Genetics and Stochastic Simulation *Do* Mix!

May 2, 2006

Abstract

The rapid development of biological technology in recent years has generated an overwhelming quantity of highly complex genetical data. In retrospect it is perhaps now obvious that standard stochastic sampling algorithms were destined to fail when applied to the type of models that these data require. During this period, both the genetical and statistical research communities have developed a range of new algorithms that have endeavoured to avoid these failures. In this article we present a brief journey through the development of some of these algorithms. We hope to demonstrate that continued collaborative research and education in both genetics and statistics disciplines is crucial for the advancement of future sampling frameworks, and call on researchers in both fields to take up this challenge.

Keywords: Approximate Bayesian Computation; Markov Chain Monte Carlo; Population Genetics; Reversible Jump; Simulated Tempering; Statistical Genetics.

1 Introduction - Initial Value

Research literature in different disciplines has a tendency to persist in parallel streams, although genuine cross-disciplinary research is not as uncommon as it once was. Unfortunately this seems to occur no matter how closely related the concepts in principle or in practice. In spite of this, the occasional meme transmission between two literatures often induces an advancement in methods or understanding in previously overlooked or abandoned areas, as suddenly fresh viewpoints on old subject matters spark new ideas and insight. While diverse disciplines, genetics and statistics are especially fortunate in this respect. In particular, in light of the creation of enormous amounts of genetic data from complex systems through recent technological advancement, researchers in both communities have been eager to take up the challenges which have followed.

In this article we examine a snapshot of the development of one aspect of research in the genetical and statistical literatures in which there is substantial overlap – namely that of novel stochastic sampling frameworks. Naive implementations of standard algorithms, such as Markov chain Monte Carlo, have faced many unforseen obstacles when applied to genetics-based problems, given the complex nature of the implemented models and the involved local dependency structures these models induce. The resulting development of algorithmic solutions to these problems has had implications in both genetics and statistics. We hope to demonstrate that progress in one field, and the transmission of these new ideas between the disciplines, has in the past proved crucial for the methodological advancement in the other. While these developments are not necessarily causal, we argue that the continuation of this interaction, and the implementation of interdisciplinary education, is crucial for technological progress in the future.

In the spirit of the sampling frameworks themselves, we take as our initial value (and admittedly not quite at random) one of the most accessible and popular algorithms to date.

2 Transmissions between literatures

2.1 Metropolis-Hastings samplers

Despite arising originally in the statistical physics literature over 50 years ago, the Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) was only brought to the attention of the mainstream statistics literature by Gelfand and Smith (1990).

This algorithm broadly states how to sample from a distribution $\pi(\theta|x)$ defined upon a parameter vector $\theta \in \Theta$ and with observed data x. Subject to a few technical requirements, a sequence of dependent vectors $\theta^{(0)}, \theta^{(1)}, \ldots$ may be constructed given an initial value $\theta^{(0)}$. For each step, a new value $\theta' \sim q(\theta'|\theta^{(t)})$ is generated by sampling from a non-empty subset of the components of $\theta^{(t)}$, for each $t = 1, 2, \ldots$, where q is an arbitrary density defined upon Θ. The resulting vector in each case is accepted as the new point, so that $\theta^{(t+1)} = \theta'$, with probability

$$\min\left\{1, A\left[\theta^{(t)} \rightarrow \theta'\right]\right\},$$

where $A\left[\theta^{(t)} \rightarrow \theta'\right] = \pi(\theta'|x)q(\theta^{(t)}|\theta')/\pi(\theta^{(t)}|x)q(\theta'|\theta^{(t)})$ in this case, otherwise $\theta^{(t+1)} = \theta^{(t)}$ remains unchanged. A special case of this algorithm, known as the Gibbs sampler, arises when $\pi(\theta'|x) = q(\theta'|\theta^{(t)})$ generating an acceptance probability of one. See, for example, Chib

and Greenberg (1995) for a useful introduction, or Robert and Casella (2004); Gilks et al. (1995) for more detailed reference texts.

A popular and conceptually accessible generalisation of the above scheme is the "reversible jump" sampler proposed by Green (1995) which extends the Metropolis-Hastings sampler to a multi-model setting. Here we now entertain a (usually) countable set of models $\mathcal{M} = \{M_1, M_2, \ldots\}$ each defined upon their own (possibly overlapping) state spaces $\Theta_1, \Theta_2, \ldots$ admitting inference on the model vector and model indicator pair (θ_k, k). The procedure is then the same as for the Metropolis-Hastings algorithm (albeit with a few more details) in that new samples are generated by proposing to move from the current pair $(\theta^{(t)}, k^{(t)})$ to a proposed pair (θ', k') for which k' may not be equal to $k^{(t)}$, and for which θ' and $\theta^{(t)}$ may be of differing lengths. This move is then similarly accepted with probability

$$\min\left\{1, A\left[(\theta^{(t)}, k^{(t)}) \rightarrow (\theta', k')\right]\right\},$$

so that $(\theta^{(t+1)}, k^{(t+1)}) = (\theta', k')$ or else $(\theta^{(t+1)}, k^{(t+1)}) = (\theta^{(t)}, k^{(t)})$ as before. The exact expression for $A\left[(\theta^{(t)}, k^{(t)}) \rightarrow (\theta', k')\right]$ is a little more detailed than for the standard Metropolis-Hastings, which we omit here for brevity. However tutorial-style expositions in the QTL and change-point modelling settings are provided respectively by Waagepetersen and Sorensen (2001) and Green (2001), or in a more research-oriented setting by Sisson (2005).

The application of these two samplers alone in the literature is immense – their endemic usage is certainly too vast to do justice here. We merely refer the interested reader in particular to the informative sources of Sorensen and Gianola (2002), Balding et al. (2003) and Gelman et al. (2004) and the extensive applications and references therein.

2.2 Naive samplers don't mix

While simple Monte Carlo methods had been established in the genetics literature for many years previously (Wright and McPhee 1925; Edwards 1968, for example), the Gibbs sampler was immediately adopted as a means to approximate probabilities on genealogies when exact computation (via the *peeling* algorithm of Cannings et al. 1978) was infeasible (Thompson 2000).

Under Mendelian laws of inheritance an individual randomly receives one allele (gene type) independently from each parent, and in turn randomly transmits an allele to each of its own offspring. For example, in the pedigree in Figure 1, individual 11 receives an A-allele from its maternal parent, and transmits either a B or C-allele to its progeny. Accordingly, individual 11 must carry either genotype AB or AC.

Given the Mendelian laws, the Markov nature of the models adopted on pedigrees made the Gibbs sampler particularly attractive from a computational viewpoint. In such a sampler, the parameters of interest were the unknown genotypes, at a single locus, of individuals related via a pedigree structure (such as in Figure 1), or similarly through the reformulation into inheritance vector representations (Lander and Green 1987). However it was soon discovered that naive application of Markov chain based samplers would lead to problems in terms of poor sampler *mixing* (whereby the sampler moves inefficiently around the parameter space), and could even result in reducible samplers – that is, samplers that were not able to visit all parts of the parameter space (Jensen and Sheehan 1998; Sisson 2002).

For the pedigree in Figure 1, while individuals 11 and (by symmetry) 12 must both carry either AB or AC genotypes, because both B and C alleles are transmitted to individual

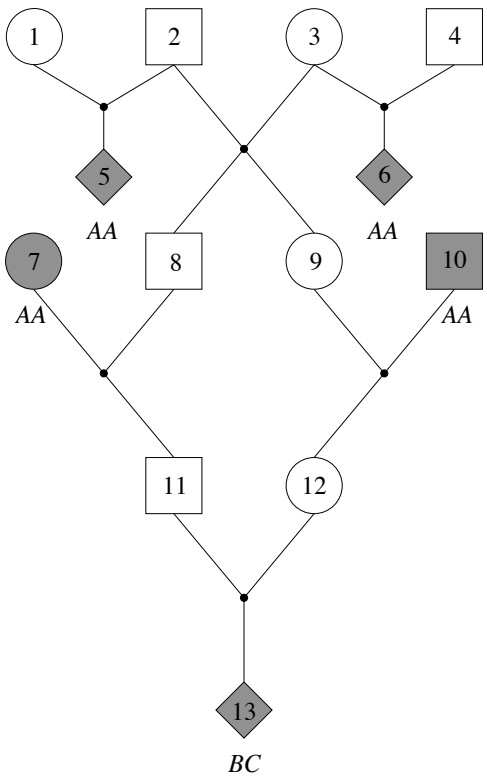

Figure 1: An inbred pedigree featuring a first-cousin mating with three alleles (A, B, C) at the locus of interest, taken from Jensen and Sheehan (1998). Circles, squares and diamonds denote female, male and individuals of unspecified gender respectively. Shaded individuals have known genotype.

13, the genotypes of the pair (11, 12) are constrained to either (AB, AC) or (AC, AB). Given these restrictions on the parameter space, a Markov chain starting at (AB, AC) for these individuals could never reach the configuration (AC, AB) with single genotype Gibbs updates as the necessary intermediary states (AB, AB) or (AC, AC) were *inconsistent* (i.e. occurred with probability 0) with the pedigree structure and observed genotypes. Hence, such a sampler is reducible. In addition for this example, because the B and C alleles must originate from individuals 2 and 3, this pair is similarly constrained to carry (AB, AC) or (AC, AB).

In general, sampler reducibility stemmed from the complex local dependency structures induced by pedigree structure, in combination with partial information observed through the genealogy. To complicate matters, in general it is not known which nor how many individuals have such constraints. As most real world studies constitute multi-locus analyses the magnitude of the problem was compounded further, affording severe problems for algorithm design.

A number of solutions to the sampler reducibility problem were proposed, either focusing on extra and (often unavailable) knowledge of the problem (Jensen and Kong 1999; Hurn et al. 1999; Lin 1995) or by adopting an augmented state space approach to improve mixing (Sheehan and Thomas 1993; Lin et al. 1993; Geyer 1991). The method of Sheehan and Thomas (1993) proposed assigning a small probability, γ, to all inconsistent configurations on the pedigree, enabling a Markov chain to step through the necessary states needed to

		Genotype					
	Individual	AA	AB	AC	BB	BC	CC
$\gamma = 0$	2	0	0	1.00	0	0	0
Rejection rate = 0%	8	0	0.26	0	0.45	0.30	0
Length of chain = 1,000	11	0	0	1.00	0	0	0
$\gamma = 0.001$	2	0	0.89	0.11	0	0	0
Rejection rate = 50.98%	8	0	0.11	0.18	0.20	0.28	0.24
Length of chain = 2,041	11	0	0.41	0.59	0	0	0
$\gamma = 0.005$	2	0	0.41	0.59	0	0	0
Rejection rate = 92.18%	8	0	0.14	0.20	0.17	0.33	0.16
Length of chain = 12,768	11	0	0.53	0.47	0	0	0
$\gamma = 0.05$	2	0	0.51	0.49	0	0	0
Rejection rate = 99.02%	8	0	0.18	0.14	0.14	0.35	0.18
Length of chain = 101,386	11	0	0.50	0.50	0	0	0
	2	0	0.50	0.50	0	0	0
True values	8	0	0.17	0.17	0.17	0.33	0.17
	11	0	0.50	0.50	0	0	0

Table 1: Marginal genotype distributions of selected individuals, estimated from 1,000 consistent simulations for varying values of $\gamma = 0, 0.001, 0.005, 0.05$.

visit each consistent configuration. Following the simulation, all inconsistent states were discarded and only the consistent configurations taken as the final sample.

Table 1 shows the results of applying this method to the pedigree in Figure 1 for varying values of γ. If for simplicity we assume that all consistent configurations on the pedigree are equally likely, then individuals 2, 3, 11 and 12 will each carry an AB genotype for exactly 50% of all consistent samples, and genotype AC for the rest (while respecting the known joint constraints). Similarly, individuals 8 and 9 may carry every genotype excluding AA, with the BC genotype being twice as common.

A Markov chain was implemented until 1,000 consistent configurations were obtained. When $\gamma = 0$ the sampler only considers consistent configurations, and so is unable to move away from the consistent starting configuration of genotype AC for individuals 2 and 11. This chain is clearly reducible, as expected. As γ increases, thereby giving greater probability to inconsistent states, two things happen. Firstly, the chain moves increasingly freely around the parameter space, enabling more frequent visits to consistent configurations and thereby enabling more accurate probability estimates. Secondly, the chain spends longer periods of time in inconsistent configurations, meaning that a longer Markov chain is required to generate a fixed number of consistent samples. Here, when $\gamma = 0.05$, although the marginal probabilities are fairly constistent with their true values, a chain of length 101,385 was needed to produce 1,000 consistent configurations. Accordingly there is a trade-off between accuracy and computational intensity.

However, of all such augmented state space algorithms, one in particular has had a hugely important influence in the mainstream statistics literature.

2.3 Simulated tempering samplers

In an application involving ancestral inference on a (then) very large genealogy, Geyer and Thompson (1995) generalised the ideas behind some of the augmented state space samplers to develop a broad sampling framework that would "mix rapidly enough to be usable for problems in which other methods would require eons of computing time... which [is] essential for attacking very hard problems, which arise in areas such as statistical genetics." Broadly the simulated tempering algorithm reformulates the multiple coupled sampler of Geyer (1991) into a single Markov chain and, given a more general framework, permits application to a wider range of modelling situations. The method has now become a standard technique to overcome problems associated with slowly mixing Markov chains.

Commonly (although not exclusively – we adopt this setting for illustration), simulated tempering comprises simulation via standard stochastic methods from a model $\pi(\theta, \tau|x) \propto \pi(\theta|x)^{1/\tau}$, defined over an augmented state space $\Theta \times \mathcal{T}$ with $\mathcal{T} = \{1, 2, \ldots, \tau_{\max}\}$, leading to a generated series $\{(\theta^{(t)}, \tau^{(t)})\}$. Here, τ is colloquially referred to as the "temperature" of the system in analogy with the physical annealing process. Thus $\tau = 1$ is the "cold" distribution, and $\tau > 1$ generates increasing degrees of "heated" distributions. Our distribution of interest is then $\pi(\theta|x) = \pi(\theta|x, \tau = 1)$. Subsetting the sampler output to obtain $\{\theta^{(t)} : \ \tau^{(t)} = 1\}$ results in a series whose stationary distribution is the target distribution. The augmented sampler will possess improved mixing properties as $\pi(\theta|x)^{1/\tau}$ is flat relative to $\pi(\theta|x)$ when τ is large, so that move proposals made when τ is large will have a greater acceptance probability. For the augmented state space sampler of Sheehan and Thomas (1993) in the previous section, γ is the augmenting parameter. Permitting γ to vary during chain implementation would be an illustration of a simulated tempering sampler in this setting.

Although a generic sampling scheme in its own right, the simulated tempering algorithm is beginning to be incorporated into other areas in the sampling literature, both genetical and statistical, that are currently active in research. One such area in genetics is discussed in the next section. Another, in the statistical literature, known as perfect or exact sampling aims to draw independent realisations exactly from $\pi(\theta|x)$. Thereby the usual Markov-sampler practicalities, such as the necessity of ensuring that the chain has converged, are circumvented, and the output realisations are independent rather than dependent, accordingly reducing the number required for inference. Perfect sampling is considered one of the "holy grails" of statistical sampling algorithms. While there are certainly other exact sampling algorithms available, we only consider those based on the tempering framework here.

Intuitively, in analogy with simulated tempering, the sampler proceeds by moving through the augmented state space of "heated" models, indexed by $\tau \in \mathcal{T}$, and in each round of parameter updates, proposes a move from $(\theta^{(t)}, \tau^{(t)})$ to the same point $(\theta^{(t)}, \tau^*)$ in a distribution indexed by $\tau = \tau^*$, from which it is possible to sample directly. Here, any of the standard statistical distributions, such as the Normal or Uniform (for which $\tau^* \to \infty$), would be sufficient. The smallest possible probability of this occurring at any stage, assuming an identical state space Θ for all tempering models, is given by

$$\epsilon = \inf_{\theta \in \Theta, \tau \in \mathcal{T}} q(\tau^*|\tau) \min\left\{1, A\left[(\theta, \tau) \to (\theta, \tau^*)\right]\right\},$$

where $q(\tau^*|\tau)$ is the probability of proposing the move from model τ to model τ^*.

We now hypothetically consider an infinite number of samplers commencing at time $t = -\infty$ initialised at every possible point $\theta \in \Theta$. If we can reasonably assume that all of

these chains coalesce into a single chain in the model τ^* with probability ϵ (and we make this assumption), focus would then naturally be upon the *first* instance this occurred in the reverse-time chain from $t = 0, -1, -2, \ldots, -T$. By logical argument it can be deduced that $T \sim \text{Geometric}(\epsilon)$. Consequently, commencing a forward-time Markov chain sampler starting in model τ^* at time $t = -T$, where T is sampled randomly from its known distribution, will have the effect of generating an independent realisation exactly from $\pi(\theta|x)$ at time $t = 0$. The process is then repeated for as many realisations as required.

The above argument is fairly subtle, but is now generally accepted. Propp and Wilson (1996) discuss the original above idea of "coupling chains from the past," whereas the tempering-based sampler is described in Brooks et al. (2006) (see also Møller and Nicholls 1999). The perfect sampler derived from the simulated tempering framework of Geyer and Thompson (1995) is unique among current perfect samplers in that it may additionally be extended to the "reversible jump" setting.

This extension is conceptually straight-forward. The Markov chain is now defined on joint parameter and model indicator space, where $\tau \in \mathcal{T}$ now indexes individual models, and the parameter $\theta_\tau \in \Theta_\tau$ may depend on the model. The description of the algorithm remains unchanged, except that the smallest probability of moving to model τ^* at any stage must now be explicitly calculated over all models of possibly varying dimension. That is

$$\epsilon = \inf_{\theta_\tau \in \Theta_\tau, \theta_{\tau^*} \in \Theta_{\tau^*}^{\tau}, \tau \in \mathcal{T}} q(\tau^*|\tau) \min\left\{1, A\left[(\theta_\tau, \tau) \to (\theta_{\tau^*}, \tau^*)\right]\right\},$$

where $\Theta_{\tau^*}^{\tau} \subseteq \Theta_{\tau^*}$ denotes the subset of parameter space in model τ^* which it is possible to reach from model τ. While a number of issues remain regarding implementation for this algorithm – mainly involving the practicality in estimating ϵ in this setting – it is clear that the tempering framework has high importance in both genetical and statistical literatures.

We now examine yet another area in which the concepts of simulated tempering are beginning to influence new types of algorithms in statistical genetics.

2.4 Bayesian computation with intractable likelihoods

While the standard Metropolis-Hastings algorithm and its variants have proved invaluable in a wide range of genetics-based analyses, many scenarios – particularly in the population genetics literature – have uncovered unforseen limitations to these methods in practice. Specifically, many studies typically consist of large numbers of nuisance parameters which in theory may be stochastically "integrated out" through the sampling procedure. However, in practice, large numbers of parameters in combination with complex models can result in likelihoods that are computationally prohibitive or even impossible to evaluate. See, for example, Beaumont et al. (2002); Marjoram et al. (2003); Estoup et al. (2004); Hamilton et al. (2005). In addition, Chikhi and Bruford (2005) note that when likelihoods are available, their evaluation can be prohibitively slow simply due to the shear size of typical genetics datasets generated in the genomic era (e.g. 10–20 microsatellite loci typed for 200-300 individuals). The practicality of likelihood-driven stochastic sampling methods is thereby severely limited.

Over the last few years a new family of stochastic simulation algorithms has been developed, primarily in the genetics literature, in order to circumvent problems faced in evaluating the likelihood in traditional sampling methods. These have since been termed "approximate

Bayesian computation" techniques (c.f. Beaumont et al. 2002), although they are by no means restricted to inference under the Bayesian paradigm.

Broadly, rather than evaluating the likelihood, these methods assess how closely a given vector of parameters, θ, corresponds to the observed data through elicitation of data summary statistics, $S = (S_1, \ldots, S_q)$, in accordance with the "model" under consideration. A basic rejection sampling algorithm has evolved following methods proposed by a number of authors (Tavaré et al. 1997; Fu and Li 1997; Weiss and von Haeseler 1998; Pritchard et al. 1999). These state that at each stage, a new point $\theta' \in \Theta$ is sampled independently from the prior (or non-Bayesian equivalent) $\pi(\theta)$. A simulated data set generated from the proposed parameter vector is then simulated according to the model, and the appropriate summary statistics $S_i(\theta')$ then evaluated. The correspondence of the vector θ' to the data is then evaluated through the summary statistics via an appropriate metric

$$\rho(S, S(\theta')) = ||(S_1, \ldots, S_q) - (S_1(\theta'), \ldots, S_q(\theta'))||,$$

and the point θ' is accepted if $\rho(S, S(\theta')) \leq \delta$ for some tolerance δ. Otherwise, the point is rejected as not residing in the portion of the state space which adequately represents the observed data.

While possessing considerable intuitive appeal, these methods unfortunately enjoy a number of impracticalities which have only recently begun to generate possible solutions. One clear problem is the inefficient nature of simulating proposal values directly from the prior, as in general prior densities do not coincide with the posterior. Another relates to the choice of tolerance, δ. If this value is taken too low, the algorithm will yield very poor acceptance rates; if too high, then the sample will generate a biased distribution as unlikely parameter vectors are accepted.

Two qualitatively different approaches have attempted to tackle these issues. The first, suggested by Beaumont et al. (2002) aims to reuse those samples otherwise rejected as being in exceedance of the tolerance δ. In this scenario, all proposed samples are transformed via a local-linear regression in the direction of the $(S - S(\theta'))$ axis, weighted according to the value of $\rho(S, S(\theta'))$ to reduce the effect of the discrepancy between S and $S(\theta')$. The transformed samples are then accepted as realisations from the distribution of interest. Thus poor acceptance rates are circumvented, and the tolerance does not become a defining efficiency parameter.

A second approach to the problem of sampling inefficiency has been the adoption of a Markov chain sampling framework (Marjoram et al. 2003), where the proposal parameter vector θ' is drawn conditionally on the previous value $\theta^{(t)}$ via some proposal density $q(\theta'|\theta^{(t)})$. A Metropolis-Hastings approach is then adopted, accepting the move from $\theta^{(t)}$ to θ' with probability $\min\left\{1, A\left[\theta^{(t)} \rightarrow \theta'\right]\right\}$, where in this situation

$$A\left[\theta^{(t)} \rightarrow \theta'\right] = \frac{\pi(\theta')q(\theta^{(t)}|\theta')}{\pi(\theta^{(t)})q(\theta'|\theta^{(t)})}\mathbf{1}(\rho(S, S(\theta')) \leq \delta),$$

where $\mathbf{1}(\mathcal{X}) = 1$ if $\mathcal{X}$ is true, and 0 otherwise. The acceptance probability is thereby equivalent to the standard Metropolis-Hastings formulation with the difference that the likelihood ratio is replaced with the equivalence to within a tolerance of the simulated summary statistics via the indicator $\mathbf{1}(\cdot)$. In this manner, the Markov chain will converge to the region of the state space with the greatest posterior density, increasing the acceptance rate of the algorithm.

As an illustration, consider the analysis of Tanaka et al. (2006), who construct a stochastic model at the genetic level, to examine the transmission and mutation of tuberculosis. Their model simulates a sample of infected individuals from which it is possible to count and group the different genotypic strains of the pathogen, and compare directly to an observed sample from a real outbreak. The "clusteredness" of the sample – a function of the number of individuals with each strain – provides information on the extent of recent transmissions, assuming the mutation rate is known. For example, a high proportion of single-individual clusters indicates a high level of recent disease transmission. The innovation of Tanaka et al. (2006) lies in using an explicit model of transmission and mutation, along with Bayesian estimation of key parameters from genotypes of tuberculosis isolates.

Denoting n_i as the number of individuals with pathogen genotype i, and $n = \sum n_i$ as the sample size, Tanaka et al. (2006) apply the Metropolis-Hastings approach of Marjoram et al. (2003) with

$$\rho(S, S(\theta)) = |H - H(\theta)| + \frac{1}{n}|g - g(\theta)|,$$

where

$$H = 1 - \sum (n_i/n)^2$$

is a summary statistic describing the genetic diversity of the sample, and g is the number of distinct genotypes in the sample. Figure 2 illustrates posterior output of this analysis.

The top image in Figure 2 illustrates the effect of the tolerance, δ, on the performance of the chain. Visually, lower tolerances yield appreciably lower acceptance rates. Tanaka et al. (2006) attain an acceptance rate of 0.3% in their final analysis (black line). The other two images in Figure 2 illustrate the relative bias induced by the tolerance on key posterior estimates of the study. The doubling time is the required duration for the number of cases in the population to double. The reproductive value is the expected number of new cases produced by a single infectious case while the primary case is still infectious. Clearly there is an evolution of each posterior as the tolerance decreases. While the true posterior can only be realised for $\delta = 0$, this is in general impractical in all but the simplest simulations. In practice, the hope is that the posterior will remain essentially unchanged beyond the adopted tolerance level.

The method of Marjoram et al. (2003) still requires *a priori* specification of the tolerance δ, and is clearly affected by mis-specification. An extension by Bortot et al. (2006) proposed improving this framework in analogy with the simulated tempering framework of Geyer and Thompson (1995) by augmenting the parameter space to include the tolerance (θ, δ). By specifying a "pseudo-prior" on δ (that is, a prior that serves only to facilitate sampler performance), the algorithm will exhibit improved mixing over a range of tolerances. This permits the final sample to be derived by either conditioning on those samples with $\rho(S, S(\theta')) \leq \delta^*$, for some δ^* chosen *a posteriori*, or by weighted transformation of the full sample in the manner of Beaumont et al. (2002). Most recently, Sisson et al. (2006) have noted that likelihood-free Bayesian computation methods based on Markov chains (that is, the methods by both Marjoram et al. 2003 and Bortot et al. 2006) are particularly inefficient at estimating the distributional tails. They advocate a population-based sequential Monte Carlo sampler as a way to realise improved and more efficient estimation, and suggest that their algorithm is more flexible than single Markov chain-based methods in terms of simulating from more complex (e.g. multi-modal) distributions.

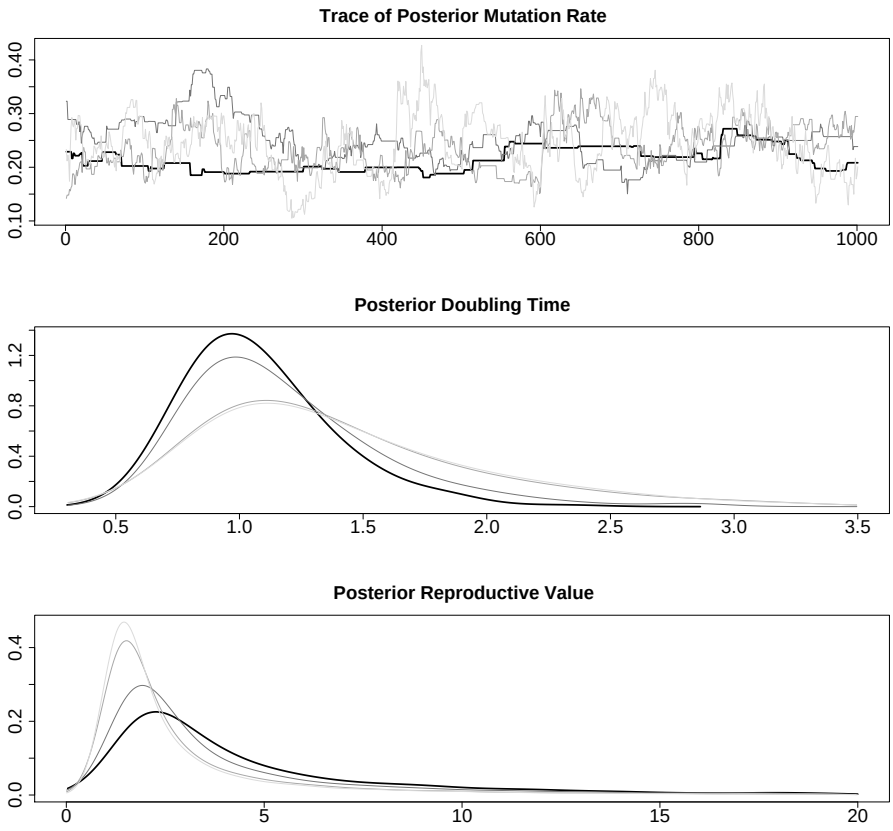

Figure 2: Various posterior estimates using the model of Tanaka et al. (2006) with a range of tolerance values, δ. Darker lines denote lower tolerances. Final posterior estimate is denoted by black line.

There remain a number of unresolved issues relating to all of the above methods, ensuring that research in this area will be active for years to come. These involve the derivation of appropriate sufficient (or approximately sufficient) summary statistics which is not necessarily a trivial procedure, the sampling variability of these statistics which will affect the acceptance rate of the algorithm, the metric ρ and the most efficient formulation of the algorithm itself.

3 Discussion - Future Proposals and Implications

In this article we have presented a small subset of the ways in which genetical and statistical research have interacted in a positive way to advance methods and tools in stochastic simulation techniques. Others undoubtedly exist. For example, the biologically inspired genetic algorithms again coupled with ideas from simulated tempering are now being adopted to perform simulation according to specified densities (Liang and Wong 2000) rather than merely finding solutions according to their fitness functional. Similarly, methods recently developed in the statistical literature to increase the acceptance rates between models in reversible jump samplers (Brooks et al. 2003) are now being taken up in the analysis of quantitative trait loci.

The implications of these interactions are multi-faceted. One important issue arises regarding within which academic journals both geneticists and statisticians should be publishing and reading. While motivated by problems in applying routine statistical methods to problems in genetics, the simulated tempering article of Geyer and Thompson (1995) appeared in the statistics literature (*Journal of the American Statistical Association*). The impact of the leap in technology was accordingly presented to an audience interested in further developing this methodology, while simultaneously highlighting the rich area of genetical problems and applications. A decade on, so-called "biostatistics" is now overwhelmingly considered an important area of statistical research in its own right.

In contrast, the early approximate Bayesian computation papers (Beaumont et al. 2002; Marjoram et al. 2003) respectively appeared in *Genetics* and the mildly biologically oriented *PNAS*. Accordingly, while many population geneticist researchers now have the tools to implement analyses that were previously intractable, these methodologies have been largely overlooked by the statistics community. While that is now beginning to change (and continuing this trend is one of the aims of this article), the obvious difficulty is that unless statisticians read a broader literature than mainstream statistics journals, they risk missing out on exciting developments in related disciplines, and detecting unforseen problems with their own methods. This process can be greatly aided, for example, by signing up to receive table-of-contents email alerts from a selection of the most suitable genetics journals. Similarly, statisticians also need to publish in a broader range of journals than the top statistics journals, to ensure that the latest methodologies are available to those who wish to use them.

Given the multidisciplinary implications for research reading and publishing, a downstream effect is that to achieve optimal research dynamics more encouragement should be given to researchers to conduct research of an interdisciplinary nature. For statisticians this directional research is sometimes self-driven; the latest example of this being the recent rush to develop methodology appropriate for the analysis of "large p, small n" microarray data. Similarly consider the number of positions open for *bio*-statisticians in recent years. However, this push can only be sustained and fully realised if statisticians are increasingly trained in relevant biological and genetics-based areas as part of their graduate education, and if the prominence of interdisciplinary programs are raised. It is no coincidence that the authors Geyer and Thompson (1995), Beaumont et al. (2002) and Marjoram et al. (2003) primarily research in the interface between statistics and genetics.

In concluding, while we have so far resisted the temptation to describe the transmission of research theory and practice between genetical and statistical literatures as analogous to a Markov chain defined on a "literature" state space, perhaps the resemblance is not only superficial. The propagation of ideas between the two strongly bimodal research streams are vital to achieving good mixing, both in terms of the metaphor and also with respect to the technologies developed. The search for a family of sampling algorithms that may be adopted for use in a broad range of modelling situations is a generic and ongoing challenge. However it is readily observed that analyses in genetic-based disciplines, by their very nature, generate some of the greatest obstacles to be faced in terms of model complexity and good sampler efficiency. Our belief is that only through continued development within both genetical and statistical literatures will these goals be achieved. Genetics and stochastic simulation *do* mix — but proposals from both research streams are necessary for future progress.

Acknowledgments

The author would like to thank M. Tanaka for helpful conversations and C-code used in performing the simulation in Section 2.4, and an Assistant Editor and two anonymous referees for their suggestions on a previous version of this article. The author is supported by the Australia Research Council through the Discovery Project scheme (DP0664970) and by the Faculty of Science, UNSW.

References

Balding, D. J., M. Bishop, and C. Cannings (Eds.) (2003). *Handbook of Statistical Genetics (2nd Edition)*. John Wiley and Sons, Ltd.

Beaumont, M. A., W. Zhang, and D. J. Balding (2002). Approximate Bayesian computation in population genetics. *Genetics 162*, 2025 – 2035.

Bortot, P., S. G. Coles, and S. A. Sisson (2006). Inference for stereological extremes. *Journal of the American Statistical Association. To appear*.

Brooks, S. P., Y. Fan, and J. S. Rosenthal (2006). Perfect forward simulation via simulated tempering. *Communications in Statistics: Simulation and Computation 35*.

Brooks, S. P., P. Guidici, and G. O. Roberts (2003). Efficient construction of reversible jump Markov chain Monte Carlo proposal distributions. *Journal of the Royal Statistical Society, B 65*, 3 – 39.

Cannings, C., E. A. Thompson, and M. H. Skolnick (1978). Probability functions on complex pedigrees. *Advances in Applied Probability 10*, 26 – 61.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician 49*, 327 – 335.

Chikhi, L. and M. Bruford (2005). Mammalian population genetics and genomics. In A. Ruvinsky and J. M. Graves (Eds.), *Mammalian Genomics*, Chapter 11, pp. 539 – 584. CAB International.

Edwards, A. W. F. (1968). Simulation studies of genealogies (presented paper abstract). *Heredity*, 628.

Estoup, A., M. Beaumont, F. Sennedot, C. Moritz, and J.-M. Cornuet (2004). Genetic analysis of complex demographic scenarios: spatially expanding populations of the cane toad, *bufo marinus*. *Evolution 58*, 2021–2036.

Fu, Y.-X. and W.-H. Li (1997). Estimating the age of the common ancestor of a sample of DNA sequences. *Molecular Biology and Evolution 14*, 195 – 199.

Gelfand, A. E. and A. F. M. Smith (1990). Sampling-based approaches to calculating marginal densities. *Journal of the Amererican Statistical Association 85*, 398 – 410.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (Eds.) (2004). *Bayesian Data Analysis (2nd Edition)*. Chapman and Hall/CRC.

Geyer, C. J. (1991). Markov chain Monte Carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface*, pp. 156 – 163.

Geyer, C. J. and E. A. Thompson (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association 90*, 909 – 920.

Gilks, W. R., S. Richardson, and D. J. Spiegelhalter (Eds.) (1995). *Markov Chain Monte Carlo in Practice*. Chapman and Hall, London.

Green, P. J. (1995). Reversible jump MCMC computation and Bayesian model determination. *Biometrika 82*, 711 – 732.

Green, P. J. (2001). In O. E. Barndorff-Nielsen, D. R. Cox, and C. Klüppelberg (Eds.), *Complex Stochastic Systems*, Number 87 in Monographs on Statistics and Probability, Chapter A primer on Markov chain Monte Carlo, pp. 1 – 62. Chapman and Hall/CRC.

Hamilton, G., M. Currat, N. Ray, G. Heckel, M. Beaumont, and L. Excoffier (2005). Bayesian estimation of recent migration rates after a spatial expansion. *Genetics 170*, 409–417.

Hastings, W. K. (1970). Monte carlo sampling methods using Markov chains and their applications. *Biometrika 57*, 97 – 109.

Hurn, M. A., H. Rue, and N. Sheehan (1999). Block updating in constrained Markov chain Monte Carlo sampling. *Statistics and Probability Letters 41*, 353 – 361.

Jensen, C. S. and A. Kong (1999). Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *American Journal of Human Genetics 65*, 885 – 901.

Jensen, C. S. and N. A. Sheehan (1998). Problems with the determination of non-communicating classes for Monte Carlo Markov chain applications in pedigree analysis. *Biometrics 54*, 416 – 425.

Lander, E. S. and P. Green (1987). Construction of moltilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences 84*, 2363 – 2367.

Liang, F. and W. H. Wong (2000). Evolutionary Monte Carlo sampling: Applications to c_p model sampling and change-point problems. *Statistica Sinica 10*, 317 – 342.

Lin, S. (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics 51*, 318 – 322.

Lin, S., E. A. Thompson, and E. Wijsman (1993). Achieving irreducibility of the Markov chain Monte Carlo method applied to pedigree data. *IMA Journal of Mathematics Applied in Medicine and Biology 10*, 1 – 17.

Marjoram, P., J. Molitor, V. Plagnol, and S. Tavaré (2003). Markov chain Monte Carlo without likelihoods. *PNAS 100*, 15324 – 15328.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics 21*, 1087 – 1091.

Møller, J. and G. K. Nicholls (1999). Perfect simulation for sample-based inference. Technical report, Aalborg University.

Pritchard, J. K., M. T. Seielstad, A. Perez-Lezaun, and M. W. Feldman (1999). Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Molecular Biology and Evolution 16*, 1791 – 1798.

Propp, J. G. and D. B. Wilson (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms 9*, 223 – 252.

Robert, C. and G. Casella (2004). *Monte Carlo Statistical Methods* (2nd ed.). Springer, New York.

Sheehan, N. and A. Thomas (1993). On the irreducibility of a Markov chain defined on a space of genotypic configurations by a sampling scheme. *Biometrics 49*, 163 – 175.

Sisson, S. A. (2002). An algorithm to characterise non-communicating classes on complex geneaolgies. Technical report, Department of Statistics, University of New South Wales.

Sisson, S. A. (2005). Trans-dimensional Markov chains: A decade of progress and future perspectives. *Journal of the American Statistical Association. 100*, 1077 – 1089.

Sisson, S. A., Y. Fan, and M. M. Tanaka (2006). Sequential Monte Carlo without likelihoods. Technical report, Department of Statistics, University of New South Wales.

Sorensen, D. and D. Gianola (2002). *Likelihood, Bayesian and MCMC Methods in Quantitative Genetics*. Springer-Verlag, New York.

Tanaka, M. M., A. R. Francis, F. Luciani, and S. A. Sisson (2006). Using approximate Bayesian computation to estimating tuberculosis transmission parameters from genotype data. *Genetics. In press*.

Tavaré, S. D., D. J. Balding, R. C. Griffiths, and P. Donnelly (1997). Inferring coalescence times from DNA sequence data. *Genetics 145*, 505 – 518.

Thompson, E. A. (2000). *Statistical Inference from Genetic Data on Pedigrees*. Institute of Mathematical Sciences and the American Statistical Association.

Waagepetersen, R. and D. Sorensen (2001). An tutorial on reversible jump MCMC with a view toward applications on QTL mapping. *International Statistical Review 69*, 49 – 62.

Weiss, G. and A. von Haeseler (1998). Inference of population history using a likelihood approach. *Genetics 149*, 1539 – 1546.

Wright, S. and H. C. McPhee (1925). An approximate method of calculating coefficients of inbreeding and relationship from livestock pedigrees. *Journal of Agricultural Research 31*, 377 – 383.