

# Approximate Bayesian inference for complex ecosystems

Michael P.H. Stumpf

Address: Centre for Integrative Systems Biology and Bioinformatics, Department of Life Sciences, Sir Ernst Chain Building, Imperial College London, London, SW7 2AZ, UK

Email: m.stumpf@imperial.ac.uk

F1000Prime Reports 2014, 6:60 (doi:10.12703/P6-60)

All F1000Prime Reports articles are distributed under the terms of the Creative Commons Attribution-Non Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/legalcode>), which permits non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

The electronic version of this article is the complete one and can be found at: <http://f1000.com/prime/reports/b/6/60>

## Abstract

Mathematical models have been central to ecology for nearly a century. Simple models of population dynamics have allowed us to understand fundamental aspects underlying the dynamics and stability of ecological systems. What has remained a challenge, however, is to meaningfully interpret experimental or observational data in light of mathematical models. Here, we review recent developments, notably in the growing field of approximate Bayesian computation (ABC), that allow us to calibrate mathematical models against available data. Estimating the population demographic parameters from data remains a formidable statistical challenge. Here, we attempt to give a flavor and overview of ABC and its applications in population biology and ecology and eschew a detailed technical discussion in favor of a general discussion of the advantages and potential pitfalls this framework offers to population biologists.

## Introduction

Theoretical population biology has been crucial for our understanding of ecosystems [1]. Mathematical models can explain elegantly what might appear as bewilderingly complex variations in species abundances. Seminal work starting in the early 20<sup>th</sup> century [2-4] has, in fact, become so familiar to population biologists and beyond that today we are hardly surprised to see complex oscillatory patterns or complex dependencies of population dynamics on a myriad of environmental and demographic factors [5]. Many of these phenomena can straightforwardly be explained in terms of relatively simple population dynamics models. The success of these models has also meant that ecological ideas are coming to pervade the analysis of other interacting systems, including cancer [6], stem cells [7,8], and even the banking system [9,10], all of which are characterized by the interactions between different entities that affect the overall dynamics of the system and its stability.

Simple models are beguiling and shape our intuition and allow us to explain trends in data. In many important

scenarios, however, different factors come together with sometimes complex patterns resulting from their interplay. Thus, understanding realistic systems—subject to a multitude of internal and external factors—is hard [11,12]. This is further complicated in situations where models are used to make predictions or assess different types of interventions *in silico* prior to their implementation in, for example, conservation biology.

These challenges are not unique to theoretical ecology, of course, and recent years have seen concerted efforts to tackle the so-called inverse problem: estimating parameters of a model from data [13]; choosing from among a set of plausible candidate models the model that is best able to explain the data [14]; or inferring mechanistic or statistical dependencies between the different state variables making up a system—in an ecological case, this would, for example, be the species considered in the model. Below, we are considering population dynamical models where a vector,  $\mathbf{x}$ , containing  $N$  species,  $x_1, x_2, \dots, x_N$  describes the abundances of species in the ecosystem. These are assumed to change as a result of

interactions among the species (and potentially external factors) according to some rate laws,

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}; \Theta), \quad (1)$$

where, with slight abuse of notations, we will also implicitly allow for stochastic dynamics. The community matrix of the ecological system (1) is, of course, given by

$$A = \frac{\partial \mathbf{f}(\mathbf{x}; \Theta)}{\partial \mathbf{x}},$$

which captures the ecological relationships among the species. Finally, the (vector-valued) parameter  $\Theta$  denotes the typically unknown demographic and system parameters (for example, birth, death, and migration rates) as well as parameters characterizing the interactions between and within species.

Below, we will discuss methods that allow us to infer the parameters,  $\Theta$ , and choose between different potential models (for example,  $f_1, f_2, \dots, f_K$ ). The statistical toolset that we will discuss, centered primarily around ABC [15,16], complements traditional mathematical approaches that have been used in theoretical population biology to great effect since the 1950s. But the aim here is—rather than to focus on general mathematical laws governing the behavior and fate of natural populations—to make models as specific to a given problem, to identify the key factors driving an ecosystem's dynamics, or to make predictions about the future of an ecosystem.

There are well-defined statistical frameworks to deal with parameter inference. Model selection—the process of comparing the ability of different models to explain some data—is continuing to attract the attention of statisticians and domain experts in different scientific disciplines [14]. But for many challenging real-world problems, conventional statistical approaches become computationally too cumbersome very quickly. This class of problems includes many stochastic processes, highly structured populations, and those where different types of data need to be considered. Often, it is still straightforward to establish simulation models—in general, real-world problems tend to defy purely analytical approaches—but conventional statistical approaches become computationally too expensive.

Arguably, many of the most contentious problems in population biology (or science in general) fall into this category of problems. A model abstracts from reality what are known or believed to be the essential features of

a real natural (or technological or social) system. This fact alone has in the past added to some controversies: as “all models are wrong” [17], it is necessary to identify the best model that captures and allows us to quantitatively and qualitatively understand the dynamics of the real system. Thus, we need statistical tools that allow us to deal with complex systems, many of which are expected to stretch conventional statistics. Here, we develop a viable alternative that maintains most if not all of the advantages of the Bayesian inferential apparatus but can be extended to problems defying conventional statistics.

### Model calibration and parameter estimation

Given a model,  $\mathbf{f}(\mathbf{x}; \Theta)$ , and some data,  $D = \{d_1, \dots, d_n\}$ , we need to infer the parameters,  $\Theta$ , from the data. The likelihood [18] is defined as the probability of obtaining some data  $D$  given a parameter value  $\Theta'$ ,

$$L(\Theta') = Pr(D|\Theta') \quad (2)$$

This is the central quantity in likelihood inference; crucially, the likelihood contains all the information about the parameter that can be extracted from the data  $D$ . In Bayesian inference [19], it, together with the prior distribution of  $\Theta$ ,  $Pr(\Theta)$ , strikes a balance between what is or can be known about the parameter prior to having seen the data, and the information contained in the data, to give rise to the posterior distribution,

$$Pr(\Theta|D) = \frac{Pr(D|\Theta)Pr(\Theta)}{Pr(D)}. \quad (3)$$

Here,  $Pr(D)$  denotes the evidence. It is often thought of as a normalization constant but does in fact contain information about the ability of a model to describe the data.

Obtaining the posterior distribution, or a sample from it, is computationally demanding. In general, computing the evidence  $Pr(D)$ , which is typically a high-dimensional integral, is complicated. Sometimes, the focus, therefore, may shift from consideration of the whole (posterior) distribution to the maximum (mode) of the posterior distribution; this maximum *a posteriori* estimate is the Bayesian equivalent to the maximum likelihood estimate.

So that the additional information contained in the distribution can be obtained, a wealth of computational statistical approaches have been developed. Markov chain Monte Carlo (MCMC) methods have become the main workhorses of computational Bayesian statistics

and have allowed us to generate samples from the posterior distribution. Recent years have witnessed increased interest in these and related methods—such as population and sequential Monte Carlo techniques—but even the most sophisticated approaches reach their limits when the number of parameters or the complexity of the model increases. The first problem, the so-called curse of dimensionality, is shared by all statistical inference procedures.

The second problem is more interesting. For example, we may ask ourselves whether there are simpler versions,  $M_a(\Theta_a)$ , of the model that we are considering,  $M_0(\Theta)$ , that would, despite the simplification or coarse-graining (by simplification we typically mean that the dimension of the parameter vector is smaller in the simplified model, that is,  $|\Theta_a| < |\Theta|$ ), allow us to draw meaningful, verifiable (or falsifiable) mechanistic insights from the available data. In principle, this might appear to exacerbate the statistical problem, for we would have to find computationally affordable and sufficiently discriminatory ways of deciding if and when a simpler model  $M_a(\Theta_a)$  is a good approximation to the original model,  $M_0(\Theta)$ . We will return to this point again below. First, however, we discuss ABC methods, which form an alternative approach to tackling statistically challenging problems in a Bayesian framework and which have become a popular alternative to conventional (or exact) Bayesian inference in many applications, especially in evolutionary, population, and systems biology.

### Approximate Bayesian computation

In ABC, we stay as close as possible [16] to the model of interest but instead forgo evaluation of the likelihood in favor of a comparison between simulated and real data [15,20,21]. For many systems, the likelihood becomes computationally intractable, either because of the model complexity or the detailed nature of the data. Nevertheless, the underlying model can still be simulated. The principal underlying insight of ABC is that we can consider

$$Pr(D|\Theta) = \lim_{\varepsilon \rightarrow 0} Pr(\Delta[D, D_\Theta] \leq \varepsilon | \Theta), \quad (4)$$

where  $D_\Theta$  is data obtained by simulating from our model with parameter  $\Theta$ ,  $\Delta[x, x']$  is a distance function that can be chosen flexibly to suit the problem at hand, and  $\varepsilon$  is a tolerance threshold that reflects the desired accuracy of our inference. The essential problem is that for any complicated problem, it is impossible to obtain the precise dataset,  $D$ , by simulating from the model  $D_\Theta$ , even if we know the true parameter (we ignore the artificial problem

of deterministic dynamics with no observational noise). By increasing the threshold  $\varepsilon$ , our inference becomes more approximate, but the chance of obtaining a simulated dataset for which  $\Delta[D, D_\Theta] < \varepsilon$  increases.

The comparison of real and simulated data is particularly straightforward for ecological time-series data, for example [22]. Here,  $D$  might take the form of vectors of population abundances,  $x_t$ , for  $n$  species collected at  $t=1, \dots, m$  time-points. In this case, the Euclidean (or any other vector) norm provides a suitable distance. The analysis of dynamical systems is thus relatively straightforward in an ABC framework. Generalization to compartmental or spatio-temporal models (or both) is straightforward [23]: if we can simulate data efficiently, we can appeal to the Bayesian inference formalism via ABC (keeping in mind the nature of the approximation and the tolerance threshold  $\varepsilon$ ).

Instead of comparing the data, we can compare aspects of the data, such as summary statistics. This has been one of the main advantages as well as sources of contention for ABC inference. We call a statistic,  $s$  of the data, *sufficient* if and only if

$$Pr(D|\Theta, s) = Pr(D|s).$$

In this case, we can replace the data by the sufficient statistic without any loss of information about the parameter,  $\Theta$ . The attraction of using sufficient summary statistics lies in the fact that their dimension,  $d_s$ , is typically much smaller than the dimension of the data itself,  $d_D$  (in the above example of  $n$  species sampled at  $m$  time points,  $d_D = m \times n$ ); that is,  $d_s \ll d_D$ . Especially in population genetics, which has inspired the rise of ABC methods since the late 1990s, the use of summary statistics has been popular (see, for example, [24-28]). With the use of summary statistics, the likelihood can be written as

$$Pr(D|\Theta) = \lim_{\varepsilon \rightarrow 0} Pr(\Delta[s(D), s(D_\Theta)] \leq \varepsilon | \Theta), \quad (5)$$

with potentially an appropriate change in the distance function and  $\varepsilon$ .

Although Equation 5 works very well for parameter inference if  $s$  is sufficient, it is important to note that sufficient statistics are few and far between for any real-world problem. Unfortunately, ABC requires appropriate sufficient statistics (or comparisons of the data directly, as in the case of time-series problems). There have been attempts to generate collections of statistics that together fulfil sufficiency properties [28-33], but these are computationally expensive in their own right.

So far, we have implicitly considered ABC in a simple rejection framework: (a) we sample a parameter from a suitable prior, (b) we simulate the model for the parameter, and then (c) we compare the simulated and the real data (or their respective summary) statistics and accept a parameter as a draw from the ABC posterior if the distance is below some threshold. Steps (a) to (c) are repeated until a sufficiently large number of parameter values have been accepted. The posterior in this case is represented as a sum over indicator functions,  $Pr_{ABC}(\Theta|D) = \sum_{i=1}^N 1(\Theta_i)$ , where either  $\Delta[D, D-\Theta_i] \leq \varepsilon$  or  $\Delta[s(D), s(D_{\Theta_i})] \leq \varepsilon$ , depending on whether the data or sufficient summaries are used in the inference process.

This framework is as simple as it is impractical: like all rejection samplers, it is limited to small problems involving less than a handful of parameters. It has been possible to construct ABC-MCMC samplers [21], but the real workhorses of most ABC approaches to real-world problems are based on sequential Monte Carlo (SMC) approaches [22,34]; ABC-SMC has become a very popular field of research (arguably inspiring more detailed analysis also in exact SMC samplers), and recent developments are allowing us to tackle larger and more complicated systems [35]. The most widely used flavor of ABC-SMC proceeds by constructing a set of intermediate distributions that start from the prior and increasingly resemble the posterior. To do so, a sequence of decreasing thresholds,  $\varepsilon_1 \geq \varepsilon_2 \geq \dots \geq \varepsilon_K$  (with  $\varepsilon_K = \varepsilon$ ), is defined and the sequence of distributions is constructed by sampling parameter vectors from the previous distribution (or the prior in the first step), perturbing them by using some perturbation kernel function, and accepting those parameter vectors for which the distance between real and simulated data falls below the threshold  $\varepsilon_k$ . Choice of the thresholds and the nature of the perturbation kernels determine the computational efficiency and runtime of the inference, but both can be tuned to speed up the process and tackle larger problems [36,37].

### Model selection and checking

So far, we have assumed that we have a single model that describes our system of interest. The Bayesian framework readily provides us with credible intervals for parameters, but it is also possible to assign probabilities for different models to be the correct model, conditional on the available data and the set of competing models,  $M_i \in \mathcal{M} = \{M_1, \dots, M_k\}$ . In the likelihood framework, the comparison of general (that is, non-nested) models is made possible only through the use of information criteria; in the Bayesian framework, the posterior

probability of a model is given analogously to Equation 3 by [14]

$$Pr(M_i|D) = \frac{Pr(D|M_i)Pr(M_i)}{Pr(D)}. \quad (6)$$

which is also known as the marginal likelihood of model  $M_i$ . In principle, Bayesian model selection allows us to compare any number of arbitrary models. An additional advantage is that the selection via the marginal likelihood, Equation 6, automatically strikes a balance between the ability of models to reproduce or explain the observed data, the complexity of the model, and the robustness of the inference.

Equation 6 can be interpreted in the ABC framework [22,38,39], and ABC model selection has been an area of great interest and activity [40-45]. Although model selection is indeed straightforward if experimental and simulated data are compared directly, it has been shown that model selection becomes unreliable when summary statistics instead of the data are compared [46,47]: summary statistics are sufficient for model selection for only a very restricted set of problems. Constructing sets of statistics that are sufficient for model selection (they must be sufficient for every model considered and across the models; this is an area of active research [48,49]), while possible in principle, is computationally enormously demanding.

In many ecological problems, however, we deal with spatio-temporal time-series data, for which model selection is possible. Our aim in such cases is typically to identify the most promising mechanistic descriptions of a complex system. If no single model emerges from such a comparison, then we need to investigate those models that have comparably high marginal likelihoods. Simulations from the respective model posteriors can then be used, for example, to develop more discriminatory experimental designs that allow us to further distinguish among these models [50]. This, too, is an area of continuing importance for ABC.

### Applicability of approximate Bayesian computation: an outlook

ABC methods were borne out of a need to tackle problems that defy conventional statistical methodologies. It has become clear, however, that whenever suitable Bayesian alternatives that do deal with the proper likelihood are available, ABC becomes computationally too expensive. The reason for this is primarily the fact that the representation of the posterior (as a weighted sum over

Dirac  $\delta$  functions) is not very efficient. So when alternatives are available, they ought to be used. In parallel to their role in computationally demanding applications, ABC techniques have, more recently, also attracted attention as an inferential framework in their own right [16,51]. From this, interesting new approaches to deal with real-world problems may well emerge [52].

In conclusion, ABC-based methods are best suited to those problems for which other likelihood-based (or exact) Bayesian inference procedures do not yet exist. This appears to still include a host of challenging and interesting problems. Many stochastic and highly structured spatio-temporal problems in ecology, epidemiology, and evolutionary genetics clearly fall into this category. The recent developments discussed above mean that ABC has become a viable new way of tackling computationally demanding parameter inference problems. Given a model—as long as we can simulate it—ABC gives us a handle to evaluate approximate posterior distributions, which then can be further evaluated. Sensitivity and robustness analyses, but also predictions of future behavior or the likely effects of any interventions or perturbations, can be analyzed by simulating the model with parameters sampled from the posterior. There is enormous scope for basing the exploration of, for example, policy or conservation measures on the available data in this way. ABC has, for example, been used in experimental design [50,53] and in synthetic biology [14,54] to generate designs of molecular pathways that exhibit certain types of behavior. In such cases, we replace the observed data,  $D$ , by a representation of the desired behavior (such as the desired abundance of a species). Then the inference procedure is used to identify the scenario for which we are most likely to observe this outcome. Such predictions then reflect the best available evidence in light of the data and the model.

As an aside, it is worth keeping in mind that the technical challenges of statistical inference and modeling can often be minor compared with the difficulties in communicating the results to policy makers or the general public. Many of the most pressing problems in ecology have become highly emotive topics as they nearly always involve a conflict between parties that have very different priorities (see, for example, [45,55]). In many complicated situations, the nuance and cautiousness that accompany how we present such analyses could be taken for wavering or lack of reliability. Here, however, ABC, with its explicit focus on simulation, may even have an advantage, as the underlying rationale is so straightforwardly explained and easy to understand.



## Abbreviations

ABC, approximate Bayesian computation; MCMC, Markov chain Monte Carlo; SMC, sequential Monte Carlo.

## Disclosures

The author declares that he has no disclosures.

## References

1. May RM: *Stability and complexity in model ecosystems*. 1st edition. Princeton: Princeton University Press; 1973.
  2. Lotka AJ: **Analytical Note on Certain Rhythmic Relations in Organic Systems**. *Proc Natl Acad Sci USA* 1920, **6**:410-5.
  3. Volterra V: **Fluctuations in the abundance of a species considered mathematically**. *Nature* 1926, **118**:558-60, 1926.
  4. Wilson EO, MacArthur RH: *The Theory of Island Biogeography*. Princeton: Princeton University Press; 1967.
  5. Pascual M: **Computational ecology: from the complex to the simple and back**. *PLoS Comput Biol* 2005, **1**:101-5.
  6. Siegmund KD, Marjoram P, Shibata D: **Modeling DNA methylation in a population of cancer cells**. *Stat Appl Genet Mol Biol* 2008, **7**:Article 18.
  7. Powell K: **Stem-cell niches: it's the ecology, stupid!**. *Nature* 2005, **435**:268-70.
  8. MacLean AL, Lo Celso C, Stumpf MPH: **Population dynamics of normal and leukaemia stem cells in the haematopoietic stem cell niche show distinct regimes where leukaemia will be controlled**. *J R Soc Interface* 2013, **10**:20120968.
  9. May RM, Arinaminpathy N: **Systemic risk: the dynamics of model banking systems**. *J R Soc Interface* 2010, **7**:823-38.
  10. Haldane AG, May RM: **Systemic risk in banking ecosystems**. *Nature* 2011, **469**:351-5.
- 
11. Pimm SL: **The complexity and stability of ecosystems**. *Nature* 1984, **307**:321-26.
  12. Jansen VAA, Kokkoris GD: **Complexity and stability revisited**. *Ecology Letters* 2003, **6**:498-502.
  13. Moles CG, Mendes P, Banga JR: **Parameter estimation in biochemical pathways: a comparison of global optimization methods**. *Genome Res* 2003, **13**:2467-74.
  14. Kirk P, Thorne T, Stumpf Michael P H: **Model selection in systems and synthetic biology**. *Curr Opin Biotechnol* 2013, **24**:767-74.
  15. Csilléry K, Blum, Michael G B, Gaggiotti OE, François O: **Approximate Bayesian Computation (ABC) in practice**. *Trends Ecol. Evol. (Amst.)* 2010, **25**:410-8.
  16. Wilkinson RD: **Approximate Bayesian computation (ABC) gives exact results under the assumption of model error**. *Stat Appl Genet Mol Biol* 2013, **12**:129-41.
- 
17. Box G: **Sampling and Bayes' inference in scientific modelling and robustness**. *J R Stat Soc Series B Stat Methodol* 1980 **143**:383-430.
  18. Cox DR: *Principles of Statistical Inference*. Cambridge: Cambridge University Press; 2006.
  19. Gelman A, Carlin JB, Stern HS, Rubin DB: *Bayesian Data Analysis*. London: Chapman & Hall; 2003.
  20. Beaumont MA, Zhang WW, Balding DJ: **Approximate Bayesian computation in population genetics**. *Genetics* 2002, **162**:2025-35.
  21. Marjoram P, Molitor J, Plagnol V, Tavaré S: **Markov chain Monte Carlo without likelihoods**. *Proc Natl Acad Sci USA* 2003, **100**:15324-8.

22. Toni T, Welch D, Strelkova N, Ipsen A, Stumpf Michael P H: **Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems.** *J R Soc Interface* 2009, **6**:187-202.
23. Liepe J, Taylor H, Barnes CP, Huvet M, Bugeon L, Thorne T, Lamb JR, Dallman MJ, Stumpf Michael P H: **Calibrating spatio-temporal models of leukocyte dynamics against in vivo live-imaging data using approximate Bayesian computation.** *Integr Biol (Camb)* 2012, **4**:335-45.
24. Chan YL, Anderson, Christian N K, Hadly EA: **Bayesian estimation of the timing and severity of a population bottleneck from ancient DNA.** *PLoS Genet* 2006, **2**:e59.
25. Hickerson MJ, Stahl E, Takebayashi N: **msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation.** *BMC Bioinformatics* 2007, **8**:268.
26. Wegmann D, Leuenberger C, Excoffier L: **Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood.** *Genetics* 2009, **182**:1207-18.
27. Rosvold J, Røed KH, Hufthammer AK, Andersen R, Stenøien HK: **Reconstructing the history of a fragmented and heavily exploited red deer population using ancient and contemporary DNA.** *BMC Evol Biol* 2012, **12**:191.
28. Soubeyrand S, Carpentier F, Guiton F, Klein EK: **Approximate Bayesian computation with functional statistics.** *Stat Appl Genet Mol Biol* 2013, **12**:17-37.
29. Nunes MA, Balding DJ: **On optimal selection of summary statistics for approximate Bayesian computation.** *Stat Appl Genet Mol Biol* 2010, **9**:Article34.
30. Jung H, Marjoram P: **Choice of summary statistic weights in approximate Bayesian computation.** *Stat Appl Genet Mol Biol* 2011, **10**.
31. Fearnhead P, Prangle D: **Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation.** *J R Stat Soc Series B Stat Methodol* 2012, **74**:419-74.
32. Aeschbacher S, Beaumont MA, Futschik A: **A novel approach for choosing summary statistics in approximate Bayesian computation.** *Genetics* 2012, **192**:1027-47.
33. Nakagome S, Fukumizu K, Mano S: **Kernel approximate Bayesian computation in population genetic inferences.** *Stat Appl Genet Mol Biol* 2013, **12**:667-78.
34. Sisson SA, Fan Y, Tanaka MM: **Sequential Monte Carlo without likelihoods.** *Proc Natl Acad Sci USA* 2007, **104**:1760-5.
35. Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf Michael P H: **A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation.** *Nat Protoc* 2014, **9**:439-56.
36. Silk D, Filippi S, Stumpf Michael P H: **Optimizing threshold-schedules for sequential approximate Bayesian computation: applications to molecular systems.** *Stat Appl Genet Mol Biol* 2013, **12**:603-18.
37. Filippi S, Barnes CP, Cornebise J, Stumpf Michael P H: **On optimality of kernels for approximate Bayesian computation using sequential Monte Carlo.** *Stat Appl Genet Mol Biol* 2013, **12**:87-107.
38. Toni T, Stumpf Michael P H: **Simulation-based model selection for dynamical systems in systems and population biology.** *Bioinformatics* 2010, **26**:104-10.
39. Leuenberger C, Wegmann D: **Bayesian computation and model selection without likelihoods.** *Genetics* 2010, **184**:243-52.
40. Peter BM, Wegmann D, Excoffier L: **Distinguishing between population bottleneck and population subdivision by a Bayesian model choice procedure.** *Mol Ecol* 2010, **19**:4648-60.
41. Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, Suerbaum S, Achtman M: **Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families.** *PLoS Genet* 2010, **6**:e1001036.
42. Estoup A, Lombaert E, Marin J, Guillemaud T, Pudlo P, Robert CP, Cornuet J: **Estimation of demo-genetic model probabilities with Approximate Bayesian Computation using linear discriminant analysis on summary statistics.** *Mol Ecol Resour* 2012, **12**:846-55.
43. Thorne T, Stumpf Michael P H: **Graph spectral analysis of protein interaction network evolution.** *J R Soc Interface* 2012, **9**:2653-66.
44. Ratmann O, Donker G, Meijer A, Fraser C, Koelle K: **Phylogenetic inference and model assessment with approximate Bayesian computation: influenza as a case study.** *PLoS Comput Biol* 2012, **8**:e1002835.
45. Mardulyn P, Goffredo M, Conte A, Hendrickx G, Meiswinkel R, Balenghien T, Sghaier S, Lohr Y, Gilbert M: **Climate change and the spread of vector-borne diseases: using approximate Bayesian computation to compare invasion scenarios for the blue-tongue virus vector *Culicoides imicola* in Italy.** *Mol Ecol* 2013, **22**:2456-66.
46. Didelot X, Everitt RG, Johansen AM, Lawson DJ: **Likelihood-free estimation of model evidence.** *Bayesian Analysis* 2011, **6**:49-76.
47. Robert CP, Cornuet J, Marin J, Pillai NS: **Lack of confidence in approximate Bayesian computation model choice.** *Proc Natl Acad Sci USA* 2011, **108**:15112-7.
48. Barnes CP, Filippi S, Stumpf MPH, Thorne T: **Considerate approaches to constructing summary statistics for ABC model selection.** *Statistics and Computing* 2012, **22**:1181-97.
49. Prangle D, Fearnhead P, Cox MP, Biggs PJ, French NP: **Semi-automatic selection of summary statistics for ABC model choice.** *Stat Appl Genet Mol Biol* 2014, **13**:67-82.
50. Drovandi CC, Pettitt AN: **Bayesian experimental design for models with intractable likelihoods.** *Biometrics* 2013, **69**:937-48.
51. Ratmann O, Andrieu C, Wiuf C, Richardson S: **Model criticism based on likelihood-free inference, with an application to protein network evolution.** *Proc Natl Acad Sci USA* 2009, **106**:10576-81.
52. Mengersen KL, Pudlo P, Robert CP: **Bayesian computation via empirical likelihood.** *Proc Natl Acad Sci USA* 2013, **110**:1321-6.



53. Liepe J, Filippi S, Komorowski M, Stumpf Michael P H: **Maximizing the information content of experiments in systems biology.** *PLoS Comput Biol* 2013, **9**:e1002888.
54. Barnes CP, Silk D, Sheng X, Stumpf Michael P H: **Bayesian design of synthetic biological systems.** *Proc Natl Acad Sci USA* 2011, **108**:15190-5.
55. Rieckermann J, Anta J, Scheidegger A, Ort C: **Assessing wastewater micropollutant loads with approximate Bayesian computations.** *Environ Sci Technol* 2011, **45**:4399-406.