Simon Tavaré

On the history of ABC

Chapter 2 in Handbook of Approximate Bayesian Computation. Eds. Sisson SA, Fan Y & Beaumont MA. Taylor and Francis, 2018.

3.1 Introduction

What follows is a personal view of the evolution of ABC – Approximate Bayesian Computation – up to 2003 and it is certainly not intended to be an exhaustive review. ABC arose as an inferential method in population genetics to address estimation of parameters of interest such as mutation rates and demographic parameters in cases where the underlying probability models had intractable likelihoods. To set the scene I will give a very brief introduction to genealogical trees and the effects of mutation, focusing on the simplest case in which a panmictic population is assumed to be very large and of constant size N, and within which there is no recombination. The treatment follows that in [**?**].

3.2 Coalescent trees and mutation

The ancestral relationships among n individuals sampled at random from the population can be described by Kingman's coalescent [**?**]. Looking back into the past, the sample has n distinct ancestors for time T_n, at which point two individuals are chosen at random to coalesce, the sample then having $n-1$ distinct ancestors. We continue merging random pairs of ancestors in such a way that for time T_j the sample has j distinct ancestors, for $j = n-1, \ldots, 2$. The times T_j are independent exponential random variables with mean

$$\mathbb{E}[T_j] = \frac{2}{j(j-1)}.$$

In this setting, time is measured in units of N generations. The height of the tree is $T_{\mathrm{MRCA}} = T_n + \cdots + T_2$, the time to the most recent common ancestor (MRCA) of the sample.

This description produces random coalescent trees, as illustrated in Figure 3.1. It is worth noting that $\mathbb{E}[T_{\mathrm{MRCA}}] = 2(1 - 1/n)$, while the average time for which the sample has just two ancestors is $\mathbb{E}[T_2] = 1$. Thus the height of the tree is influenced most by T_2, as Figure 3.1 clearly shows.

Conditional on the coalescent tree of the sample, mutations in the genomic region of interest are modelled in two steps. In the first, potential mutations are poured down the branches of the coalescent tree from the MRCA according to independent Poisson processes of rate $\theta_0/2$, where $\theta_0 = 2Nu$ is the compound mutation parameter, u being the chance of

FIGURE 3.1
Six realizations of coalescent trees for a sample of size $n = 5$, drawn on the same scale.

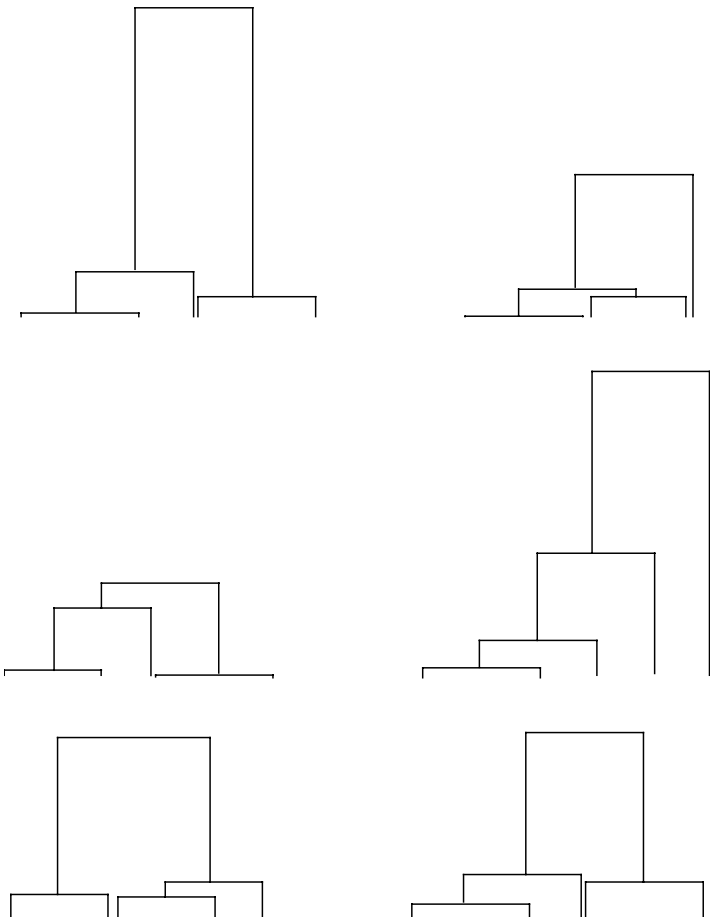
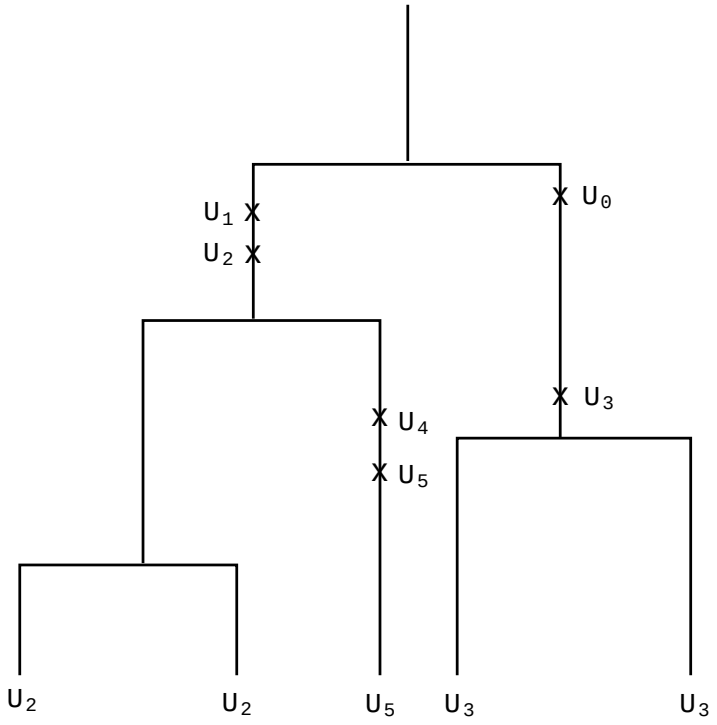


FIGURE 3.2
A coalescent tree for $n = 5$ with mutations $U_0, U_1, \ldots, U_5$ marked on the branches. The labels at the bottom of the tree give the type of each individual, assuming the infinitely many alleles model. Two types (U_2, U_3) are represented twice, and one type (U_5) once.



a mutation occurring in the genomic region in a given generation. Once the locations of mutations are determined, their effects are modeled by a mutation process that changes the current type.

I will describe three mutation models, the first being the so-called *infinitely many alleles model*, used originally to study the behaviour of allozyme frequencies. Mutations arising on the branches of the coalescent tree are marked by a sequence $U_j, j = 0, 1, \ldots$ of distinct labels, a mutation on a branch replacing the current label with the next available U. An example is given in Figure 3.2, which shows the sample of size $n = 5$ represented by labels U_2, U_2, U_5, U_3, U_3 respectively. The particular values of the types observed in a sample are not of interest; rather,

it is the number of types $C_j(n)$ represented j times in the sample, for $j = 1, 2, \ldots, n$, that records the information in these data. In the example above there are $K_5 = 3$ types, with $C_1(5) = 1, C_2(5) = 2$. Note that $C_1(n) + 2C_2(n) + \cdots + nC_n(n) = n$.

In the second example, known as the *infinitely many sites model*, we think of the genomic region of interest as the unit interval (0,1), and assume that each mutation in the coalescent tree arises at a novel position in the genomic region. These positions may be realised as values in a sequence $U_j, j = 0, 1, \ldots$ of distinct labels (generated, for example, as independent random variables uniformly distributed in $(0, 1)$). Figure 3.2 can be used to illustrate this model too. Each individual in the sample is represented as a sequence of mutations back to the root. Reading from left to right in the figure, we see that the first two individuals in the sample have mutations at locations $\{U_1, U_2\}$, the third at $\{U_1, U_2, U_4, U_5\}$, and the fourth and fifth at $\{U_0, U_3\}$.

We can write these in rather more conventional "DNA style" by representing the ancestral type as 0, mutants as 1, and recording the mutation status of each individual i in an $n \times s$ matrix, where s is the number of mutations, or *segregating sites*, that have occurred in the coalescent tree. For our example, the sequences are

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} \begin{array}{c} \begin{array}{cccccc} U_0 & U_1 & U_2 & U_3 & U_4 & U_5 \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array} \tag{3.1}$$

In practice, of course, the ancestral labeling is often not known, and neither is the time-ordered labeling of the mutations; the sequences would be recorded by ordering the columns according to positions along the genome. Any sequence dataset consistent with the infinitely many sites model, such as that in (3.1), can be represented by a rooted tree if the ancestral labeling is known, and as an unrooted tree if it is not. See Chapter 5 of [**?**] for further details.

In the previous examples mutations have a simple structure, in that they do not allow for back mutations for example. More detailed models of sequence evolution have also been developed. For example, rather than representing the DNA region as a unit interval, it might be described as a series of m completely linked sites, each site containing one of the letters A, C, G or T. There are now $M = 4^m$ possible values (or haplotypes) for each sequence. Mutations are laid down on the coalescent tree as before, the results of each mutation being given by an $M \times M$ mutation probability matrix $\mathcal{P}$. The (i, j)th element of $\mathcal{P}$ gives the probability

that sequence i mutates to sequence j when a mutation occurs. These models are referred to collectively as *finite sites models*. For historical amusement, when these models were used in the early days of sequence data the sample sizes were $n \approx 60$ and the length of the sequences $m \approx 360$ [**?**]. How things have changed!

3.3 Statistical Inference

Statistical inference for the parameter θ_0 for the infinitely many alleles model was the subject of Ewens' celebrated paper [**?**]. Ewens established that $K_n := C_1(n)+\cdots+C_n(n)$, the number of types observed in the sample, is sufficient for θ_0, and that the (moment and) maximum likelihood estimator $\widehat{\theta}_{\mathrm{E}}$ of θ_0 is the solution of the equation

$$K_n = \sum_{j=0}^{n-1} \frac{\theta}{\theta + j}.$$

In large samples, $\widehat{\theta}_{\mathrm{E}}$ is asymptotically Normally distributed with mean θ_0, and variance $\theta_0/\log n$. This last result goes some way to explaining why accurate estimation of θ_0 is hard; even modern-day sample sizes do not make much progress.

For the infinitely many sites mutation model, θ_0 has traditionally been estimated by making use of the summary statistic S_n, the number of segregating sites observed in the sample. Watterson's classic paper [**?**] derived the basic results. We note first that

> Conditional on the total length $L_n = 2T_2 + \cdots + nT_n$ of the branches of the coalescent tree, S_n has a Poisson distribution with mean $L_n\theta_0/2$.

Hence, unconditionally,

$$\mathbb{E}[S_n] = \frac{\theta_0}{2}\mathbb{E}[L_n] = \frac{\theta_0}{2}\, 2\left(1 + \frac{1}{2} + \cdots + \frac{1}{n-1}\right) = \theta_0 \sum_{j=1}^{n-1} \frac{1}{j}. \qquad (3.2)$$

This gives Watterson's unbiased estimator,

$$\widehat{\theta}_{\mathrm{W}} = S_n \Bigg/ \sum_{j=1}^{n-1} \frac{1}{j}\ .$$

In large samples, $\widehat{\theta}_{\mathrm{W}}$ is approximately Normally distributed with mean θ_0 and variance $\theta_0/\log n$. The rate of decay of the variance of $\widehat{\theta}_{\mathrm{W}}$ and $\widehat{\theta}_{\mathrm{E}}$, the reciprocal of the logarithm of the sample size n (rather than of the sample size itself, as might have been anticipated), reflects the dependence among the observations arising from the tree structure of the coalescent.

3.4 Computationally intensive methods

Computationally intensive methods have been used to fit stochastic models to data for many years. Among the early examples are [**?**, **?**, **?**, **?**]. Edwards noted that

> Particular emphasis will be placed on the need to formulate sound methods of 'estimation by simulation' on complex models.

[**?**] explored approximate likelihood methods to fit models to data, and a systematic treatment was provided in the paper by [**?**].

By the early 1990s the emergence of DNA sequence data led to a number of computational inference methods in population genetics. Among these is [**?**], who analysed mitochondrial data by using a finite sites model to describe the behaviour of purine-pyrimidine sites across the region. Their first approach compared the expected number of sites of different types with the observed numbers, and estimated parameters by matching expected to observed numbers as closely as possible. Their second approach was a composite likelihood method that treated the sites as independent. [**?**] developed an ingenious Metropolis-Hastings Monte Carlo method to estimate the parameter θ in another finite sites model, exploiting the coalescent structure to generate a likelihood curve from which inference could be made.

[**?**, **?**] introduced another approach to full likelihood based inference by exploiting a classical result about Markov chains. For a discrete-time Markov chain $\{X_k, k \geq 0\}$ with state space $\mathcal{S}$ and transition matrix $P = (p_{xy}, x, y \in \mathcal{S})$, let $\mathcal{A}$ be a set of states for which the hitting time

$$\eta = \inf\{k \geq 0 : X_k \in \mathcal{A}\}$$

is finite with probability one starting from any $x \in \mathcal{T} := \mathcal{S} \setminus \mathcal{A}$. Let f be a function on $\mathcal{S}$, and define

$$u_x(f) = \mathbb{E}\left[\prod_{k=0}^{\eta} f(X_k) \mid X_0 = x\right]$$

for all $X_0 = x \in \mathcal{S}$ (so that $u_x(f) = f(x), x \in \mathcal{A}$). Then for all $x \in \mathcal{T}$,

$$u_x(f) = f(x) \sum_{y \in \mathcal{S}} p_{xy} u_y(f). \tag{3.3}$$

A simulation approach to solve equations such as the one in (3.3) follows: simulate a trajectory of the chain X starting at x until it hits $\mathcal{A}$ at time η, compute the value of product $\prod_{k=0}^{\eta} f(X_k)$, and repeat this many times to obtain an estimate of $u_x(f)$. In the applications in [**?**, **?**], coalescent-based recursions for likelihoods were reduced to this form. The method is essentially a version of von Neumann and Ulam's suggestion for matrix inversion, as described in [**?**], and improved by sequential Monte Carlo by [**?**, **?**]. Further examples may be found in Chapter 6 of [**?**]. [**?**] showed how to exploit importance sampling to design more efficient ways to (in our language) choose the process X, and this resulted in a number of more effective inference methods; see, for example, [**?**] and [**?**]. A Markov chain Monte Carlo approach to inference for the infinitely many sites model appears in [**?**].

Summary statistics continued to be used for inference, as illustrated by [**?**] who described what is essentially the frequentist version of ABC in the context of inference about population history, based on the number of segregating sites and the mean pairwise distance among the sequences. They produced a likelihood surface over a grid of parameter values, approximating the likelihood by repeated simulation of the model and recording the proportion of simulated values of the statistics that were sufficiently close to the observed values.

The distributions of unobservable features of coalescent models, such as T_{MRCA}, conditional on observed values of the data, have also been studied by Monte Carlo methods. [**?**] considered inference for T_{MRCA} under the infinitely many sites model, using data of the form (3.1) and exploiting a version of the approach outlined in (3.3). [**?**] studied a similar problem, but using the maximal value of the number of nucleotide differences between any pair of sequences in the dataset as the observed statistic. Their method uses a simple form of density estimation to approximate T_{MRCA}.

3.5 A Bayesian approach

Bayesian methods provide a natural setting for inference not just about model parameters, but also about unobservables in the underlying model. [**?**] illustrated this for the infinitely many sites model by de-

veloping a rejection algorithm for simulating observations from T_{MRCA} and θ, conditional on the number of segregating sites $S_n = s$ seen in the data. The method is based on the observation made above (3.2) that, conditional on the times $T_2, \ldots, T_n$ and θ, the number of segregating sites in the sample of size n has a Poisson distribution with mean $\mathbb{E}[S_n|T_n, \ldots, T_2, \theta] = \theta\, L_n/2$; we write $S \sim \mathrm{Po}(\theta L_n/2)$

Suppose then that θ has prior distribution $\pi()$, and let $p(s|\lambda)$ denote the probability that a Poisson random variable with mean λ has value s:

$$p(s|\lambda) = \frac{e^{-\lambda}\lambda^s}{s!}, s = 0, 1, \ldots.$$

The rejection algorithm is

A1 Generate $\theta \sim \pi(\cdot)$

A2 Generate $T_n, \ldots, T_2$ from the coalescent model. Calculate $L_n = \sum_{j=2}^n jT_j$ and $T_{\mathrm{MRCA}} = \sum_{j=2}^n T_j$.

A3 Accept $(\theta, T_{\mathrm{MRCA}})$ with probability proportional to

$$\alpha = p(s|\theta L_n/2)$$

Accepted values of this algorithm have the required distribution, that of $(\theta, T_{\mathrm{MRCA}})$ given $S_n = s$.

The previous method may be viewed as an application of the rejection algorithm, which proceeds as follows. For discrete data $\mathcal{D}$, probability model $\mathcal{M}$ with parameters θ having prior $\pi()$, we can simulate observations from

$$f(\theta|\mathcal{D}) \propto \mathbb{P}(\mathcal{D}|\theta)\, \pi(\theta) \tag{3.4}$$

via

B1 Generate $\theta \sim \pi(\cdot)$

B2 Accept θ with probability proportional to the likelihood $\mathbb{P}(\mathcal{D}|\theta)$.

This method can be extended dramatically in its usefulness using the following, stochastically equivalent, version:

C1 Generate $\theta \sim \pi(\cdot)$

C2 Simulate an observation $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ

C3 Accept θ if $\mathcal{D}' = \mathcal{D}$

For the example in algorithm A above, C3 takes the form

C3 Simulate an observation $S \sim \text{Po}(\theta L_n/2)$, and accept $(T_{\text{MRCA}}, \theta)$ if $S = s$.

While algorithms B and C are probabilistically identical, C is much more general in that one does not need to compute probabilities explicitly to make it work; only simulation is needed. Version C is due to [**?**]. Surprisingly, the result does not seem to be described in text books that focus on simulation.

The drawback in C is clear. It will typically be the case that for a given value of θ the chance of the outcome $\mathcal{D}' = \mathcal{D}$, namely $\mathbb{P}(\mathcal{D}|\theta)$, is either vanishingly small, or very time consuming to compute, resulting in an algorithm that does not work effectively. This is where ABC finally comes into play, in the form of the following scheme. We start with a metric ρ to compare data sets and a tolerance $h \geq 0$, and then

D1 Generate $\theta \sim \pi(\cdot)$

D2 Simulate an observation $\mathcal{D}'$ from model $\mathcal{M}$ with parameter θ

D3 Compute $\rho := \rho(\mathcal{D}', \mathcal{D})$, and accept θ as an approximate draw from $f(\theta|\mathcal{D})$ if $\rho \leq h$.

The parameter h measures the tension between computability and accuracy. If ρ is a metric, then $\rho = 0 \implies \mathcal{D}' = \mathcal{D}$, so that such an accepted θ is indeed an observation from the true posterior.

[**?**] were the first to describe a version of this scheme, in which the data sets in D3 were compared through a choice of summary statistics. Thus ρ compares how well a set of simulated summary statistics matches the observed summary statistics. If the statistics are sufficient for θ, then when $h = 0$ the accepted values of θ are still from the true posterior based on the full data. This begs the question of how one might identify 'approximately sufficient' statistics, a topic covered elsewhere in this book. The method is also applicable to continuous data.

3.6 ABC takes off

[**?**] showed that it might be better to soften the hard cut-off suggested in algorithm D, by making use of the all the simulated values. They proposed to weight values of θ by the size of the corresponding distance ρ; smaller values of ρ suggest an observation whose distribution is closer to the required posterior. They made a number of suggestions for how

the weights might be chosen, and then used to produce a sample with better sampling properties than the original hard cut-off method.

The development of ABC was predicated on the availability of computational power and the lack of tractable likelihoods. The latter is also an issue for Markov Chain Monte Carlo methods, and this motived [**?**] to suggest an MCMC method that does not need likelihoods in its implementation.

In the present setting the idea behind classical MCMC [**?**] is to construct an ergodic Markov chain that has $f(\theta|\mathcal{D})$ as its stationary distribution. In skeleton form, it works as follows:

E1 The chain is now at θ

E2 Propose a move to θ' according to a proposal distribution $q(\theta, \theta')$

E3 Calculate the Hastings ratio

$$\alpha = \min\left(1, \frac{\mathbb{P}(\mathcal{D}|\theta')\pi(\theta')q(\theta', \theta)}{\mathbb{P}(\mathcal{D}|\theta)\pi(\theta)q(\theta, \theta')}\right) \tag{3.5}$$

E4 Move to θ' with probability α, else remain at θ.

It is the ratio of likelihoods in E3 that might cause problems. [**?**] proposed the following:

F1 The chain is now at θ

F2 Propose a move to θ' according to $q(\theta, \theta')$

F3 Generate $\mathcal{D}'$ using parameter θ'

F4 If $\mathcal{D}' = \mathcal{D}$, go to F5, else remain at θ

F5 Calculate

$$\alpha = \min\left(1, \frac{\pi(\theta')q(\theta', \theta)}{\pi(\theta)q(\theta, \theta')}\right)$$

F6 Move to θ' with probability α, else remain at θ.

This likelihood-free method does indeed have the correct stationary distribution. In practice the rejection step is often replaced by a version of D3:

F4 If $\rho(\mathcal{D}', \mathcal{D}) \leq h$, go to next step, else return θ,

and this too might involve a comparison of summary statistics. [**?**] were able to assess the effect of summary statistics in a population genetics problem, and [**?**] used the method in a problem concerning divergence times of primate species.

[**?**] also suggested that the likelihood terms in (3.5) be approximated by estimates of the form

$$\hat{\mathbb{P}}(\mathcal{D}|\theta) = \frac{1}{B}\sum_{j=1}^{B}\mathbb{I}(\mathcal{D}'_j = \mathcal{D})$$

for B independent simulations of the model with parameter θ; algorithm F is the special case $B = 1$. [**?**] made a similar suggestion, and this motivated the development of the 'pseudo-marginal method' [**?**], discussed elsewhere in this book.

3.7 Conclusions

The term "Approximate Bayesian computation" has arisen more than once. For example, [**?**] used it to describe computation based on the asymptotic behaviour of signed roots of log-density ratios. He argued that

> ... analytic approximation still has an important role to play in Bayesian statistics.

In the setting of the present handbook, and in some sense at the other end of the analytical spectrum, it was [**?**] who coined the term 'Approximate Bayesian Computation,' in the article that made ABC the popular technique it has become.

Where did the acronym 'ABC' arise? By the time we submitted [**?**] for publication at the end of 2002, the USC group had held many meetings on what we then called ABC. In the submitted version, the term ABC appeared twice:

> ... we have the following approximate Bayesian computation (ABC) scheme for data $\mathcal{D}$ summarized by S

and

> ... and it is often useful to replace the full data by a number of judiciously chosen summary statistics. The resulting approximate Bayesian computation, which we dub ABC, allows us to explore scenarios which are intractable if the full data are used.

In the published version ABC does not appear, because of house style at the time in the Proceedings of the National Academy of Sciences. This from the proofs:

> D – AU: Per PNAS style, nonstandard abbreviations are allowed only when used at least 5 times in the main text.

A missed opportunity for the National Academy of Sciences! In 2003 I gave an invited lecture at the Royal Statistical Society entitled (with a certain amount of bravado) 'Who needs likelihoods', in which ABC appeared several times, as the write-up in the RSS News in October 2003 showed. It concluded:

> The lively discussion that followed reinforced our feeling that we were not hearing the last of ABC.

This observation turned out to be true, and ABC has become a standard approach in the statistician's toolbox. New areas of application arise frequently, as the rapidly expanding literature shows. One area that would repay deeper analysis is that of cancer evolution, a field that is producing enormous amounts of DNA sequence and phenotype data and for which there is a dearth of inference methods. For an early application see [**?**]. Inference for agent-based models in stem cell biology appears in [**?**], which motivated the approach in [**?**] for colorectal cancer.

3.8 Acknowledgements

I thank Dr. Andy Lynch and two anonymous reviewers for helpful comments on an earlier version of this article, and Paul Gentry, Conference and Events Manager at the Royal Statistical Society, for tracking down the RSS News report on ABC.