BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 1 2010, pages 104–110
doi:10.1093/bioinformatics/btp619

Systems biology

Simulation-based model selection for dynamical systems in systems and population biology

Tina Toni[1,2,*] and Michael P. H. Stumpf[1,2,*]

[1]Division of Molecular Biosciences, Imperial College London, Wolfson Building, SW7 2AZ and [2]Institute of Mathematical Sciences, Imperial College London, 53 Prince's Gate, London SW7 2PG, UK

Received on September 28, 2009; revised on October 23, 2009; accepted on October 25, 2009

Advance Access publication October 29, 2009

Associate Editor: Jonathan Wren

ABSTRACT

Motivation: Computer simulations have become an important tool across the biomedical sciences and beyond. For many important problems several different models or hypotheses exist and choosing which one best describes reality or observed data is not straightforward. We therefore require suitable statistical tools that allow us to choose rationally between different mechanistic models of, e.g. signal transduction or gene regulation networks. This is particularly challenging in systems biology where only a small number of molecular species can be assayed at any given time and all measurements are subject to measurement uncertainty.

Results: Here, we develop such a model selection framework based on approximate Bayesian computation and employing sequential Monte Carlo sampling. We show that our approach can be applied across a wide range of biological scenarios, and we illustrate its use on real data describing influenza dynamics and the JAK-STAT signalling pathway. Bayesian model selection strikes a balance between the complexity of the simulation models and their ability to describe observed data. The present approach enables us to employ the whole formal apparatus to any system that can be (efficiently) simulated, even when exact likelihoods are computationally intractable.

Contact: ttoni@imperial.ac.uk; m.stumpf@imperial.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Mathematical models are widely used to describe and analyse complex systems and processes. Formulating a model to describe, e.g. a signalling pathway or host parasite system, requires us to condense our assumptions and knowledge into a single coherent framework (May, 2004). Mathematical analysis and computer simulations of such models then allow us to compare model predictions with experimental observations in order to test, and ultimately improve these models. The continuing success, e.g. of systems biology, relies on the judicious combination of experimental and theoretical lines of argument.

Because many of the mathematical models in biology (as in many other disciplines) are too complicated to be analysed in a closed form, computer simulations have become the primary tool in the quantitative analysis of very large or complex biological systems. This, however, can complicate comparisons of different candidate models in light of (frequently sparse and noisy) observed data. Whenever probabilistic models exist, we can employ standard model selection approaches of either a frequentist, Bayesian or information theoretic nature (Burnham and Anderson, 2002; Vyshemirsky and Girolami, 2008). But if suitable probability models do not exist, or if the evaluation of the likelihood is computationally intractable, then we have to base our assessment on the level of agreement between simulated and observed data. This is particularly challenging when the parameters of simulation models are not known but must be inferred from observed data as well. Bayesian model selection side-steps or overcomes this problem by marginalizing (i.e. integrating) over model parameters, thereby effectively treating all model parameters as nuisance parameters.

For the case of parameter estimation when likelihoods are intractable, approximate Bayesian computation (ABC) frameworks have been applied successfully (Beaumont *et al.*, 2002; Marjoram *et al.*, 2003; Ratmann *et al.*, 2007, 2009; Sisson *et al.*, 2007; Toni *et al.*, 2009). In ABC, the calculation of the likelihood is replaced by a comparison between the observed data and simulated data. Given the prior distribution $P(\theta)$ of parameter θ, the goal is to approximate the posterior distribution, $P(\theta|D_0) \propto f(D_0|\theta)P(\theta)$, where $f(D_0|\theta)$ is the likelihood of θ given the data D_0. ABC methods have the following generic form:

(1) Sample a candidate parameter vector θ^* from prior distribution $P(\theta)$.

(2) Simulate a dataset D^* from the model described by a conditional probability distribution $f(D|\theta^*)$.

(3) Compare the simulated dataset, D^*, to the experimental data, D_0, using a distance function, d, and tolerance ϵ; if $d(D_0,D^*) \leq \epsilon$, accept θ^*. The tolerance $\epsilon \geq 0$ is the desired level of agreement between D_0 and D^*.

The output of an ABC algorithm is a sample of parameters from the distribution $P(\theta|d(D_0,D^*) \leq \epsilon)$. If ϵ is sufficiently small then this distribution will be a good approximation for the 'true' posterior distribution, $P(\theta|D_0)$. A tutorial on ABC methods is available in the Supplementary Material.

Such a parameter estimation approach can be used whenever the model is known. However, when several plausible candidate models are available we have a model selection problem, where

*To whom correspondence should be addressed.

both the model structure and parameters are unknown. In the Bayesian framework, model selection is closely related to parameter estimation, but the focus shifts onto the marginal posterior probability of model m given data D_0,

$$P(m|D_0)=\frac{P(D_0|m)P(m)}{P(D_0)}$$

where $P(D_0|m)$ is the marginal likelihood and $P(m)$ the prior probability of the model (Gelman *et al.*, 2003). This framework has some conceptual advantages over classical hypothesis testing: for example, we can rank an arbitrary number of different non-nested models by their marginal probabilities; and rather than only considering evidence against a model the Bayesian framework also weights evidence in a model's favour (Jeffreys, 1939). In practical applications, however, a range of potential pitfalls need considering: model probabilities can show strong dependence on model and parameter priors; and the computational effort needed to evaluate these posterior distributions can make these approaches cumbersome.

The computationally expensive step in Bayesian model selection is the evaluation of the marginal likelihood, which is obtained by marginalizing over model parameters; i.e. $P(D_0|m)=\int f(D_0|m,\theta)P(\theta|m)\mathrm{d}\theta$, where $P(\theta|m)$ is the parameter prior for model m. Here, we develop a computationally efficient ABC model selection formalism based on a sequential Monte Carlo (SMC) sampler. We show that our ABC SMC procedure allows us to employ the whole paraphernalia of the Bayesian model selection formalism, and illustrate the use and scope of our new approach in a range of models: chemical reaction dynamics, Gibbs random fields and real data describing influenza spread and JAK-STAT signal transduction.

2 ABC FOR MODEL SELECTION

Our goal is to estimate the marginal posterior distribution of a model, $P(m|D_0)$, and in this section we explain two ways in which this problem can be approached. In the *joint space-based approach*, we define a joint space of model indicators, $m=1,2,...,|\mathcal{M}|$, and corresponding model parameters, θ, obtain the joint posterior distribution over the combined space of models and parameters, $P(\theta,m|D_0)$, and finally marginalize over parameters to obtain $P(m|D_0)$. In the second, *marginal likelihood-based approach*, we estimate marginal likelihoods (also called the *evidence*), $P(D_0|m)$, for each given model, and use these to calculate the marginal posterior model distributions through

$$P(m|D_0)=\frac{P(D_0|m)P(m)}{\sum_{m'}P(D_0|m')P(m')}.$$

Both approaches have been applied under the ABC rejection scheme, which is computationally prohibitive for models with even an only moderate number of parameters (Grelaud *et al.*, 2009; Wilkinson, 2007). Here, we incorporate ideas from SMC to both of the above approaches, making them computationally more efficient. In this section, we present only the more powerful approach *ABC SMC model selection on the joint space*. We refer the reader to the Supplementary Material for derivations and details, as well as discussion on the ABC SMC model selection algorithm based on the marginal likelihood approach.

In model selection based on ABC rejection, we adapt the basic ABC procedure (presented in Section 1) to the joint space, where *particles* (m,θ) consist of a model indicator m and a parameter θ. The ABC rejection model selection algorithm on the joint space proceeds as follows (Grelaud *et al.*, 2009):

(1) Draw m^* from the prior $P(m)$.

(2) Sample θ^* from the prior $P(\theta|m^*)$.

(3) Simulate a candidate dataset $D^*\sim f(D|\theta^*,m^*)$.

(4) Compute the distance. If $d(D_0,D^*)\leq\epsilon$, accept (m^*,θ^*), otherwise reject it.

(5) Return to 1.

Once a sample of N particles has been accepted, the marginal posterior distribution is approximated by

$$P(m=m'|D_0)\approx\frac{\#\text{accepted particles}(m',.)}{N}.$$

In the ABC SMC model selection algorithm on the joint space, particles (parameter vectors) $\{(m_1,\theta_1),...,(m_N,\theta_N)\}$ are sampled from the prior distribution, $P(m,\theta)$, and propagated through a sequence of intermediate distributions, $P(m,\theta|d(D_0,D^*)\leq\epsilon_i)$, $i=1,...,T-1$, until they represent a sample from the target distribution, $P(m,\theta|d(D_0,D^*)\leq\epsilon_T)$. The tolerances ϵ_i are chosen such that $\epsilon_1>\cdots>\epsilon_T\geq 0$, and the distributions thus gradually evolve towards the target posterior distribution.

The algorithm is presented below (and explained in the Supplementary Material).

2.1 ABC SMC model selection algorithm on the joint space

MS1 Initialize $\epsilon_1,...,\epsilon_T$.
Set the population indicator $t=1$.

MS2.0 Set the particle indicator $i=1$.

MS2.1 If $t=1$, sample (m^{**},θ^{**}) from the prior distribution $P(m,\theta)$.
If $t>1$, sample m^* with probability $P_{t-1}(m^*)$ and draw $m^{**}\sim KM_t(m|m^*)$.
Sample θ^* from previous population $\{\theta(m^{**})_{t-1}\}$ with weights w_{t-1} and draw $\theta^{**}\sim KP_{t,m^{**}}(\theta|\theta^*)$.
If $P(m^{**},\theta^{**})=0$, return to **MS2.1**.
Simulate a candidate dataset $D^*\sim f(D|\theta^{**},m^{**})$.
If $d(D_0,D^*)>\epsilon_t$, return to **MS2.1**.

MS2.2 Set $(m_t^{(i)},\theta_t^{(i)})=(m^{**},\theta^{**})$ and calculate the weight of the particle as

$$w_t^{(i)}(m_t^{(i)},\theta_t^{(i)})=\begin{cases} b_t(m_t^{(i)},\theta_t^{(i)}), & \text{if } t=1\\ \dfrac{P(m_t^{(i)},\theta_t^{(i)})b_t(m_t^{(i)},\theta_t^{(i)})}{S}, & \text{if } t>1.\end{cases}$$

where

$$b_t(m_t^{(i)},\theta_t^{(i)})\ =\ \frac{1}{B_t}\sum_{b=1}^{B_t}\mathbb{1}(d(D_0,D_b^*)\leq\epsilon_t)$$

$$S=\sum_{j=1}^{|\mathcal{M}|}P_{t-1}(m_{t-1}^{(j)})KM_t(m_t^{(i)}|m_{t-1}^{(j)})\times$$

$$\sum_{k;m_{t-1}=m_t^{(i)}}\frac{w_{t-1}^{(k)}KP_{t,m_t^{(i)}}(\theta_t^{(i)}|\theta_{t-1}^{(k)})}{P_{t-1}(m_{t-1}=m_t^{(i)})}$$

If $i<N$ set $i=i+1$, go to **MS2.1**.

MS3 Normalize the weights w_t.
Sum the particle weights to obtain marginal model probabilities,

$$P_t(m_t=m)=\sum_{i;m_t^{(i)}=m}w_t^{(i)}(m_t^{(i)},\theta_t^{(i)}).$$

If $t<T$, set $t=t+1$, go to **MS2.0**.

Particles sampled from a previous distribution are denoted by a single asterisk, and after perturbation by a double asterisk. *KM* is a model perturbation kernel which allows us to obtain model m from model m^* and *KP* is the parameter perturbation kernel. $B_t\geq 1$ is the number of replicate simulation run for a fixed particle (for deterministic models $B_t=1$) and $|\mathcal{M}|$ denotes the number of candidate models.

The output of the algorithm, i.e. the set of particles $\{(m_T,\theta_T)\}$ associated with weights w_T, is the approximation of the full posterior distribution on the joint model and parameter space. The approximation of the marginal posterior distribution of the model obtained by marginalization is

$$P_T(m_T=m)=\sum_{i;m_T^{(i)}=m}w_t^{(i)}(m_T^{(i)},\theta_T^{(i)}),$$

and we can also straightforwardly obtain the marginalized parameter distributions.

The algorithm requires the user to define the prior distribution, distance function, tolerance schedule and perturbation kernels. In all the examples presented in the results section, we choose uniform prior distributions for all parameters and models; that is, all models are a priori equally plausible. Such priors are informative in a sense that they define a feasible parameter region (e.g. reaction rates are positive), but they are predominantly non-informative as they do not specify any further preference for particular parameter values. This way the inference will mostly be informed by the information contained in the data. A good tolerance can be found empirically by trying to reach the lowest distance feasible and arrive at the posterior distribution in a computationally efficient way. Our perturbation kernels are component-wise truncated uniform or Gaussian and are automatically adapted by feeding back information on the obtained parameter ranges from the previous population. Distance functions are defined for each model as specified in Section 3. The algorithm presented in Toni *et al.* (2009) is a special case of the above algorithm for discrete uniform *KM* kernel and uniform prior distribution of the model $P(m)$.

3 RESULTS

In this section, we illustrate ABC SMC for model selection on a simple example of stochastic reaction kinetics. We then compare the computational efficiency of ABC SMC for stochastic models of Gibbs random fields with that of the ABC rejection model selection method. Finally, we apply the algorithm to several real datasets: first, we select between different stochastic models of influenza epidemics (where we can compare our approach with previously published results obtained using exact Bayesian model selection), and then apply our approach to choose from among different mechanistic models for the STAT5 signalling pathway.

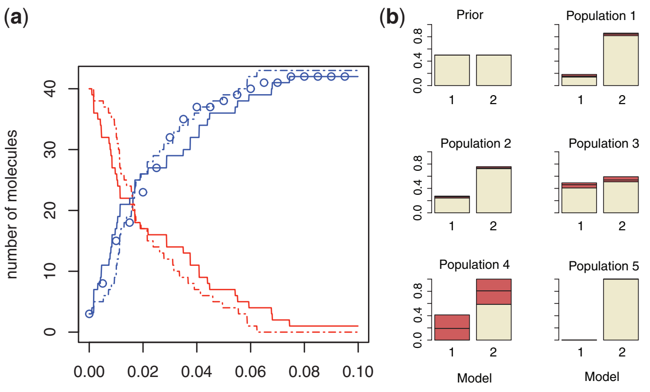

Fig. 1. (**a**) Stochastic trajectories of species X (red) and Y (blue). Model 1 is simulated for $k_1=2.1$ (dashed line), model 2 for $k_2=30$ (full line). Data points are represented by circles. (**b**) We have repeated the model selection run 20 times; the red sections present 25% and 75% quantiles around the median. Prior distribution $P(m)$ is chosen uniform and k_1, $k_2\sim U(0,100)$. Perturbation kernels are chosen as follows: $KP_t(k|k^*)=U(-\sigma,\sigma)$, $\sigma=2(\max\{k\}_{t-1}-\min\{k\}_{t-1})$ and $KM_t(m|m^*)=0.7$ if $m=m^*$ and 0.3 otherwise. Number of particles $N=1000$. $B_t=1$. Distance function is mean squared error and tolerance schedule $\epsilon=\{3000,1400,600,140,40\}$.

3.1 Chemical reaction kinetics

We illustrate our algorithm for the stochastic reaction kinetic models $X+Y\xrightarrow{k_1}2Y$ and $X\xrightarrow{k_2}Y$. The first is a model of an autocatalytic reaction, where the reaction product Y is the catalyst for the reaction. In the second, molecules Y do not need to be present for a change from X to Y to occur. Such models have, for example, been considered in the context of prion replication dynamics (Eigen, 1996; Prusiner, 1982), where X represents a healthy form of a prion protein and Y a diseased form.

We simulate synthetic datasets of Y measured at 20 time points using Gillespie algorithm (Gillespie, 1977) from model 2 with parameter $k_2=30$ and initial conditions $X_0=40$, $Y_0=3$ (Fig. 1a; Supplementary Table 1). We apply our ABC SMC algorithm for model selection, which identifies the correct model with high confidence (Fig. 1b).

3.2 Gibbs random fields

Gibbs random fields have become staple models in machine learning, including applications in computational biology and bioinformatics [see, e.g. Grelaud *et al.* (2009); Wei and Li (2007)]. Here, we use two Gibbs random field models (Møller, 2003), for which closed form posterior distributions are available. This allows us to compare the ABC SMC approximated posterior distributions of the models with true posterior distributions, and to demonstrate the computational efficiency of our approach when compared with model selection based on ABC rejection sampling.

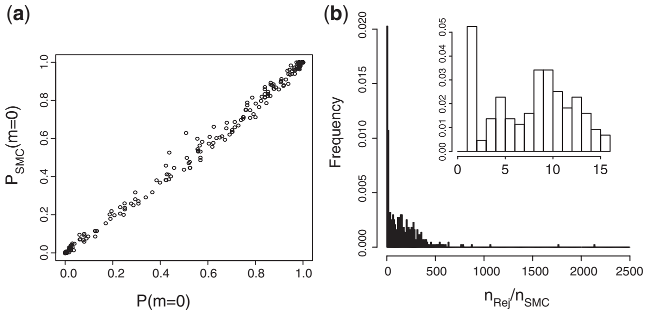

Fig. 2. (**a**) True versus inferred posterior distribution of model m_0. In ABC SMC, we use the Euclidian distance $d(D_0,x)=\sqrt{(S_0(D_0)-S_0(x))^2+(S_1(D_0)-S_1(x))^2}$. $N=500$. $B_t=1$. Tolerance schedule: $\epsilon=\{9,4,3,2,1,0\}$. Perturbation kernels: $KM_t(m|m^*)=0.75$ if $m=m^*$ and 0.25 otherwise; $KP_t(\theta|\theta^*)=U(-\sigma,\sigma)$, $\sigma=0.5(\max\{\theta\}_{t-1}-\min\{\theta\}_{t-1})$. We have excluded those datasets for which all states are in 0 or 1 (for which $P(m=0)\approx 0.3094$ is also correctly inferred) from the analysis. (**b**) Comparison of the number of simulation steps needed by ABC rejection (n_{Rej}) and ABC SMC (n_{SMC}); ABC SMC yields ~50-fold speed-up on average.

Both models, m_0 and m_1, are defined on a sequence of n binary random variables, $x=(x_1,\ldots,x_n)$, $x_i\in\{0,1\}$; m_0 is a collection of n i.i.d. Bernoulli random variables with probability $\theta_0/(1+\exp(\theta_0))$; m_1 is equivalent to a standard Ising model, i.e. x_1 is taken to be a binary random variable and $P(x_{i+1}=x_i|x_i)=\theta_1/(1+\exp(\theta_1))$ for $i=2,\ldots,x_n$. The likelihood functions are

$$f_0(x|\theta_0)=\frac{e^{\theta_0 S_0(x)}}{(1+e^{\theta_0})^n} \quad \text{and} \quad f_1(x|\theta_1)=\frac{e^{\theta_1 S_1(x)}}{2(1+e^{\theta_1})^{n-1}},$$

where $S_0(x)=\sum_{i=1}^{n}\mathbb{1}(x_i=1)$ and $S_1(x)=\sum_{i=2}^{n}\mathbb{1}(x_i=x_{i-1})$ are sufficient statistics, respectively.

We simulate 1000 datasets from both models for different values of parameters $\theta_0\sim U(-5,5)$, $\theta_1\sim U(0,6)$ and $n=100$. Using ABC SMC for model selection allows us to estimate posterior model distributions correctly and demonstrate a considerable computational speed-up in ABC SMC compared with ABC rejection (Fig. 2).

3.3 Infuenza infection outbreaks

We next apply ABC SMC for model selection to models of the spread of different strains of the influenza virus. We use data from influenza A (H3N2) outbreaks that occurred in 1977–1978 and 1980–1981 in Tecomseh, Michigan (Addy *et al.*, 1991, Supplementary Table 2), and a second dataset of an influenza B infection outbreak in 1975–1976 and influenza A (H1N1) infection outbreak in 1978–1979 in Seattle, Washington (Longini and Koopman, 1982, Supplementary Table 3). The basic questions to be addressed here are whether (i) different outbreaks of the same strain and (ii) outbreaks of different molecular strains of the influenza virus can be described by the same model of disease spread.

We assume that virus can spread from infected to susceptible individuals and distinguish between spread inside households or across the population at large (Longini and Koopman, 1982). Let q_c denote the probability that a susceptible individual does not get infected from the community and q_h the probability that a susceptible individual escapes infection within their household. Then w_{js}, the probability that j out of the s susceptibles in a household become infected, is given by

$$w_{js}=\binom{s}{j}w_{jj}(q_c q_h^j)^{s-j}, \tag{1}$$

where $w_{0s}=q_c^s$, $s=0,1,2,\ldots$ and $w_{jj}=1-\sum_{i=0}^{j-1}w_{ij}$. We are interested in inferring the pair of parameters q_h and q_c of the model (1) using the data from Supplementary Table 2. These data were obtained from two separate outbreaks of the same strain, H3N2, and the question of interest is whether these are characterized by the same epidemiological parameters [this question was previously considered in Clancy and O'Neill (2007) and O'Neill *et al.* (2000)]. To investigate this issue, we consider two models: one with four parameters, q_{h1}, q_{c1}, q_{h2}, q_{c2}, which describes the hypothesis that each outbreak has its own characteristics; the second models the hypothesis that both outbreaks share the same epidemiological parameter values for q_h and q_c. Prior distributions of all parameters are chosen to be uniform over the range $[0,1]$.

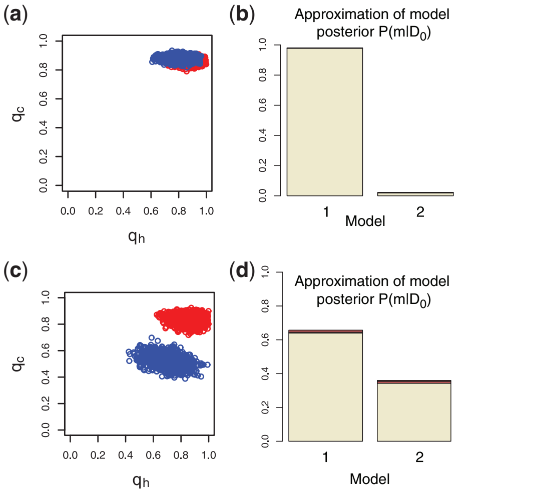

Fig. 3. (**a**) ABC SMC posterior distributions for parameters inferred for a four-parameter model from the data in Supplementary Table 2. Marginal posterior distributions of parameters q_{c1}, q_{h1} (red) and q_{c2}, q_{h2} (blue). (**b**) Approximation of a posterior marginal distribution $P(m|D_0)$. Model 1 is a two-parameter and model 2 a four-parameter model (1). All intermediate populations are shown in Supplementary Figure 1a. (**c**) The same as (a) but here the data used is from Supplementary Table 3. (**d**) Estimation of a posterior marginal distribution. Model 1 is a two-parameter and model 2 a three-parameter model (1). All intermediate populations are shown in Supplementary Figure 1b.

To apply ABC SMC, we use a distance function

$$d(D_0,D^*)=\frac{1}{2}(||D_1-D^*(q_{h1},q_{c1})||_F+||D_2-D^*(q_{h2},q_{c2})||_F),$$

where $||\ ||_F$ denotes the Frobenius norm, $D_0=D_1\cup D_2$ with D_1 the 1977–1978 outbreak and D_2 the 1980–1981 outbreak datasets from Supplementary Table 2, and D^* is the simulation output from model (1). The results we obtain are summarized in Figure 3a and b and strongly suggest that the two outbreaks appear to have shared the same epidemiological characteristics. Figure 3a shows the posterior distribution of the four-parameter model. The marginal posterior distributions of q_{h1} and q_{c1} are largely overlapping with the marginal posterior distributions of q_{h2} and q_{c2} and we therefore, unsurprisingly, get strong evidence in favour of the two-parameter model. Figure 3b shows the marginal posterior distribution of the model; the posterior probability of model 1 is 0.98 (median over 10 runs), which gives unambiguous support to model 1, meaning that outbreaks of the same strain share the same dynamics.

Outbreaks due to a different viral strain (Supplementary Table 3) have different characteristics as indicated by the posterior distribution of the four-parameter model presented in Figure 3c. This was confirmed by applying our model selection algorithm; the inferred posterior marginal model probability of a two-parameter model was negligible (data not shown). From Figure 3c, we also see that these differences are due to differences in viral spread across the community whereas within-household dynamics are comparable. We thus explore a further model with three parameters, q_{c1}, q_{c2}, q_h (model 1), where the two outbreaks share the same within-household characteristics (q_h), and compare it against and the four-parameter model (model 2). The obtained Bayes factor suggests that there is only very week evidence in favour of model 1 (Fig. 3d), which is in agreement with the result of Clancy and O'Neill (2007).

In general genetic predisposition, differences in immunity and lifestyle, etc., will lead to heterogeneity in susceptibility to viral infection among the host population. Such a model can be written as (O'Neill *et al.*, 2000)

$$w_{js}(v) = \sum_{i=0}^{s-j} \binom{s}{i} v^i (1-v)^{s-i} w_{j,s-i}. \qquad (2)$$

On the basis of the previous results, we combine both outbreak datasets from Supplementary Table 2, and find some evidence that model (2) explains the data better than model (1), suggesting that the host-virus dynamics are shaped by the molecular nature of the viral strain, as well as by variability in the host population (Supplementary Fig. 2).

3.4 JAK-STAT signalling pathway

Having convinced ourselves that the novel ABC SMC model selection approach agrees with the analytical model probabilities and those obtained using conventional Bayesian model selection, while outperforming conventional ABC rejection model selection approaches, we can now turn our attention to real world scenarios that have not previously been considered from a Bayesian (exact or approximate) perspective. Here, we consider models of signalling though the erythropoietin receptor (EpoR), transduced by STAT5 (Fig. 4a) (Darnell, 1997; Horvath, 2000). Signalling through this receptor is crucial for proliferation, differentiation and survival of erythroid progenitor cells (Klingmüller *et al.*, 1996). When the Epo hormone binds to the EpoR receptor, the receptor's cytoplasmic domain is phosporylated, which creates a docking site for signalling molecules, in particular STAT5. Upon binding to the activated receptor, STAT5 first becomes phosphorylated, then dimerizes and translocates to the nucleus, where it acts as a transcription factor. There have been competing hypotheses about what happens with the STAT5 in the nucleus. Originally, it had been suggested that STAT5 gets degraded in the nucleus in an ubiquitin-associated way (Kim and Maniatis, 1996), but other evidence suggests that they are dephosphorylated in the nucleus and then trafficked back to the cytoplasm (Köster and Hauser, 1999).

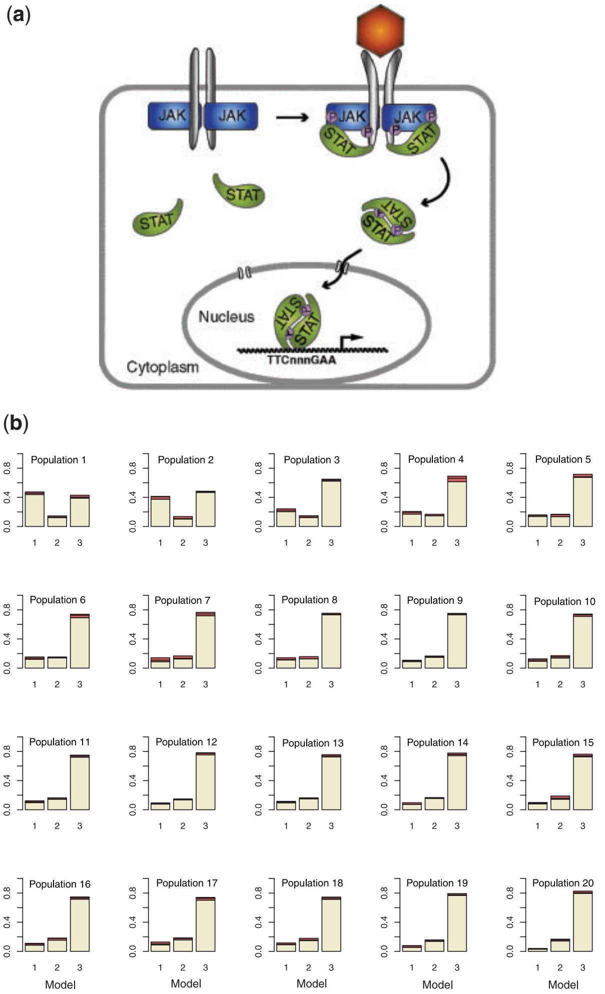

Fig. 4. (**a**) STAT5 signalling pathway. Adapted from (Arbouzova and Zeidler, 2006). (**b**) Histograms show populations of the model parameter m. Population 20 represents the approximation of the marginal posterior distribution of m. Tolerance schedule: $\epsilon = \{200, 100, 50, 35, 30, 25, 22, 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, 10, 9, 8\}$. Perturbation kernels: $KM_t(m|m^*) = 0.6$ if $m = m^*$ and 0.2 otherwise; $KP_t(\theta|\theta^*) = U(-\sigma, \sigma)$, $\sigma = 0.5(\max\{\theta\}_{t-1} - \min\{\theta\}_{t-1})$. $N = 500$. Distance function: $d(D_0, D^*) = \sqrt{\sum_t \left(\frac{y_0^{(1)}(t) - y^{*(1)}(t)}{\sigma_{D_0}^{(1)}(t)}\right)^2 + \left(\frac{y^{(2)}(t) - y^{*(2)}(t)}{\sigma_{D_0}^{(2)}(t)}\right)^2}$, with $D_0 = \{y_0^{(1)}, y_0^{(2)}\}$, $D^* = \{y^{*(1)}, y^{*(2)}\}$ and $y^{(1)}$ the total amount of phosphoryalated STAT5 in the cytoplasm and $y^{(2)}$ the total amount of STAT5 in the cytoplasm. $\sigma_{D_0}^{(1)}$ and $\sigma_{D_0}^{(2)}$ are the associated confidence intervals; reassuringly, other distance functions, e.g. the square root of the sum of squared errors yield identical model selection results (data not shown).

The ambiguity of the shutoff mechanism of STAT5 in the nucleus triggered the development of several mathematical models (Müller *et al.*, 2004; Swameye *et al.*, 2003; Timmer and Müller, 2004) describing different hypotheses. All models assume mass action kinetics and denote the amount of activated Epo-receptors by

$EpoR_A$, monomeric unphosphorylated and phosphorylated STAT5 molecules by x_1 and x_2, respectively, dimeric phosphorylated STAT5 in the cytoplasm by x_3 and dimeric phosphorylated STAT5 in the nucleus by x_4. The most basic model developed by Timmer *et al.*, under the assumption that phosphorylated STAT5 does not leave the nucleus, consists of the following kinetic equations,

$$\dot{x}_1 = -k_1 x_1 EpoR_A \tag{3}$$

$$\dot{x}_2 = -k_2 x_2^2 + k_1 x_1 EpoR_A$$

$$\dot{x}_3 = -k_3 x_3 + \frac{1}{2} k_2 x_2^2$$

$$\dot{x}_4 = k_3 x_3. \tag{4}$$

One can then assume that phosphorylated STAT5 dimers dissociate and leave the nucleus; this is modelled by adding appropriate kinetic terms to the Equations (3) and (4) of the basic model to obtain

$$\dot{x}_1 = -k_1 x_1 EpoR_A + 2k_4 x_4$$

$$\dot{x}_4 = k_3 x_3 - k_4 x_4.$$

The cycling model can be developed further by assuming a delay before STAT5 leaves the nucleus:

$$\dot{x}_1 = -k_1 x_1 EpoR_A + 2k_4 x_3(t-\tau)$$

$$\dot{x}_4 = k_3 x_3 - k_4 x_3(t-\tau). \tag{5}$$

This model was chosen as the best model in the original analyses (Müller *et al.*, 2004; Swameye *et al.*, 2003) based on a numerical evaluation of the likelihood, followed by a likelihood ratio test and bootstrap procedure for model selection. The data are partially observed time course measurements of the total amount of STAT5 in the cytoplasm, and the amount of phosphorylated STAT5 in the cytoplasm; both are only known up to a normalizing factor.

We propose a further model with clear physical interpretation where the delay acts on STAT5 inside the nucleus (x_4) rather than on x_3 [in Equation (5)], for which a biological interpretation is difficult. Instead of $x_3(t-\tau)$, we propose to model the delay of phosphorylated STAT5 x_4 in the nucleus directly and obtain (Zi and Klipp, 2006):

$$\dot{x}_1 = -k_1 x_1 EpoR_A + 2k_4 x_4(t-\tau)$$

$$\dot{x}_4 = k_3 x_3 - k_4 x_4(t-\tau).$$

We perform the ABC SMC model selection algorithm on the following non-nested models: (i) cycling delay model with $x_3(t-\tau)$, (ii) cycling delay model with $x_4(t-\tau)$ and (iii) cycling model without a delay. The model parameter m can therefore take values 1–3.

For each proposed model and parameter combination we numerically solve the ordinary differential equations and add $\epsilon \sim N(0,\sigma)$ to obtain the simulated time course data. The noise parameter σ can be either fixed or treated as another parameter to be estimated; we consider the latter option, under the assumption that the experimental noise is independent and identically distributed for all time points.

Figure 4b shows intermediate populations leading to the ABC SMC marginal posterior distribution over the model parameters m (population 20). Bayes factors can be calculated from the last population and according to the conventional interpretation of Bayes factors (Kass and Raftery, 1995), it can be concluded that there is strong evidence in favour of model 3 compared to model 1, positive evidence in favour of model 3 compared to model 2, and positive evidence in favour of model 2 compared to model 1. Thus, cycling appears to be clearly important and the model that receives the most support is the cycling model without a time delay. Here, the flexibility of ABC SMC has allowed us to perform simultaneous model selection on non-nested models of ordinary and time-delay differential equations.

4 DISCUSSION

We have developed a novel model selection methodology based on ABC and SMC. The results obtained here illustrate the usefulness and wide applicability of our ABC SMC method, even when experimental data are scarce, when no measurements are available for some species, when temporal data are not measured at equidistant time points and when parameters such as kinetic rates are unknown. In the context of dynamical systems, our method can be applied across all simulation and modelling (including qualitative modelling) frameworks; for JAK-STAT signal transduction dynamics, for example, we have been able to compare the relative explanatory power of ordinary and time-delay differential equation models. Our model selection procedure is also not confined to dynamical systems; in fact the scope for application is immense and limited only by the availability of efficient simulation approaches.

Routine application to complex models in systems, computational and population biology with hundreds or thousands of parameters (Chen *et al.*, 2009), will require further numerical developments due to the high-computational cost of repeated simulations. SMC-based ABC methods are, however, highly parallelizable and we believe that future work should exploit this property to make these methods computationally more efficient. Further potential improvements might come from (i) regression adjustment techniques that have so far been applied in the parameter estimation ABC framework (Beaumont *et al.*, 2002; Blum and François, 2009; Excoffier, 2009); (ii) from automatic generation of the tolerance schedules (Del Moral *et al.*, 2009); and (iii) by developing more sophisticated perturbation kernels that exploit inherent properties of biological dynamical systems such as sloppiness (Gutenkunst *et al.*, 2007; Secrier *et al.*, 2009); here especially we feel that there is substantial room for improvement as the likelihoods of dynamical systems contain information about the qualitative behaviour (Kirk *et al.*, 2008) which can also be exploited in ABC frameworks.

5 CONCLUSIONS

We conclude by emphasizing the need for inferential methods which can assess the relative performance and reliability of different models. The need for such reliable model selection procedures can hardly be overstated: with an increasing number of biomedical problems being studied using simulation approaches, there is an obvious and urgent need for statistically sound approaches that allow us to differentiate between different models. If parameters are known or the likelihood is available in a closed form, then the model selection is generally straightforward. However, for many of the most interesting systems biology (and generally, scientific) problems this is not the case and here ABC SMC can be employed.

ACKNOWLEDGEMENTS

We are especially grateful to Paul Kirk for his insightful comments and many valuable discussions. We furthermore thank the members of the *Theoretical Systems Biology group* for discussions and comments on earlier versions of this article.

Funding: Medical Research Council priority studentship (to T.T.) and Biotechnology and Biological Sciences Research Council (grant BB/G009374/1).

Conflict of Interest: none declared.

REFERENCES

Addy,C. Jr, *et al.* (1991) A generalized stochastic model for the analysis of infectious disease final size data. *Biometrics*, **47**, 961–974.

Arbouzova,N.I. and Zeidler,M.P. (2006) JAK/STAT signalling in drosophila: insights into conserved regulatory and cellular functions. *Development*, **133**, 2605–2616.

Beaumont,M.A. *et al.* (2002) Approximate Bayesian computation in population genetics. *Genetics*, **162**, 2025–2035.

Blum,M.G. and François,O. (2009) Non-linear regression models for approximate Bayesian computation. *Stat. Comput.*, [Epub ahead of print, doi:10.1007/s11222-009-9116-0].

Burnham,K. and Anderson,D. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer, New York.

Chen,W.W. *et al.* (2009) Input-output behavior of ErbB signaling pathways as revealed by a mass action model trained against dynamic data. *Mol. Syst. Biol.*, **5**, 239.

Clancy,D. and O'Neill,P.D. (2007) Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households. *Scand. J. Stat.*, **34**, 259–274.

Darnell,J.E. (1997) STATs and gene regulation. *Science*, **277**, 1630–1635.

Del Moral,P. *et al.* (2009) An adaptive sequential Monte Carlo method for approximate Bayesian computation. Imperial College Technical Report.

Eigen,M. (1996) Prionics or the kinetic basis of prion diseases. *Biophys. Chem.*, **63**, A1–A18.

Excoffier,C.L.D.W.L. (2009) Bayesian computation and model selection in population genetics. *arXiv:0901.2231v1 [stat.ME]*.

Gelman,A. *et al.* (2003) *Bayesian Data Analysis*, 2nd edn. Chapman & Hall/CRC, London.

Gillespie,D. (1977) Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, **81**, 2340–2361.

Grelaud,A. *et al.* (2009) ABC likelihood-free methods for model choice in Gibbs random fields. *Bayesian Anal.*, **4**, 317–336.

Gutenkunst,R. *et al.* (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.*, **3**, e189.

Horvath,C.M. (2000) STAT proteins and transcriptional responses to extracellular signals. *Trends Biochem. Sci.*, **25**, 496–502.

Jeffreys,H. (1939) *Theory of Probability*, 1st edn. The Clarendon Press, Oxford.

Kass,R. and Raftery,A. (1995) Bayes factors. *J. Am. Stat. Assoc.*, **90**, 773–795.

Kim,T.K. and Maniatis,T. (1996) Regulation of interferon-gamma-activated STAT1 by the ubiquitin-proteasome pathway. *Science*, **273**, 1717–1719.

Kirk,P.D.W. *et al.* (2008) Parameter inference for biochemical systems that undergo a Hopf bifurcation. *Biophys. J.*, **95**, 540–549.

Klingmüller,U. *et al.* (1996) Multiple tyrosine residues in the cytosolic domain of the erythropoietin receptor promote activation of STAT5. *Proc. Natl Acad. Sci. USA*, **93**, 8324–8328.

Köster,M. and Hauser,H. (1999) Dynamic redistribution of STAT1 protein in IFN signaling visualized by GFP fusion proteins. *Eur. J. Biochem.*, **260**, 137–144.

Longini,I.L. Jr. and Koopman, J. (1982) Household and community transmission parameters from final distributions of infections in households. *Biometrics*, **38**, 115–126.

Marjoram,P. *et al.* (2003) Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA*, **100**, 15324–15328.

May,R.M. (2004) Uses and abuses of mathematics in biology. *Science*, **303**, 790–793.

Møller,J. (2003) *Spatial Statistics and Computational Methods*. Springer, New York.

Müller,T.G. *et al.* (2004) Tests for cycling in a signalling pathway. *J. R. Stat. Soc. Ser. C*, **53**, 557.

O'Neill,P. *et al.* (2000) Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *Appl. Stat.*, **49**, 517–542.

Prusiner,S.B. (1982) Novel proteinaceous infectious particles cause scrapie. *Science*, **216**, 136–144.

Ratmann,O. *et al.* (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H. pylori and P. falciparum. *PLoS Comput. Biol.*, **3**, 2266–2278.

Ratmann,O. *et al.* (2009) Model criticism based on likelihood-free inference, with an application to protein network evolution. *Proc. Natl Acad. Sci. USA*, **106**, 10576–10581.

Secrier,M. *et al.* (2009) The ABC of reverse engineering biological signalling systems. *Mol. BioSyst.* **5**, 1925–1935.

Sisson,S.A. *et al.* (2007) Sequential Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA*, **104**, 1760–1765.

Swameye,I. *et al.* (2003) Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proc. Natl Acad. Sci. USA*, **100**, 1028–1033.

Timmer,J. and Müller,T. (2004) Modeling the nonlinear dynamics of cellular signal transduction. *Int. J. Bifurcat. Chaos*, **14**, 2069–2079.

Toni,T. *et al.* (2009) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, **6**, 187–202.

Vyshemirsky,V. and Girolami,M.A. (2008) Bayesian ranking of biochemical system models. *Bioinformatics*, **24**, 833–839.

Wei,Z. and Li,H. (2007) A Markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**, 1537–1544.

Wilkinson,R.D. (2007) Bayesian inference of primate divergence times. PhD Thesis, University of Cambridge.

Zi,Z. and Klipp,E. (2006) SBML-PET: a Systems Biology Markup Language-based parameter estimation tool. *Bioinformatics*, **22**, 2704–2705.