Model criticism based on likelihood-free inference, with an application to protein network evolution

Oliver Ratmann[a,1], Christophe Andrieu[b], Carsten Wiuf[c], and Sylvia Richardson[d]

[a]Department of Public Health and Epidemiology, Imperial College London, London W2 1PG, United Kingdom; [b]Department of Mathematics, University of Bristol, Bristol BS8 1TW, United Kingdom; [c]Bioinformatics Research Center, University of Aarhus, 8000 Aarhus C, Denmark; and [d]Centre for Biostatistics, Imperial College London, London W1 1PG, United Kingdom

Edited by Elizabeth A. Thompson, University of Washington, Seattle, WA, and approved March 26, 2009 (received for review August 13, 2008)

Mathematical models are an important tool to explain and comprehend complex phenomena, and unparalleled computational advances enable us to easily explore them without any or little understanding of their global properties. In fact, the likelihood of the data under complex stochastic models is often analytically or numerically intractable in many areas of sciences. This makes it even more important to simultaneously investigate the adequacy of these models—in absolute terms, against the data, rather than relative to the performance of other models—but no such procedure has been formally discussed when the likelihood is intractable. We provide a statistical interpretation to current developments in likelihood-free Bayesian inference that explicitly accounts for discrepancies between the model and the data, termed Approximate Bayesian Computation under model uncertainty (ABCμ). We augment the likelihood of the data with unknown error terms that correspond to freely chosen checking functions, and provide Monte Carlo strategies for sampling from the associated joint posterior distribution without the need of evaluating the likelihood. We discuss the benefit of incorporating model diagnostics within an ABC framework, and demonstrate how this method diagnoses model mismatch and guides model refinement by contrasting three qualitative models of protein network evolution to the protein interaction datasets of *Helicobacter pylori* and *Treponema pallidum*. Our results make a number of model deficiencies explicit, and suggest that the *T. pallidum* network topology is inconsistent with evolution dominated by link turnover or lateral gene transfer alone.

Bayesian inference | intractable likelihoods | Markov chain Monte Carlo | Approximate Bayesian Computation | model uncertainty

In the quest to comprehend complex observations, hypotheses about underlying mechanisms are formalized in terms of precise mathematical models (1). Much of statistical reasoning then proceeds in an iterative process between data acquisition, data analysis, and model development (2). At the ith iteration, the interpretation of observed data x_0 in terms of some target parameters θ conditional on a tentative, probabilistic model M_i has a long tradition in Bayesian inference (3). The focus is typically on the posterior density $f(\theta \mid x_0, M_i)$, which is related to the likelihood $f(x_0 \mid \theta, M_i)$ and the prior $\pi_\theta(\theta \mid M_i)$ via Bayes' Theorem:

$$f(\theta \mid x_0, M_i) = f(x_0 \mid \theta, M_i)\pi_\theta(\theta \mid M_i)/f(x_0). \qquad [1]$$

To *explore* whether the current model M_i is consonant with x_0, and to guide further model development, Bayesian predictive diagnostics (4, 5) ask whether x_0 can be viewed as a random observation from a predictive distribution $m(x|M_i)$ in terms of a chosen discrepancy function $\rho(x,x_0)$, $x \sim m$; interpretation of such diagnostics has been the subject of lively debate (6, 7). Application of this machinery for complex models is provided by the workhorses of Bayesian inference (8), such as Markov Chain Monte Carlo (MCMC), as long as the likelihood is readily evaluable up to a normalizing constant.

In many areas of science, such as econometrics (9), molecular genetics (10), epidemiology (11), and evolutionary systems biology (12), *the likelihood is sometimes intractable*. Nevertheless, given a value of θ, it is typically easy to simulate data from $f(\cdot|\theta, M_i)$. Approximate Bayesian Computation (ABC), reviewed in ref. 10, proposes to infer θ by comparing simulated data x to the observed data x_0, in terms of a (real-valued) univariate discrepancy ρ that combines a set of (computationally tractable) summaries $\mathbb{S} = (S_1, \ldots, S_k, \ldots, S_K)$. In its simplest form, values of θ for which the discrepancies are within $\tau \geq 0$ are retained to define the "approximate likelihood"

$$t_\tau(\theta) = \frac{1}{\tau} \int \mathbf{1}\{|\rho(\mathbb{S}(x), \mathbb{S}(x_0))| \leq \tau/2\} f(x|\theta, M_i)dx \qquad [2]$$

in the sense that as $\tau \to 0$, $t_\tau(\theta)$ should approach the likelihood of the summaries, $f(\mathbb{S}(x_0)|\theta, M_i)$; see Fig. 1*A* and the supporting information (SI) *Appendix*, subsection S1.2. ABC may be embedded into Bayesian methods to formally select one model from a specified collection of models, or to average them (13–15); see Fig. 1*B*. However, *relative* comparisons between models do not convey whether models correspond adequately to the observed data and, without exploring the adequacy of models to explain the data, the *meaning* of reporting θ from Eq. **2** remains unclear, see Fig. 2.

We continue in believing that "all models are wrong but some are useful" (2), which prompts us to interpret *several* $\rho_k(S_k(x), S_k(x_0))$ as *realizations* of real-valued error terms, denoted by $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_K)$ (16). Error terms are not observed, and must be estimated from the data; we develop a theoretical framework and provide an algorithm, ABCμ, for this purpose when the likelihood is intractable. We intentionally focus on the posterior distributions of components of $\boldsymbol{\varepsilon}$ to make probabilistic statements of mismatch between the model and the data (17) and hence to facilitate model criticism, as summarized in Fig. 1*C*.

Postgenomic data such as protein interaction networks (PINs) are now available for a growing number of organisms, (e.g., refs. 18 and 19). They offer a new perspective on the function of all organisms, and are, in addition to individual gene or genomic approaches, increasingly useful to elucidate the evolution of living systems, (e.g., refs. 12, 20, and 21), despite being noisy, incomplete, and static descriptions of the real, transient protein network (22). To elucidate the network evolution of prokaryotes, we here analyze the compatibility of the *Treponema pallidum* and *Helicobacter pylori* PIN datasets with a set of competing models inspired by fundamentally different modes of network evolution.

Author contributions: O.R. and S.R. designed research; O.R. performed research; O.R., C.A., and S.R. analyzed data; and O.R., C.A., C.W., and S.R. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: oliver.ratmann@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/cgi/content/full/0807882106/DCSupplemental.

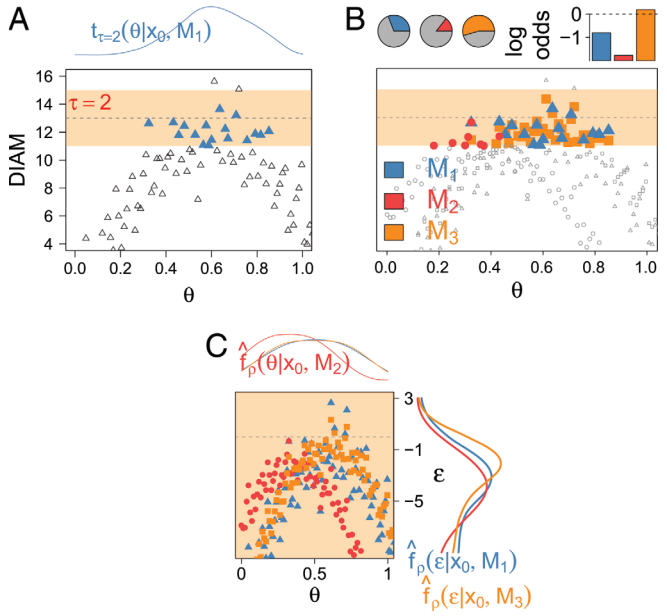


Fig. 1. Comparison of ABC versus our implementation of likelihood-free inference, on a fictitious PIN dataset x_0, fictitious models with a single, common parameter θ, and one summary, DIAM, with observed value 13. The points represent posterior samples of θ and DIAM and resemble more a "bouncy castle" than a likelihood surface. (*A*) In "standard" ABC, inference proceeds by retaining those samples (blue triangles) for which the realized errors are smaller than τ (here, $\tau = 2$), and are taken to approximate the posterior density of θ (*Top*). (*B*) In "standard" ABC, different models (blue, red, yellow) may be compared based on the number of retained samples under one model relative to that number under all other models (*Top Left*), for instance, in terms of the log odds ratio (*Top Right*), here indicating that model M_3 performs best. (*C*) We propose to use the generated data to a fuller extent by augmenting the likelihood (vertical dimension, ε). Discrepancies between the data and the models are made explicit in terms of posterior quantities of ε.

ABC under Model Uncertainty*

Joint Posterior Density of Model Parameters and Summary Errors. For the purpose of model criticism in situations where the likelihood is intractable, define the unknown error ε as the random variable with conditional probability distribution

$$\mathbb{P}_{\theta,x_0}(\varepsilon \le e) = \int_{\mathcal{X}} \mathbf{1}\{\rho(\mathbb{S}(x), \mathbb{S}(x_0)) \le e\} f(x|\theta, M_i)dx^{\dagger}. \quad [3]$$

Next, we assume that $\mathbb{P}_{\theta,x_0}(\varepsilon \le e)$ has a density ξ_{θ,x_0} with respect to an appropriate measure for ρ.† It is natural to suggest using this quantity as an *augmented* likelihood for x_0 under ρ while *adhering* to the current model,

$$\theta, \varepsilon \to f_\rho(x_0|\theta, \varepsilon, M_i) = \xi_{\theta,x_0}(\varepsilon). \quad [4]$$

We thus capture the direct information brought by the discrepancies ρ on θ and/or model M_i in a scalar value. For a given prior $\pi_{\theta,\varepsilon}(\theta, \varepsilon|M_i)$, we embrace two aspects of statistical reasoning, parameter inference and model criticism, *simultaneously* by the joint posterior density

$$f_\rho(\theta, \varepsilon|x_0, M_i) = \xi_{\theta,x_0}(\varepsilon)\pi_{\theta,\varepsilon}(\theta, \varepsilon|M_i)/f_\rho(x_0|M_i), \quad [5]$$

using the data *once*.‡ In practice, we take $\pi_{\theta,\varepsilon} = \pi_\varepsilon \times \pi_\theta$,§ reflecting our inability to quantify a priori model adequacy for a value of θ. The posterior relationship Eq. **5** exploits the dependence between model error and model parameterization. ABC only infers model parameterization from realized model errors after simulation and does not question the adequacy of the likelihood model.

The simplest algorithm to sample from Eq. **5** is:

Std-ABCμ1 Sample $\theta \sim \pi_\theta(\theta \,|\, M_i)$, simulate $x \sim f(\cdot \,|\, \theta, M_i)$ and compute $\varepsilon = \rho(\mathbb{S}(x), \mathbb{S}(x_0))$.

Std-ABCμ2 Accept (θ, ε) with probability proportional to $\pi_\varepsilon(\varepsilon \,|\, M_i)$, and go to Std-ABCμ1.

Interpretation of the Marginals: Parameter Inference and Model Criticism. The thrust of this article is to recognize the utility of the unknown error ε for model criticism. By design, nonzero values of ρ indicate discrepancies between the model and the data, so that intuitively, only if the model matches the data, we expect the mode of $\xi_{\theta,x_0}(\varepsilon)$ to be on average zero for some value of θ if the summaries behave sufficiently well. Parameter inference based on the *marginal* posterior distribution $f_\rho(\theta|x_0, M_i)$ is justified in an "approximate likelihood" sense, because, under regularity assumptions (*SI Appendix*, S1.1) on $\xi_{\theta,x_0}(\varepsilon)$ which we assume throughout this section,

$$f_\rho(\theta|x_0, M_i) \propto \int \pi_\varepsilon(\rho(\mathbb{S}(x), \mathbb{S}(x_0))|M_i) f(x|\theta, M_i)\pi_\theta(\theta|M_i)dx. \quad [6]$$

Setting $\pi_\varepsilon(\varepsilon|M_i) = \mathbf{1}\{|\varepsilon| \le \tau/2\}/\tau$, we recover the "standard" ABC approximation Eq. **2**; please see *SI Appendix*, S1.2 for more details and examples. We can interpret the variety of ABC kernels as exerting a particular prior belief on the adequacy of the current model (23). In agreement with methods of ABC (*SI Appendix*, S1.2), we always choose a prior $\pi_\varepsilon(\varepsilon|M_i)$ with mode at zero to accommodate a prior belief that the model is plausible. For the purpose of parameter inference, it is sufficient to "plug-in" realized errors in Eq. **6**, but here we also focus on the *marginal* posterior error $f_\rho(\varepsilon|x_0, M_i)$. For the prior predictive error density $L_\rho(\varepsilon) = \int \delta\{\rho(\mathbb{S}(x), \mathbb{S}(x_0)) = \varepsilon\}\pi(x|M_i)dx$ we have that

$$f_\rho(\varepsilon|x_0, M_i) = L_\rho(\varepsilon)\pi_\varepsilon(\varepsilon|M_i)/f_\rho(x_0|M_i) \quad [7]$$

(see *SI Appendix* S1.1). Hence, $f_\rho(\varepsilon|x_0, M_i)$ can be understood as an error density under the prior predictive distribution that is *weighted* according to error magnitude. Small error boosts the prior belief for a particular value of θ, see Eq. **6**. We thus prefer model criticism based on Eq. **7** rather than $L_\rho(\varepsilon)$ as it focuses on those θ actually inferred from the perspective of Eq. **5**, *and* attenuates the dependence of $L_\rho(\varepsilon)$ on $\pi_\theta(\theta|M_i)$. This dependence is undesirable in that a faultless model could appear questionable under unfortunate prior choice (*SI Appendix*, S1.3). For the practitioner, we provide a computationally feasible method for model criticism within the prior predictive setting as an alternative to using data-splitting techniques that are here difficult or too expensive to construct (5, 7, 17).

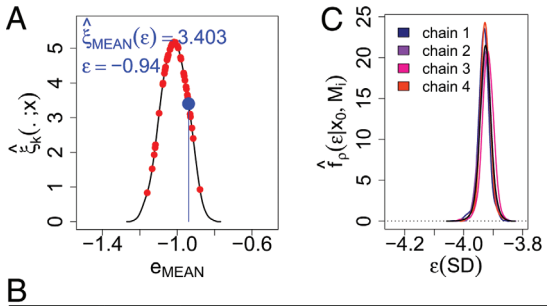
Multidimensional Error Terms $\boldsymbol{\varepsilon}$. The complexity of the settings to which ABC is typically applied makes it difficult to think of a universal discrepancy function ρ. The joint posterior distribution of multiple errors $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_K)$, corresponding to K discrepancies

*For ease of exposition, we start with a *scalar* error term ε corresponding to a univariate discrepancy ρ, and later generalize to multidimensional error terms. In ABC, a set $\mathbb{S}$ of summaries is commonly combined into the univariate ρ; at a first reading it may help to think of $\mathbb{S}$ as a single summary. In particular, it may be useful to take $f(x|\theta, M_i)$ as the one-dimensional Gaussian density with mean θ and fixed variance, and $\rho(\mathbb{S}(x), \mathbb{S}(x_0))$ as the difference $x - x_0$.

†We denote the Indicator function with **1**, and particular limits of a sequence of functions with δ (see Eq. **S1** in *SI Appendix*, S1.1). If ρ is continuous, ξ_{θ,x_0} is taken with respect to the Lebesgue measure; in many applications, $\mathcal{X}$ is a finite set and ξ_{θ,x_0} is then understood with respect to a counting measure.

‡Our developments are subject to the integrability of Eq. **5**.

§For clarity, we subscript π_θ and π_ε to denote the priors in θ and ε, respectively. From now on, we drop the conditioning of π_ε on θ. Finally, we denote with $\pi(x|M_i)$ the prior predictive density $\int f(x|\theta, M_i)\pi(\theta|M_i)d\theta$.



B

	$\theta\|x_0$	$\varepsilon_{MEAN}\|x_0$	$\varepsilon_{SD}\|x_0$	$\varepsilon_{QU0.25}\|x_0$
ABCμ	5.0	[-0.26,0.29]		
ABCμ	3.2	[-2.6,-1.02]	[-3.95,-3.93]	[0.2,1.8]
		τ_{MEAN}	τ_{SD}	$\tau_{QU0.25}$
ABC	3.2	2.25	4.0	1.5

Fig. 2. $\mathcal{N}(\theta, 1)$ toy example to illustrate our approach to diagnose model mismatch with posterior densities of multiple error terms when the likelihood is intractable. Suppose we have obtained a dataset x_0 of 200 independent samples. We believe each sample of x_0 to be generated from $f(\cdot|\theta, M_1) = \mathcal{N}(\theta, 1)$ with θ unknown, whereas in reality x_0 is exponential with rate $\theta_t = 0.2$ (denoted with M_t). By construction, the sample mean (MEAN) is a sufficient statistic to estimate both θ and θ_t. To illustrate one iteration of ABCμ, suppose we sample $\theta = 3$, $\varepsilon_{MEAN} = -0.94$ from the priors (respectively, uniform and exponential). We generate 50 errors $\overline{x_b} - \overline{x_0}$ (*A*, red points for $x_b \sim f(\cdot|\theta = 3, M_1)$), estimate the associated error density $\hat{\xi}(\cdot;\boldsymbol{x})$ (A, black line) with a biweight kernel, and compute $\hat{\xi}(\varepsilon_{MEAN};\boldsymbol{x})$ (*A*, blue point). Using algorithm ABCμ, we estimated posterior quantities of θ and ε under various summary statistics. Summarizing x only with MEAN, the method indicates no model mismatch (*B*, row 1, column 2; 95% high posterior density interval (HPD)) and θ is estimated as if x_0 were indeed $\mathcal{N}(1/\theta_t, 1)$ (*B*, row 1, columen 1; posterior mean). Reminding ourselves that likelihood-free inference is honest in that inference is here based on $\overline{x_b} - \overline{x_0}$, we see that the algorithm samples correctly from Eq. **9**. Observing that under M_1 the standard deviation (SD) is 1 independently of θ and that SD is $1/\theta_t$ under M_t, we recognize that progress is possible when a comprehensive set of summaries is employed. Repeating our method based on MEAN, SD, and the 0.25 quantile (QU0.25), targetting Eq. **9**, we find that all error terms indicate model mismatch (*B*, row 2, column 2-4; and *C*, posterior error densities for 4 runs of ABCμ starting from overdispersed initial values (colored solid lines) versus the (dashed) prior $\pi_{\varepsilon_k}(\varepsilon_k|M_i)$). If posterior quantities of comprehensive error terms clearly diagnose model mismatch as in this example, we recommend questioning the interpretability of $\theta|x_0$ in terms of the likelihood model. For reference, we applied standard (Rejection) ABC to this example; conditioning on the summaries MEAN, SD, and QU0.25, we find that the numerical estimates of $\theta|x_0$ agree between ABC and ABCμ (with τ fixed as in *B*, third row).

$\rho_k(S_k(x), S_k(x_0))$ (*SI Appendix*, S1.4), facilitates to *diagnose* model mismatch more systematically and comprehensively; we have

$$f_\rho(\theta, \boldsymbol{\varepsilon}\,|\,x_0, M_i) \propto \xi_{\theta,x_0}(\boldsymbol{\varepsilon})\pi_\theta(\theta\,|\,M_i)\pi_\varepsilon(\boldsymbol{\varepsilon}\,|\,M_i). \quad [8]$$

In the ABC literature, it has been recognized that attempting to match jointly a set of summaries is too conservative, and instead a linear combination of summaries is typically employed (14). Nonetheless, we believe that each summary captures aspects of model discrepancy. To control several summaries stringently for *accurate* and *robust* parameter inference (12), $\min_k \xi_{k,\theta,x_0}(\varepsilon_k)$ here supersedes $\xi_{\theta,x_0}(\boldsymbol{\varepsilon})$ (see *Materials and Methods*, section 3, and *SI Appendix*, S1.6).

Algorithm. The major impediment in ABC—that the likelihood surface is turned into a "bouncy castle," see Fig. 1—is in the multivariate case exacerbated by the fact that the unknown error terms are correlated *by design*, and easily outnumber the θ's. To obtain a *smoothed, stabilized* approximation to $\xi_{k,\theta,x_0}(\xi_k)$ that better controls the volatility of the simulated datasets, we employ kernel density estimates $\hat{\xi}_k(\varepsilon_k;\boldsymbol{x}) := 1/(Bh_k)\sum_{b=1}^{B} K([\varepsilon_k - \rho_k(S_k(x_b), S_k(x_0))]/h_k)$ in line with ABC (*SI Appendix*, S1.5). In theory, this corresponds to replacing $\xi_{\theta,x_0}(\boldsymbol{\varepsilon})$ in Eq. 8 with $\min_k \int h_k^{-1} K((\varepsilon_k - v_k)/h_k)\xi_{\theta,x_0}(\boldsymbol{v})d\boldsymbol{v}$. In practice (*SI Appendix*, S1.7), we set $B = 50$ and attain under technical modifications (see *Material and Methods*, section 3) a *smoothing approximaion*

$$\hat{f}_{\boldsymbol{\rho}}(\theta, \boldsymbol{\varepsilon}, \boldsymbol{x}|x_0, M_i) \propto \pi_\theta(\theta|M_i)\pi_{\boldsymbol{\varepsilon}}(\boldsymbol{\varepsilon}|M_i)\min_k \hat{\xi}_k(\varepsilon_k;\boldsymbol{x})f(\boldsymbol{x}|\theta, M_i) \quad [9]$$

on the auxiliary space $(\theta, \boldsymbol{\varepsilon}, \boldsymbol{x})$. Various Monte Carlo strategies (8, 24) may be devised to sample from Eq. **9** (*SI Appendix*, S1.8); our MCMC implementation (*SI Appendix*, S1.9), particularly addresses the codependencies of ρ_k with a careful choice of $q(\boldsymbol{\varepsilon} \to \boldsymbol{\varepsilon}')$. Suppose an initial sample $(\theta, \boldsymbol{\varepsilon})$ and prior specifications (*SI Appendix*, S1.10);

ABCμ1 if now at θ, move to θ' according to $q(\theta \to \theta')$ (*SI Appendix*, S1.12).

ABCμ2 Generate $\boldsymbol{x}' \sim f(\cdot|\theta', M_i)$, and construct $\hat{\xi}_k(\cdot;\boldsymbol{x}')$ for all k. If now at $\boldsymbol{\varepsilon}$, move to $\boldsymbol{\varepsilon}'$ according to $q(\boldsymbol{\varepsilon} \to \boldsymbol{\varepsilon}')$. We guide this proposal with $\hat{\xi}_k(\cdot;\boldsymbol{x})$ and $\hat{\xi}_k(\cdot;\boldsymbol{x}')$ (*SI Appendix*, S1.12).

ABCμ3 Accept $(\theta', \boldsymbol{\varepsilon}', \boldsymbol{x}')$ with probability

$$\min\left\{1, \frac{\pi(\theta', \boldsymbol{\varepsilon}'|M_i)q(\theta' \to \theta)q(\boldsymbol{\varepsilon}' \to \boldsymbol{\varepsilon})}{\pi(\theta, \boldsymbol{\varepsilon}|M_i)q(\theta \to \theta')q(\boldsymbol{\varepsilon} \to \boldsymbol{\varepsilon}')} \times \frac{\min_k \hat{\xi}_k(\varepsilon_k';\boldsymbol{x}')}{\min_k \hat{\xi}_k(\varepsilon_k;\boldsymbol{x})}\right\},$$

and otherwise stay at $(\theta, \boldsymbol{\varepsilon}, \boldsymbol{x})$, then return to ABC$\mu$1.

Please see *Materials and Methods*, section 4, for a technical discussion and Tables 1 and 2 for a comparison of the efficiency of ABCμ with related samplers.

Model Criticism by Revealing Model Inconsistency Across Discrepancies. For large datasets and/or complex models, the discrepancies ρ_k are often codependent (5, 17). Our approach to model criticism capitalizes on the fact that the co-dependencies among

Table 1. Acceptance rate and average mixing quality in θ

Algorithm	B	acc.prob	Burn-in	$10^3 \frac{n_{eff}(\theta)}{n}$	$\frac{n_{eff}(\theta)}{CPU\ hr}$
ABC-MCMC (26)	1	0.002	10^5	0.7	13.4
Zoom-ABC-MCMC (12)	50/1	0.002	800	0.7	23.8
AUX-ABC (27)	1	0.01	10^5	0.6	4.5
ABCμ with asymmetric walk in ε, Eq. **S12**	50	0.36	771	25.6	24.6

Performance results are obtained from inference from the *H. pylori* PIN dataset; tuning parameters have been optimized for each algorithm separately (*SI Appendix*, S1.15). The effective sample size n_{eff} is taken as an indicator of mixing quality across n iterations (*SI Appendix*, S1.14). Importantly, $n_{eff}(\theta)$/CPU hr must be compared relative to the achieved absolute errors (see further *SI Appendix*, Table S2). Higher acceptance rates are not necessarily desirable. As a rule of thumb, we found that, here, rates >0.45 reduced the effective sample size.

Table 2. Mixing quality of the unknown error terms

		$10^3 n_{eff}/n$ for ε of the summaries				
Algorithm	B	WR	DIA	CC	$\overline{ND}$	FRAG
AUX-ABC (27) with GRW in ε	1	0.1	0.23	0.13	0.07	0.05
ABCμ with GRW in ε	50	10.6	3.1	14.9	3.1	2.6
ABCμ with asymmetric walk in ε, Eq. **S12**	50	140	41.7	48.1	71.6	101

Mixing quality is quantified with the effective sample size n_{eff} for each error ε_k (ABCμ) or random mismatch threshold (AUX-ABC), standardized per 1,000 iterations; tuning parameters have been optimized for each algorithm (*SI Appendix*, S1.15).

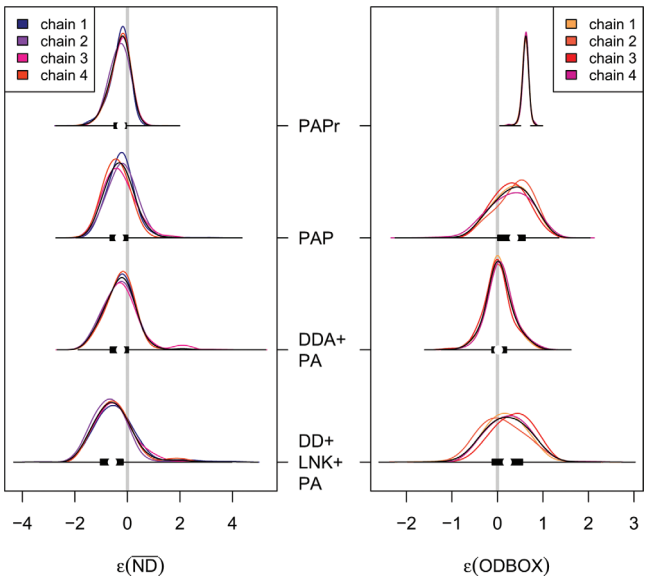


Fig. 3. Numerical estimates, obtained from ABCμ, of the approximate posterior error densities $\hat{f}_{\rho}(\varepsilon_k|x_0, M_i)$ combined with 50% box plots (black bars) for 2 of 7 summaries to quantify departures of 3 competing models of network evolution to the *T. pallidum* PIN dataset; PAPr employs sampling scheme RS1, whereas all others use sampling scheme RS2. ODBOX of PAPr is miniaturized by a factor of 5 to improve the visualization of differences across models for RS2. Whereas PAPr and PAP depart in ODBOX from the data, DD+LNK+PA departs (slightly) in $\overline{\text{ND}}$ from the observed PIN, suggesting that only DDA+PA provides an adequate fit the *T. pallidum* dataset.

$S_k(x)$ under the predictive distribution $\pi(x|M_i)$ are typically different from those among $S_k(x_0)$ if the model is not adequate, revealing *model inconsistency* in terms of conflicting, codependent summaries. As exemplified in Fig. 2, only a *comprehensive* set of summaries may enable model criticism; it is our view that choosing comprehensive summaries and discrepancy functions is crucial to ensure the approximation quality of Eq. **6** to the likelihood (12) as well as for model criticism based on posterior densities of summary errors (Eq. **7**). To explore model adequacy, we recommend investigating various posterior quantities of $\hat{f}_{\rho}(\boldsymbol{\varepsilon}|x_0, M_i)$ and using centrality measures such as high probability density (HPD) intervals; we remark at this point that marginal properties are in our setting typically not independent and caution against the overinterpretation of marginal diagnostics (see further Fig. 3).

Example: Criticizing Models of PIN Evolution

The structure of PINs derives from multiple stochastic processes over evolutionary timescales, and a number of mechanisms, based on randomly growing graphs, have been proposed to capture aspects of network growth (ref. 29 and references therein). We briefly motivate three models of network evolution. Recent comprehensive analyses across 181 prokaryotic genomes suggest that lateral gene transfer probably occurs at a low rate, but that, cumulatively, ≈80% of all genes in a prokaryotic genome are involved in lateral gene transfer (30); model PAP[¶] is inspired by this scenario as it proposes network evolution in terms of attachment processes only (*Materials and Methods*, section 2). At least 40% of genes in prokaryotes appear to be products of gene duplication (31). Model DDA+PA (*Materials and Methods*, section 2) is designed to quantify the potential role of duplication and divergence in network evolution (12). At least for eukaryotes, the formation or degeneration of functional links between proteins (link turnover) is estimated to occur at a fast rate of $\approx 10^{-5}$ changed interactions per My per protein pair (20). We extend model DDA+PA into Model DD+LNK+PA (*Materials and Methods*, section 2), which includes link turnover in terms of preferential loss and gain of protein interactions. Crucially, ABC enables us to account for data incompleteness. Previously, we modeled missing data by randomly sampling proteins from the simulated data (RS1) (12). Here, we examine an alternative model that randomly samples from those proteins that have an interaction in the simulated data (RS2). (For the former, we add "r" to the model acronyms.) Necessarily, all models remain conceptually limited and must be cautiously interpreted; for example, the assumption that the network as a whole evolves at homogeneous rates has been questioned (32).

Models of Network Evolution Inspired by Horizontal Gene Transfer, Duplication-Divergence, and Link Turnover. We ask whether the *T. pallidum* PIN topology is compatible with a number of fundamentally different modes of network evolution, in guise of simplified models. We successfully checked (*Materials and Methods*, section 5) and applied ABCμ (*Materials and Methods*, section 6) to sample from $\hat{f}_{\rho}(\theta, \boldsymbol{\varepsilon}|x_0, \cdot)$ for the models PAP, DDA+PA, and DD+LNK+PA (under both sampling schemes RS1 and RS2); see Fig. **3** and *SI Appendix*, Figs. S5 and S6. Based on RS1 (12), all models depart significantly in FRAG as exemplified for PAPr in Table 3. This motivated us to consider alternative models of missing data, and we found no significant departures in FRAG for any of the considered evolution models under RS2; see Table 3. Turning to the 6 remaining discrepancy functions, we observe (Table 3 and Fig. 3), that only model DDA+PA matches the *Treponema pallidum* PIN adequately, suggesting that an evolutionary mode of duplication-divergence is most consistent with the *T. pallidum* PIN dataset. Repeating our analysis based on RS2 for the *Helicobacter pylori* PIN dataset, we could not substantiate our results further, because all considered models provide an adequate fit to the data. This is surprising, because we expect a similar power of our method on both datasets (*SI Appendix*, S1.19), and may point to qualitative differences among the two PINs owing to different, underlying experimental protocols.

Conclusion

The growing complexity of realistic models renders Bayesian model criticism increasingly important and difficult. In this article, we provide Bayesian techniques to *comprehensively quantify*

[¶] Model acronyms are explained in Materials and Methods, section M2, with underlined characters.

Table 3. Fifty percent high probability density intervals of $\varepsilon_k|x_0$, indicating model mismatch relative to the *T. pallidum* PIN

M_i	$\varepsilon_{\text{CONN}}$	ε_{WR}	$\varepsilon_{\text{ODBOX}}$	ε_{DIA}	ε_{CC}	$\varepsilon_{\overline{\text{ND}}}$	$\varepsilon_{\text{FRAG}}$
PAPr	[−0.11, 0.11]	[−0.11, 0.15]	**[0.28, 0.72]**	[−0.16, 0.92]	**[-0.012,-0.002]**	[−0.20, 0.05]	**[0.08, 0.18]**
PAP	[−0.15, 0.14]	[−0.12, 0.16]	**[0.02, 0.61]**	[−0.60, 0.91]	**[-0.010,-0.003]**	[−0.66, 0.01]	[−0.16, 0.01]
DDA+PA	[−0.24, 0.29]	[−0.34, 0.48]	[−0.12, 0.20]	[−0.48, 0.50]	[−0.002, 0.027]	[−0.66, 0.05]	[−0.01, 0.29]
DD+LNK+PA	[−0.27, 0.21]	[−0.20, 0.17]	[−0.11, −0.55]	[−0.59, 0.32]	[**−0.018**,**−0.006**]	[**−1.03**,**−0.16**]	[0,0.11]

Note that the scales of ε_k correspond to the scales of the summaries, so that small numbers are meaningful. We caution against overinterpretation because the errors are not independent. See *Materials and Methods*, section 6, for further details.

discrepancies between the likelihood model and the data, simultaneously with parameter inference in situations *when the likelihood is intractable*, thus providing valuable *guidance* on the interpretability of parameter estimates and on how to improve models. We found this methodology helpful in iteratively identifying an adequate model of network evolution in terms of a large number of summaries; in particular, the PIN topology of the prokaryote *T. pallidum* provides little support for network evolution dominated by link turnover, or by lateral gene transfer alone. We close by cautioning that it is difficult to convincingly associate a formal framework to our high probability density intervals on multiple diagnostic error terms (6, 7). The presented methods will be useful in the initial stages of model and data exploration (16), and in particular, in efficiently scrutinizly several models by direct, inspection of their summary errors (5), prior to more formal analyses (14).

Materials and Methods

1. Summaries. PINs can be described as graphs that contain a set of nodes, interacting proteins, and undirected binary edges, representing the observed interactions between the proteins. We consider the following topological summary statistics of PINs: Order, the number of nodes; size, the number of edges; node degree, the number of edges associated with a node; $\overline{\text{ND}}$, average node degree; distance, the minimum number of edges that have to be visited to reach a node j from node i; WR, within-reach distribution, the mean probability of how many nodes are reached from one node within distance $k = 1, 2, \ldots$ (12); DIA, diameter, the longest minimum path among pairs of nodes in a connected component; CC, cluster coefficient, the mean probability that 2 neighbors of a node are themselves neighbours; BOX, the number of 4-cycles with 4 edges among the 4 nodes; FRAG, fragmentation, the percentage of nodes not in the largest connected component; CONN, log connectivity distribution, $\log(p(k_1, k_2)\overline{\text{ND}}^2)/(k_1 p(k_1) k_2 p(k_2))$, the depletion or enrichment of edges ending in nodes of degree k_1, k_2 relative to the uncorrelated network with same node degree distribution; ODBOX, BOX degree distribution, the probability distribution of BOXes with k edges to nodes outside the BOX. Examples are provided in *SI Appendix*, Table S1.

2. Algorithmic Details of the Models of Network Evolution. Given a PIN x_0, we simulate a network under a given model to the number of genes in the respective genome, and account for incompleteness in x_0 by either RS1 or RS2. In model PAP evolution proceeds only by preferential attachment (33); at each step the number of attachments minus one is Poisson distributed with mean m. DDA+PA (12) features preferential attachment of a new node to one node of the existing network with probability α, or, with probability $1 - \alpha$, a step of node duplication and immediate link divergence. In the latter case, a parent node is randomly chosen and its edges are duplicated. For each parental edge, the parental and duplicated one are then lost with probability δ_{Div} each, but not both; moreover, at least one link is retained to any node. The parent node may be attached to its child with probability δ_{Att}. DD+LNK+PA is a mixture of duplication-divergence (as above with parameter δ_{Div} but fixed $\delta_{\text{Att}} = 0$), link addition and deletion, and preferential attachment as in DDA+PA. Link addition (deletion) proceeds by choosing a node randomly, and attaching it preferentially to another node (deleting it preferentially from its interaction partners) (20). At each step unnormalized weights are calculated as follows. For duplication-divergence, the rate κ_{Dup} is multiplied by the order of the current network; for link addition, the rate κ_{LnkAdd} is multiplied by $\binom{\text{Order}}{2} - \text{size}$; for link deletion, the unnormalized weight of link addition is multiplied by κ_{LnkAdd}. Preferential attachment occurs at a constant frequency α, and the weights of duplication, link addition, and link deletion are normalized so that their sum equals $1 - \alpha$. Each of the components is chosen with these weights; the parameter ranges are determined by the prior (*SI Appendix*, S1.10).

3. Combining Multiple Error Terms. It is difficult to compare $\hat{\xi}_{k,\theta,x_0}(\varepsilon_k)$ across k without further transformation, because summaries differ in their sensitivity to changes in θ (12) so that the scales of the density estimates vary across summaries and (to a lesser extent) across θ; see *SI Appendix*, Fig. S1. In Rejection-ABCμ (*SI Appendix*, S1.8), summaries may be precomputed and standardized, but this is not applicable in MCMC. We propose to standardize the variance of each $\hat{\xi}_k$ to one, bearing in mind that this might reduce approximation quality in some cases.

4. Details of ABCμ. ABCμ is similar to the MCMC algorithm proposed in ref. 27; the latter also extends the state space, but includes a *scalar* τ (circumventing the need to design an efficient proposal $q(\boldsymbol{\varepsilon} \to \boldsymbol{\varepsilon}')$). We show that ABCμ eventually samples from $\hat{f}_{\boldsymbol{\rho}}(\theta, \boldsymbol{\varepsilon}, \boldsymbol{x})$, provide convergence results for the smoothing approximation (*SI Appendix*, S1.5 and S1.11), discuss our non-standard proposal kernel (*SI Appendix*, S1.12), and provide final details (*SI Appendix*, S1.13). We do not claim that our smoothing approach based on repeated sampling from $f(\cdot|\theta, M_i)$ comes at no cost. What we contend is that (*i*) we obtain *improved mixing quality* relative to ABC within MCMC, owing to a *stabilized, numerical* approximation of the likelihood with Eq. **9**, and (*ii*) that we can construct *more efficient proposal kernels*, a point particularly relevant for ABCμ, where the number of error terms easily exceeds the dimensionality of θ. With respect to (*i*), we compared ABCμ with ABC-MCMC (26), zoomABC-MCMC (12), and AUX-ABC (27) on the *H. pylori* PIN dataset (*SI Appendix*, S1.15). Table 1 illustrates that ABCμ results in much improved acceptance rates and better mixing; this has already been suggested by Becquet and Przeworski (28) when $f(\mathbb{S}(\boldsymbol{x})|\theta, M_i)$ is available in closed form. As for (*ii*), we devised a guided, asymmetric random walk in $\boldsymbol{\varepsilon}$ (*SI Appendix*, S1.12). This greatly improved both overall acceptance rate and mixing in $\boldsymbol{\varepsilon}$ compared to AUX-ABC and ABCμ with a (symmetric) Gaussian random walk (GRW) in $\boldsymbol{\varepsilon}$ (Table 2), exemplifying that *effectively*, repeated sampling may improve the efficiency of standard MCMC methods.

5. Testing ABCμ on PINs. It was unclear whether our implementation is efficient enough to sample from Eq. **9**. First, estimates of Eq. **7** might be inherently biased as for technically similar algorithms, and/or the PIN topology, in terms of the chosen summaries, might not be informative enough to evidence discrepancies between the model and the data. Second, given our smoothing approximation based on an adaptively chosen bandwidth $h = h(\boldsymbol{x})$, we might be worried that posterior quantities of θ may be unreliable. We have addressed both concerns empirically (*SI Appendix*, S1.16), comforting that ABCμ provides useful numerical estimates of $\hat{f}_{\boldsymbol{\rho}}(\boldsymbol{\varepsilon} \mid x_0, M_i)$ to criticize the models of network evolution considered here, and suggesting that samples $\theta \mid x_0$ from the marginal of Eq. **9**, obtained by ABCμ, provide a good approximation to $f_{\boldsymbol{\rho}}(\theta \mid x_0, M_i)$.

6. Criticizing Models of Network Evolution. To contrast models PAP, DDA+PA, and DD+LNK+PA to the *T. pallidum* PIN dataset, ABCμ based on the summaries CONN, WR, ODBOX, DIA, CC, $\overline{\text{ND}}$, FRAG, and $\tau(\varepsilon|\text{PAP}) = (0.2, 0.2, 1.4, 1, 0.007, 0.7, 0.4)$, $\tau(\varepsilon|\text{DDA} + \text{PA}) = (1, 0.7, 0.8, 0.5, 0.05, 0.5, 0.4)$, and $\tau(\varepsilon|\text{DD} + \text{LNK} + \text{PA}) = (0.3, 0.3, 0.5, 0.7, 0.02, 1.1, 0.25)$ were used to generate 4 Markov chains as in *SI Appendix*, S1.15.

ACKNOWLEDGMENTS. We thank M.P.H. Stumpf for stimulating discussions, T. Hinkley for providing an efficient C++ library to evaluate network summaries, and M. Sternberg for comments on an earlier version of the manuscript, and two anonymous referees for valuable comments on an earlier version of this article. Computations were performed at the Imperial College High Performance Computing Centre http://www3.imperial.ac.uk/ict/services/teachingandresearchservices/highperformancecomputing. O.R. was supported by the Wellcome Trust; C.A. by an Advance Research Fellowship from the Engineering and Physical Sciences Research Council, C.W. by the Danish Cancer Society and the Danish Research Councils, and S.R. by the Biotechnology and Biological Sciences Research Council and the Centre for Integrative Systems Biology at Imperial College.

1. May RM (2004) Uses and abuses of mathematics in biology. *Science* 303: 790–793.
2. Box GEP (1976) Science and statistics. *J Am Stat Assoc* 71:791–799.
3. Bernado JM, Smith AFM (1994) *Bayesian Theory* (Wiley & Sons, Chichester, UK), 1st Ed.
4. Box GEP (1980) Sampling and Bayes' inference in scientific modelling and robustness. *J R Soc A (General)* 143:383–430.
5. Gelfand AE, Dey DK, Chang H (1992) *Bayesian Statistics 4*, eds Bernardo JM, Berger JO, Dawid AP, Smith AFM (Oxford Univ Press, Oxford), pp 147–167.
6. Meng XL (1994) Posterior predictive p-values. *Ann Stat* 22(3):1142–1160.
7. Bayarri MJ, Berger JO (1999) *Bayesian Statistics 6*, eds Bernardo JM, Berger JO, Dawid AP, Smith AFM (Oxford Univ Press, Oxford), pp 53–82.
8. Liu JS (2001) *Monte Carlo Strategies in Scientific Computing* (Springer, New York), 343 pp.
9. Gouriéroux C, Monfort A (1996) *Simulation-Based Econometric Methods* (Oxford Univ Press, Oxford).
10. Marjoram P, Tavaré S (2006) Modern computational approaches for analysing molecular genetic variation data. *Nat Rev Genet* 7:759–770.
11. Riley S, et al. (2003) Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of public health interventions. *Science* 300:1961–1966.
12. Ratmann O, et al. (2007) Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H.pylori and P.falciparum. *PLoS Comp Biol* 3:e230.
13. Wilkinson RD (2007) Bayesian inference of primate divergence times. PhD thesis (Univ of Cambridge, Cambridge).
14. Fagundes NJR, et al. (2007) Statistical evaluation of alternative models of human evolution. *Proc Natl Acad Sci USA* 104:17614–17619.
15. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH (2008) Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J R Soc Interf*, 10.1098rsif. 2008.0172.
16. Zellner A (1975) Bayesian analysis of regression error terms. *J Am Stat Assoc* 70:138–144.
17. O'Hagan A (2003) *Highly Structured Stochastic Systems*, eds Green PJ, Hjort NL, Richardson S (Oxford Univ Press, Oxford), pp 423–453.
18. Rain JC, et al. (2001) The protein-protein interaction map of Helicobacter pylori. *Nature* 409:211–215.

19. Titz B, et al. (2008) The binary protein interactome of Treponema pallidum—The Syphilis spirochete. *PLoS ONE* 3:e2292.
20. Beltrao P, Serrano L (2007) Specificity and evolvability in eukaryotic protein interaction networks. *PLoS Comp Biol* 3:e25.
21. Pinney JW, Amoutzias GD, Rattray M, Robertson DL (2007) Reconstruction of ancestral protein interaction networks for the bZIP transcription factors. *Proc Natl Acad Sci USA* 104:20449–20453.
22. Hakes L, Pinney JW, Robertson DL, Lovell SC (2008) Protein-protein interaction networks and biology—What's the connection? *Nat Biotechnol* 26: 69–72.
23. Wilkinson RD (2008) Approximate Bayesian Computation (ABC) gives exact results under the assumption of model error. arXiv:0811.3355v1 [stat.CO].
24. Sisson SA, Fan Y, Tanaka MM (2007) Sequential Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 104:1760–1765.
25. Gilks WR, Richardson S, Spiegelhalter DJ (1996) *Markov Chain Monte Carlo in Practice* (Chapman & Hall, London).
26. Marjoram P, Molitor J, Plagnol V, Tavaré S (2003) Markov Chain Monte Carlo without likelihoods. *Proc Natl Acad Sci USA* 100:15324–15328.
27. Bortot P, Coles S, Sisson S (2007) Inference for stereological extremes. *J Am Stat Assoc* 102:84–92.
28. Becquet C, Przeworski M (2007) A new approach to estimate parameters of speciation models with application to apes. *Genome Res* 17:1505–1519.
29. Knudsen M, Wiuf C (2008) A Markov chain approach to randomly grown graphs. *J Appl Math*, 10.1155/2008/190836.
30. Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 105:10039–10044.
31. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300:1701–1703.
32. Davidson EH, Erwin DH (2006) Gene regulatory networks and the evolution of animal body plans. *Science* 311:796–800.
33. Barabási A, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512.