastic Modelling for Systems Biology*, London: Chapman & Hall/CRC.
- Wilkinson, D. J. (2011): “Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology (with discussion),” in J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. David, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 9*, Oxford University Press, 679–705.
- Yao, G., T. J. Lee, S. Mori, J. R. Nevins, and L. You (2008): “A bistable Myc-Rb-E2F switch: a model for the restriction point,” *Nature Cell Biology*, 10, 476–482.
- Yao, G., C. Tan, M. West, J. R. Nevins, and L. You (2011): “Origin of bistability underlying mammalian cell cycle entry,” *Molecular Systems Biology*, 7, 485.
- You, L., A. Hoonlor, and J. Yin (2003): “Modeling biological systems using Dynetica— a simulator of dynamic networks,” *Bioinformatics*, 19, 435–436.