LIKELIHOOD-FREE MODEL CRITICISM APPLIED TO STOCHASTIC MODELS OF PROTEIN NETWORK EVOLUTION

Oliver Ratmann[1]*, Christophe Andrieu*[2]*, Carsten Wiuf*[3]*, and Sylvia Richardson*[4]

[1]Department of Public Health and Epidemiology, Imperial College London, London, UK
[2]Department of Mathematics, University of Bristol, Bristol, UK
[3]Bioinformatics Research Center, University of Aarhus, Aarhus, Denmark
[4]Centre for Biostatistics, Imperial College London, London, UK

ABSTRACT

Various models of network evolution are used in systems biology with little or no statistical justification. These may be described as generative as they are framed in terms of incremental network growth, node-by-node, on an abstract, discrete timeline, and neglect the actual history of the proteome. However, despite their simplicity the likelihood of an observed network is computationally difficult to calculate; thus goodness-of-fit and model criticism are difficult to perform and rarely attempted. Here, we provide a general Bayesian framework for model criticism and apply it to two real PIN data sets using different models of network evolution. Using novel statistical techniques appropriate for the analysis of generative models of network evolution with network data, our results support the view that (1) even in abstract terms network evolution cannot be understood as a process of repeated preferential attachment and that (2) a better appreciation of the various forms of measurement error and the reported data are essential to making inference of existing network data.

1. INTRODUCTION

Much of statistical reasoning proceeds in an iterative process between data acquisition, data analysis, and model development [1]. At the ith iteration, the interpretation of observed data x_0 in terms of some parameter θ under model M_i has a long tradition in Bayesian inference [2]. The focus is typically on the posterior density $f(\theta|x_0, M_i)$, which is related to the likelihood $f(x_0|\theta, M_i)$ and the prior $\pi(\theta|M_i)$ via Bayes' Theorem:

$$f(\theta|x_0, M_i) = f(x_0|\theta, M_i)\pi(\theta|M_i) \,/\, f(x_0). \quad (1)$$

Whether the current model M_i is consonant with x_0 can be difficult to explore if the likelihood is computationally intractable. In systems biology this is often the case; see e.g. [3, 4]. 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 [5], proposes to infer θ by comparing simulated data x to the observed data x_0, in terms of a real-valued discrepancy ρ that combines a set of (computationally tractable) summaries $\mathbb{S} = (S_1, \ldots, S_k, \ldots, S_K)$. For convenience, define $s_0 = \mathbb{S}(x_0)$ and $s_x = \mathbb{S}(x)$.

In its simplest form, values of θ for which the discrepancies are within $\tau \geq 0$ are retained to define the "approximate likelihood"

$$t_\tau(\theta) = \frac{1}{\tau}\int_{\mathcal{X}} \mathbf{1}\Big\{|\rho(s_x, s_0)| \leq \tau/2\Big\} f(x|\theta, M_i)dx \quad (2)$$

in the sense that as $\tau \to 0$, $t_\tau(\theta)$ should approach the likelihood of the summaries, $f_{\mathbb{S}}(s_0|\theta, M_i)$. ABC may be embedded into Bayesian methods to formally select one model from a specified collection of models, or to average them [6, 4]. 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.

Here we interpret $\rho(s_x, s_0)$ as a *realization* of a real-valued error term [7]. The error term is not observed (only x_0 is) and must be estimated from the data; we develop a theoretical framework and provide a novel algorithm for this purpose. We focus on the posterior distribution of the error term to make probabilistic statements of mismatch between the model and the data [8] and hence to facilitate model criticism [9].

Postgenomic data such as protein interaction networks (PINs) are now available for a growing number of organisms, e.g. [10, 11]. 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. [3, 12, 13], despite being noisy, incomplete and static descriptions of the real, transient protein network [14]. To elucidate the network evolution of prokaryotes, we here analyze the compatibility of the *T. pallidum* and *H. pylori* PIN datasets ([10, 11]) with a set of competing models inspired by fundamentally different models of network evolution.

2. ABC UNDER MODEL UNCERTAINTY

For ease of exposition, we consider a *scalar* error term corresponding to a univariate discrepancy ρ that combines all summaries. Multi-dimensional error terms will be discussed in a subsequent section.

2.1. Joint posterior density of model parameters and summary errors

For the purpose of model criticism, define the unknown error ε as the random variable with probability distribution

$$\mathbb{P}_{\theta,x_0}(\varepsilon \leq e) = \int_{\mathcal{X}} \mathbf{1}\Big\{\rho(s_x, s_0) \leq e\Big\} f(x|\theta, M_i)dx; \quad (3)$$

in many applications $\mathcal{X}$ is finite and the integral is replaced by a sum. Assume that $\mathbb{P}_{\theta,x_0}(\varepsilon \leq e)$ has density ξ_{θ,x_0}. It is natural to consider this quantity as an *augmented* likelihood for x_0 under the current model,

$$\theta, \varepsilon \rightarrow f_\rho(x_0|\theta, \varepsilon, M_i) := \xi_{\theta,x_0}(\varepsilon). \quad (4)$$

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 $\pi(\theta, \varepsilon|M_i)$, we embrace two aspects of statistical reasoning, parameter inference and model criticism, *simultaneously* by the joint posterior density

$$f_\rho(\theta, \varepsilon|x_0, M_i) = \xi_{\theta,x_0}(\varepsilon)\pi(\theta, \varepsilon|M_i) \;/\; f_\rho(x_0|M_i). \quad (5)$$

In practice, we adopt the prior on θ independently of that of ε. The posterior relationship Eq. 5 exploits the dependence between model error and model parameterization.

2.2. Parameter inference and model criticism

Non-zero 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 zero for some value of θ. Parameter inference based on the *marginal* posterior distribution $f_\rho(\theta|x_0, M_i)$ can be justified in an "approximate likelihood" sense, because

$$f_\rho(\theta|x_0, M_i) \propto \pi(\theta|M_i) \int_{\mathcal{X}} \pi(\varepsilon_x|M_i) f(x|\theta, M_i)dx, \quad (6)$$

where $\varepsilon_x = \rho(s_x, s_0)$. Setting $\pi(\varepsilon|M_i) = \mathbf{1}\{|\varepsilon| \leq \tau/2\}/\tau$, we recover the "standard" ABC approximation Eq. 2. We can interpret the variety of ABC kernels as exerting a particular prior belief on the adequacy of the current model [15]. We always choose a prior $\pi(\varepsilon|M_i)$ with mode at zero to accommodate a prior belief that the model is plausible.

For the purpose model criticism, we also focus on the *marginal* posterior error distribution $f_\rho(\varepsilon|x_0, M_i)$. Noting that the prior predictive error density fulfills

$$L_\rho(\varepsilon) = \int_{\{\varepsilon_x = \varepsilon\}} \pi(\theta|M_i) f(x|\theta, M_i)dx, \quad (7)$$

we find that

$$f_\rho(\varepsilon|x_0, M_i) = L_\rho(\varepsilon)\pi(\varepsilon|M_i) \;/\; f_\rho(x_0|M_i). \quad (8)$$

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. Model criticism based on Eq. 8 rather than $L_\rho(\varepsilon)$ is appealing from the perspective of Eq. 5, as it focuses on those θ actually inferred *and* attenuates the dependence of $L_\rho(\varepsilon)$ on $\pi(\theta|M_i)$.

2.3. Algorithm

We propose to replace Eq. 4 with a kernel density estimate

$$\hat{\xi}(\varepsilon; \mathbf{x}) := \frac{1}{Bh} \sum_{b=1}^{B} K\big([\varepsilon - \rho(s_b, s_0)]/h\big), \quad (9)$$

(here suppressing the dependence of $\hat{\xi}$ on x_0) where K is a kernel, $\mathbf{x} = (x_1, \ldots, x_B)$ and $s_b = \mathbb{S}(x_b)$, obtaining a *smoothed* approximation to the acceptance probability $\xi_{\theta,x_0}(\varepsilon)$. Suppose an initial sample (θ, ε) and prior specifications;

1) If now at θ, move to θ' according to $q(\theta \rightarrow \theta')$
2) Generate $\mathbf{x}' \sim f(\cdot|\theta', M_i)$, and construct $\hat{\xi}(\,\cdot\,; \mathbf{x}')$
3) If now at ε, move to ε' according to $q(\varepsilon \rightarrow \varepsilon')$. We guide this proposal with $\hat{\xi}_k(\,\cdot\,; \mathbf{x})$ and $\hat{\xi}_k(\,\cdot\,; \mathbf{x}')$
3) Accept $(\theta', \boldsymbol{\varepsilon}', \mathbf{x}')$ with probability

$$\min\left\{1\,,\, \frac{\pi(\theta', \varepsilon'|M_i)q(\theta' \rightarrow \theta)q(\varepsilon' \rightarrow \varepsilon)}{\pi(\theta, \varepsilon|M_i)q(\theta \rightarrow \theta')q(\varepsilon \rightarrow \varepsilon')} \frac{\hat{\xi}(\varepsilon'; \mathbf{x}')}{\hat{\xi}(\varepsilon; \mathbf{x})}\right\},$$

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

The ABC sampler operates on the augmented space $(\theta, \varepsilon, \mathbf{x})$ and the posterior density Eq. 5 is obtained by marginalization with respect to (θ, ε). Details of how we choose $q(\theta \rightarrow \theta')$ and $q(\varepsilon \rightarrow \varepsilon')$ can be found in [9].

2.4. Multi-dimenisonal discrepancies and error terms

The complexity of the settings to which ABC is typically applied, makes it difficult to think of a universal scalar error term. However, approximating ξ_{θ,x_0} by a kernel estimate is difficult to do reliably for higher dimensional error terms unless B is very large. Alternatively, one might put $\hat{\xi}(\boldsymbol{\varepsilon}; \mathbf{x}) = \prod_k \hat{\xi}_k(\varepsilon_k; \mathbf{x})$ (assuming independence) or $\hat{\xi}(\boldsymbol{\varepsilon}; \mathbf{x}) = \min_k \hat{\xi}_k(\varepsilon_k; \mathbf{x})$, which takes into account dependencies between summaries. We use the latter in our example below.

Our approach to model criticism capitalizes on the fact that the co-dependencies among $S_k(x)$ under the predictive distribution are typically different from those among $S_k(x_0)$ if the model is not adequate, revealing *model inconsistency* in terms of conflicting, co-dependent summaries. For model criticism in situations where the likelihood cannot be evaluated, it is therefore essential that a large set of summaries is considered, see [9].

3. CRITICIZING MODELS OF PIN EVOLUTION

The structure of PINs derives from multiple stochastic processes over evolutionary time-scales, and various models, based on randomly growing graphs, have been proposed to capture aspects of network growth [16] (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 about 80% of all

Table 1. 50% confidence intervals of $\varepsilon_k|x_0$, indicating model mismatch relative to the *T. pallidum* PIN†

M_i	ε_{WR}	$\varepsilon_{\text{ODBOX}}$	ε_{DIA}	ε_{CC}	$\varepsilon_{\overline{\text{ND}}}$	$\varepsilon_{\text{FRAG}}$
PAP (S1)‡	[-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	[-0.12,0.16]	**[0.02,0.61]**	[-0.60,0.91]	**[-0.010,-0.003]**	[-0.66,0.01]	[-0.16,0.01]
DDA+PA	[-0.34,0.48]	[-0.12,0.20]	[-0.48,0.50]	[-0.002,0.027]	[-0.66,0.05]	[-0.01,0.29]
DD+LNK+PA	[-0.20,0.17]	[-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.
‡ PAP (S1) uses sampling scheme S1; all others use S2.

genes in a prokaryotic genome are involved in lateral gene transfer [17]; model PAP (details below) is inspired by this scenario. At least 40% of genes in prokaryotes appear to be products of gene duplication [18]. Model DDA+PA is designed to quantify the potential role of duplication-divergence in network evolution [3]. At least for eukaryotes, the formation or degeneration of functional links between proteins (link turnover) is estimated to occur at a fast rate of ca 10^{-5} changed interactions per My per protein pair [12]. Model DD+LNK+PA includes link turnover in terms of preferential loss and gain of interactions.

3.1. Specification of models and sampling schemes

We consider the following models.

PAP. Evolution proceeds only by **p**referential **a**ttachment[19]; at each step the number of attachments minus one is **P**oisson distributed with mean m.

DDA+PA. Preferential attachment (PA) of a new node to one node of the existing network with probability α, or, with probability $1-\alpha$, node **d**uplication with immediate link **d**ivergence. A parent node is randomly chosen, its edges are duplicated and the parental and new edges 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 **a**ttached to its child with probability δ_{Att}.

DD+LNK+PA. A mixture of DDA with $\delta_{\text{Att}} = 0$ (DD), **l**i**nk** addition and deletion, and PA. Link addition (deletion) proceeds by choosing a node randomly, and attaching it preferentially to another node (deleting it preferentially from its interaction partners) [12]. 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}. PA 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.

A network is grown to the order of the (estimated) number of protein coding genes in the genome; subsequently nodes are sampled corresponding to the number of proteins in the PIN. Sampling is done in either two ways; by randomly sampling proteins from the simulated data (S1) or by randomly sampling among those proteins that have an interaction in the simulated data (S2).

3.2. Summary statistics

PINs can be described as graphs which contain a set of nodes, interacting proteins, and undirected binary edges, representing the observed interactions between the proteins. We consider the following summary statistics. 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; WR, within-reach distribution, the mean probability of how many nodes are reached from one node within distance $k = 1, 2, \ldots$ [3]; DIA, diameter; CC, cluster coefficient; 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; ODBOX, BOX degree distribution, the probability distribution of BOXes with k edges to nodes outside the BOX.

3.3. Analysis of two PIN data sets

We successfully applied the algorithm to sample from the posterior $\hat{f}_\rho(\theta, \boldsymbol{\varepsilon}|x_0, \cdot)$ for the models PAP, DDA+PA, and DD+LNK+PA and sampling schemes S1 and S2. Based on S1, all models depart significantly in FRAG as exemplified for PAP in Table 1. 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 S2; see Table 1. Turning to the five remaining summaries, we observe, that only model DDA+PA matches the *T. 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 S2 for the *H. pylori* PIN dataset, we could not substantiate our results further, as all considered models provide an adequate fit to this data. This is surprising, because we expect a similar power of our method on both datasets and may point to qualitative differences among the two PINs owing to different, underlying experimental protocols.

Enumerating subgraph counts may provide deeper insights into the dynamics of network evolution [20]. The fact that the box degree distribution ODBOX of the observed PIN dataset is not reproduced with model PAP under several sampling schemes shows that in abstract terms, network evolution cannot be understood as a process of repeated preferential attachment alone.

Under the assumption that each protein had the same chance of being identified as an interaction partner S1, we

expect no mismatch in terms of the fragmentation of the PIN dataset. The significant departures in terms of FRAG under S1 may thus reflect sampling bias [14], which can be alleviated by simply accounting for those proteins that are tested but not reported (S2), see Table 1. In line with other researchers [21], we suggest that a deeper appreciation of the systematic and stochastic measurement errors in high-throughput experiments and the limitations of the reported data is key to making better use of existing data.

4. ACKNOWLEDGMENTS

OR gratefully accepts funding from the Wellcome Trust; CA an Advanced Research Fellowship from the EPSRC; CW from the Danish Research Councils; and SR from the BBSRC and the Centre for Integrative Systems Biology at Imperial College.

5. REFERENCES

[1] G. E. P. Box, "Science and statistics," *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.

[2] J. M. Bernado and A. F. M. Smith, *Bayesian Theory*, Wiley & Sons, Chichester, UK, 1st edition, 1994.

[3] O. Ratmann, O. Jørgensen, T. Hinkley, M. P. Stumpf, S. Richardson, and C. Wiuf, "Using likelihood-free inference to compare evolutionary dynamics of the protein networks of H.pylori and P.falciparum," *PLoS Computational Biology*, vol. 3, no. 2007, pp. e230, 11 2007.

[4] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf, "Approximate bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of The Royal Society Interface*, 2008.

[5] P. Marjoram and S. Tavaré, "Modern computational approaches for analysing molecular genetic variation data," *Nat Rev Genet*, vol. 7, no. 10, pp. 759–770, 2006.

[6] N. J. R. Fagundes, N. Ray, M. Beaumont, S. Neuenschwander, F. M. Salzano, S. L. Bonatto, and L. Excoffier, "Statistical evaluation of alternative models of human evolution," *Proceedings of the National Academy of Sciences*, vol. 104, no. 45, pp. 17614–17619, 2007.

[7] A. Zellner, "Bayesian analysis of regression error terms," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 138–144, 1975.

[8] A. O'Hagan, "HSSS model criticism (with discussion)," In *Highly Structured Stochastic Systems,* Oxford University Press, pp. 423–453, 2003.

[9] O. Ratmann, C. Andrieu, C. Wiuf, and S. Richardson, "Model Criticism based on likelihood-free inference, with an application to protein network evolution," *submitted*, 2009.

[10] J.-C. Rain, L. Selig, H. De Reuse, V. Battaglia, C. Reverdy, S. Simon, G. Lenzen, F. Petel, J. Wojcik, V. Schachter, Y. Chemama, A. Labigne, and P. Legrain, "The protein-protein interaction map of Helicobacter pylori," *Nature*, vol. 409, pp. 211–215, 2001.

[11] B. Titz, S. V. Rajagopala, J. Goll, R. Häuser, M. T. McKevitt, T. Palzkill, and P. Uetz, "The binary protein interactome of Treponema pallidum – the Syphilis spirochete," *PLoS ONE*, vol. 3, no. 5, pp. e2292, 2008.

[12] P. Beltrao and L. Serrano, "Specificity and evolvability in eukaryotic protein interaction networks.," *PLoS Comput Biol*, vol. 3, no. 2, pp. e25, 2007.

[13] J. W. Pinney, G. D. Amoutzias, M. Rattray, and D. L. Robertson, "Reconstruction of ancestral protein interaction networks for the bZIP transcription factors.," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 20449–20453, 2007.

[14] L. Hakes, J. W. Pinney, D. L. Robertson, and S. C. Lovell, "Protein-protein interaction networks and biology - what's the connection?," *Nat Biotech*, vol. 26, no. 1, pp. 69–72, 2008.

[15] R. D. Wilkinson, "Approximate Bayesian computation (ABC) gives exact results under the assumption of model error," *arXiv.org*, 2008.

[16] M. Knudsen and C. Wiuf, "A Markov chain approach to randomly grown graphs," *Journal of Applied Mathematics*, p. 190836, 2008.

[17] T. Dagan, Y. Artzy-Randrup, and W. Martin, "Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution," *Proceedings of the National Academy of Sciences*, vol. 105, no. 29, pp. 10039–10044, 2008.

[18] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, "Evolution of the Protein Repertoire," *Science*, vol. 300, no. 5626, pp. 1701–1703, 2003.

[19] A. L. Barabási and R. Albert, "Emergence of scaling in random networks.," *Science*, vol. 286, pp. 509–512, 1999.

[20] J. Rice, A. Kershenbaum, and G. Stolovitzky, "Lasting impressions: Motifs in protein-protein maps may provide footprints of evolutionary events," *Proc Natl Acad Sci U S A.*, vol. 102, no. 9, pp. 3173–3174, 2005.

[21] R. Gentleman and W. Huber, "Making the most of high-throughput protein-interaction data," *Genome Biology*, vol. 8, no. 10, pp. 112, 2007.