

IMPROVED MCMC METHODS FOR SOME POPULATION GENETIC MODELS

M-1556

July 2000

Rasmus Nielsen

Keywords: Markov Chain Monte Carlo, DNA sequence evolution, population genetics.

Abstract:

We present a new approach to the estimation of likelihood surfaces from a single run of a Markov chain in some Markov Chain Monte Carlo (MCMC) methods. It alleviates the need to multiple chains applied in some previous applications of MCMC in population genetics. We also present a new MCMC method applicable to DNA sequence data, based on data augmentation. In this method, mutations in the geneology are treated as missing data. this method facilitates inferences regarding the age and identity of specific mutations while taking the full complexities of the mutational process in DNA sequences into account.

JULY 2000

1153

1556

IMPROVED MCMC METHODS FOR SOME POPULATION GENETIC

MODELS

Rasmus Nielsen

Dept. of Biometrics
439 Warren Hall
Cornell University
Ithaca, NY 14853-7801

Summary

We present a new approach to the estimation of likelihood surfaces from a single run of a Markov chain in some Markov Chain Monte Carlo (MCMC) methods. It alleviates the need for multiple chains applied in some previous applications of MCMC in population genetics. We also present a new MCMC method applicable to DNA sequence data, based on data augmentation. In this method, mutations in the genealogy are treated as missing data. This method facilitates inferences regarding the age and identity of specific mutations while taking the full complexities of the mutational process in DNA sequences into account.

Key words: Markov chain Monte Carlo; DNA sequence evolution; Population Genetics

1. Introduction

There has recently been a dramatic increase in the use of Markov Chain Monte Carlo (MCMC) methods in population genetics (e.g. Kuhner *et al.*(1995), Kuhner *et al.*(1998), Beerli and Felsenstein (1999), Wilson and Balding (1998). The objective of these methods is to estimate parameters of the demographic process in the population from which the sample has been obtained or of the mutational process in the genetic data analyzed. For example, in the method of Kuhner *et al.*(1995), the objective is to make inferences on the fundamental population genetical parameter $\theta = 4N_e\mu$, where N_e is the effective inbreeding population size (Wright 1931) and μ is the mutation rate per site in the analyzed molecular marker. For most mutational models, it is only possible to obtain the likelihood function by conditioning on the underlying gene genealogy, i.e. the likelihood is obtained as

$$L(\Theta | X) = \int \Pr(X | G, \Theta) dP_{\Theta}(G), \quad (1)$$

where Θ is the vector of parameters, X is the observed genetic data, G denotes the gene genealogy and $P_{\Theta}(G)$ is the probability distribution of G given Θ . It is necessary to take account of the gene genealogy because it summarizes information regarding the correlation among individuals in the population due to shared common ancestry.

A major breakthrough in population genetics was achieved when it was demonstrated how to derive distributions of gene genealogies, $P_{\Theta}(G)$, from classical population genetical models (Kingman 1982a,b). In brief, the ancestry of a sample of n genes obtained from a population of size N_e is considered. In its most simple form it is assumed that individuals in the population are sampled randomly and are mating randomly and that the population is of

constant size with no population structure. In the limit of $N_e \rightarrow \infty$, a coalescence process then arises (Kingman 1982a,b) that allow probabilities to be assigned to genealogies. A genealogy (G) is here a labeled history in the sense of Thompson (1975) and it can be described by a tree where each leaf is associated with one of the sampled haplotypes (chromosomes). It consists of a topology [of which there are $(n!)^2/(n2^{n-1})$] and a set of coalescence times $\tau = \{\tau_2, \dots, \tau_n\}$, where τ_i is the time in G in which there are i ancestors of the sample, i.e. the length of the edges in the genealogy are proportional to the time the genes have diverged from each other (Figure 1). Here and in the following, time is measured in number of generations scaled by the mutation rate μ . For example, for a class of neutral population genetical models in which the genes segregating in the population are exchangeable, and the distribution offspring number is constant in time

$$dP_\theta(G) = \left(\frac{2}{\theta}\right)^{n-1} \exp\left[\sum_{i=2}^n -\frac{i(i-1)\tau_i}{\theta}\right] d\tau \quad (2)$$

(Kingman 1982a, Felsenstein 1992). Remarkably, this distribution of genealogies arises both from the neutral Wright-Fisher model (Fisher 1930, Wright 1931), the Moran model (Moran 1962) and any other exchangeable models, i.e. models that assumes that all individuals in the population have the same constant distribution of off-spring numbers. By relaxing this assumption it is also possible to analyze models that include population structure, changes in population size and selection. However, in the following we will just concentrate on the estimation of the parameter θ in the model given by (2).

In these models $\Pr(X | G, \Theta)$ is calculated by superimposing a model of mutation on the gene genealogy. The mutation models are usually time reversible continuous time Markov

chains. For example, for DNA sequence data, models such as the F84 model (Felsenstein 1984) may be appropriate. In this model the substitutional process in each site along an edge of the genealogy is modeled as a continuous time Markov chain. The infinitesimal generator (Q) is given by a 4x4 matrix with off-diagonal elements

$$q_{ij} = \begin{cases} \lambda[1 + \kappa / (\pi_A + \pi_G)]\pi_j & \text{if } i, j \in \{A, G\} \\ \lambda[1 + \kappa / (\pi_C + \pi_T)]\pi_j & \text{if } i, j \in \{C, T\}, \\ \lambda\pi_j & \text{otherwise} \end{cases} \quad (3)$$

and diagonal elements determined by the mathematical requirement that the row sums should be zero. π_j is the stationary frequency of nucleotide j (there are 4 nucleotides: A, C, T and G) and λ is chosen such that

$$-\sum_i \pi_i q_{ii} = 1. \quad (4)$$

Then, the transition probabilities of this Markov chain along an edge of length t in the genealogy is given by $P(t) = \{p_{ij}(t)\} = e^{Qt}$. In the case of the F84 model these transition probabilities can be calculated analytically and can be found in Kishino and Hasegawa (1989). $\Pr(X | G, \Theta)$ can then be calculated for any G assuming stationarity of the process by summing over the ancestral states at each node of the genealogy. Because the length of the edges in the gene genealogy are scaled by θ , $\Pr(X | G, \Theta)$ does not depend on θ .

In models, such as the finite state space DNA models, evaluation of (1) is only possible by simulation techniques for realistic sized data sets. Several approaches to this problem have been published, the most successful being the method by Kuhner *et al.* (1995). The objective in this method is to estimate the single parameter θ in model (2). In other applications Θ may contain parameters regarding migration between populations (Beerli and

Felsenstein 1999) or population growth (Kuhner *et al.* 1998). However, the models are always parameterized such that $\Pr(X | G, \Theta) = \Pr(X | G)$ in this framework. A Markov chain with state space on G and stationary distribution proportional to $\Pr(X | G)P_{\Theta}(G)$ is established using the Metropolis-Hastings method (Metropolis 1953, Hastings 1970). Updates to G are proposed according to proposal kernel $h(G, G')$ and these updates are accepted with probability

$$\min\{1, r\}, \quad r = \frac{\Pr(X | G')P_{\Theta_0}(G')h(G', G)}{\Pr(X | G)P_{\Theta_0}(G)h(G, G')}. \quad (5)$$

Under assumptions of ergodicity, this Markov chain has the desired stationary distribution (e.g. Hastings 1970, Ripley 1987, Tierney 1996).

The likelihood surface for multiple values of the parameter Θ are obtained by running a Markov chain at a fixed value Θ_0 close to the mode of the likelihood function while evaluating the likelihood for multiple value of Θ using importance sampling (Thompson and Guo, 1991). Assuming $P_{\Theta_0}(G)$ dominates $P_{\Theta}(G)$ for all Θ

$$\Pr(X | \Theta) = \int \Pr(X | G)w(\Theta, \Theta_0, G)dP_{\Theta_0}(G), \quad w(\Theta, \Theta_0, G) = \frac{p_{\Theta}(G)}{p_{\Theta_0}(G)}. \quad (6)$$

The likelihood function for Θ can, therefore, be evaluated by sampling n values of $w(\Theta, \Theta_0, G)$ from a Markov chain with stationary distribution of G proportional to $\Pr(X | G)P_{\Theta_0}(G)$ as

$$\frac{L(\Theta | X)}{L(\Theta_0 | X)} \approx \frac{1}{k} \sum_{i=1}^k w_i(\Theta, \Theta_0, G^{(i)}), \quad (7)$$

where k is the number of sampled steps in the Markov chain and $w_i(\Theta, \Theta_0, G)$ and $G^{(i)}$ are the value of $w(\Theta, \Theta_0, G)$ and G , respectively, in the i th sampled step of the chain. The

symbol \approx means here approximated by and is used if the expression on the right hand side converges to the expression on the left hand side by a law of large numbers.

The method of Kuhner *et al.*(1995) provided a major breakthrough in the analysis of genetic data. It demonstrated that likelihood inference is possible for complex mutational models, provided the first true MCMC method in population genetics and lead to the subsequent adoption of similar methods by Wilson and Balding (1998), Beaumont (1999) and Nielsen (2000). It also preceded, and possibly inspired, the use of closely related MCMC methods in the field of phylogenetic inference (Yang and Rannala 1997, Larget and Simon 1999, Mau *et al.*1999). However, some of the details of the method may be improved. Most importantly, it has been suggested in the literature that the importance sampling scheme (equation 6) may not be very reliable. It was in Nielsen (2000) found to perform poorly except when the value of Θ is very close to the value of Θ_0 and was also criticized in Stephens (1999). The problem is that the distribution of $w(\Theta, \Theta_0, G)$ will be very skewed when Θ differs from Θ_0 . The variance in the estimate of the likelihood will therefore be very large and difficult to estimate when $|\Theta - \Theta_0|$ is large. Because of the skewed distribution of $w(\Theta, \Theta_0, G)$, the method will tend to underestimate the likelihood when Θ differs from Θ_0 . The effect is to bias the estimator towards values close to Θ_0 . Stephens (1999) points out, that the under some very simple conditions, the variance of the $w_i(\Theta, \Theta_0, G)$ may be infinite when $\theta > 2\theta_0$.

Geyer (1991) suggests running multiple chains at different values of Θ_0 to overcome this problem. The global estimate of the likelihood function can be estimated using the results from all chains by reweighting the results from each chain according to the likelihood. In Kuhner *et al.* (1995), the problem of large variance when $|\Theta - \Theta_0|$ is large, is addressed by

running multiple chains and updating Θ_0 each time the chain is restarted, i.e. a type of optimization based on multiple chains. This is an approach suggested in Geyer and Thompson (1992). The problems with this approach is that correct implementation may be very computationally intensive because it requires multiple chains reaching stationarity. Even if the estimator is guaranteed to converge to the MLE, the associated likelihood surface may not be efficiently estimated. Also, this method does not solve the fundamental problems, particular to some population genetical models, raised in Stephens (1999).

The alternative approach in Wilson and Balding (1998) and Nielsen (2000) was to assume a uniform prior distribution of Θ and approximating the likelihood function by the distribution of values of Θ sampled from a Markov chain with stationary distribution of Θ and G proportional to $P(\Theta, G | X)$. This is the method most often used in Bayesian applications of MCMC. One problem with this approach, is that the approximation to the resulting posterior distribution/likelihood function will not be smooth when only a limited number of samples from the Markov chain have been obtained. It is therefore necessary to use non-parametric smoothing methods to obtain maximum likelihood estimates based on this approach. In addition, convergence of the ergodic averages may be slow and the shape of the estimated likelihood surface may depend on the details of the implementation of the smoothing method.

In the following, we present a new MCMC based method for evaluation of the likelihood function that always produces smooth likelihood surfaces based on a single run of a Markov chain. The method will be evaluated and compared to method (6-7) by the application to a simulated and a real data set.

We also present a new MCMC method applicable to finite sites models of DNA evolution based on data augmentation (Tanner and Wong 1987). This method is based on a

mapping of mutations on the genealogy, i.e. we treat the mutations as missing data. In many biological applications there has been a strong interest in making inferences regarding the identity of specific mutations. For example, there has recently been an interest in methods for estimating the age of specific mutations (e.g. Slatkin and Rannala, 1997, Nielsen and Weinreich 1999). Treating the mutations as missing data facilitates Bayesian inferences regarding specific mutations. Also, in models of sequence evolution where the transition probabilities cannot be calculated analytically, such as the models considered by Goldman and Yang (1994) and Nielsen and Yang (1998), treating the mutations as missing data may lead to a marked speed up of the MCMC method.

2. An improved MCMC method

The strategy for evaluating likelihood surfaces we will use in the following is to use a pseudoprior for Θ_0 . This will allow us to evaluate the likelihood function for multiple values of Θ simultaneously, while weighting the contribution of $w_i(\Theta, \Theta_0^{(i)}, G)$ to $L(\Theta | X)$ according to the distance $|\Theta - \Theta_0^{(i)}|$, where $\Theta_0^{(i)}$ is the value of Θ_0 at the i th step of the chain. This approach can be used to generate smooth reliable likelihood surfaces in a single run of the Markov chain. The likelihood function can be written as

$$L(\Theta | X) = \frac{\int \int \Pr(X | G) r(\Theta, \Theta_0) w(\Theta, \Theta_0, G) dP_{\Theta_0}(G) dP(\Theta_0)}{\int r(\Theta, \Theta_0) dP(\Theta_0)}. \quad (8)$$

[compare to (6)]. $r(\Theta, \Theta_0)$ is a weighting function, and is used to down weight values of $w(\Theta, \Theta_0, G)$ for which $|\Theta - \Theta_0|$ is large. The only condition on it is that it must be chosen such that $\int r(\Theta, \Theta_0) dP(\Theta_0) < \infty$. The likelihood function for Θ can therefore be evaluated

by simulating a Markov chain with stationary distribution of G and Θ_0 proportional to $\Pr(X | G)P_{\Theta_0}(G)P(\Theta_0)$ and sample n values of $r(\Theta, \Theta_0)w(\Theta, \Theta_0, G)$ form this chain, so that

$$L(X | \Theta) \approx c \frac{\sum_{i=1}^n r_i(\Theta, \Theta_0^{(i)})w_i(\Theta, \Theta_0^{(i)}, G^{(i)})}{n \int r(\Theta, \Theta_0)dP(\Theta_0)}, \quad (9)$$

where c is a scaling factor equal to $\Pr(X) = \int \int \Pr(X | G)dP(\Theta_0)dP(G | \Theta_0)$. The estimation of $L(X | \Theta)$ can be done simultaneously using one chain for many values of Θ , and the scaling factor c needs not be estimated. The advantage of this method is that it allows the weight, $r(\Theta, \Theta_0)$, to be specified for any Θ such that the highest weight is given to values of Θ_0 close to Θ . For example, it was found in preliminary studies that for the population genetical problems studied here, weights of the form $r(\Theta, \Theta_0) = e^{-((\Theta_0 - \Theta)^2 / s)}$, $\theta < 2\theta_0$, allowed relative fast convergence over a wide range of values.

The prior distribution $P(\Theta_0)$ can be any desired distribution. It determines how much computational time is spend evaluating the likelihood at particular values of Θ . For example, if a uniform distribution is chosen, the time the chain spends at a particular value of Θ is proportional to the likelihood function for Θ . This may in many cases be desirable because it ensures that most accuracy in the estimate is obtained near the mode of the likelihood function. It is therefore the approach chosen in the following.

3. Mutations as missing data

Notice that he likelihood can be written as

$$L(\Theta | X) = \int \int I(X, G, \eta)dP_{G, \Theta}(\eta)dP_{\Theta}(G), \quad (10)$$

where η is an assignment of a set of mutations to G , $P_{G, \Theta}(\eta)$ is the probability distribution of η on G given Θ , and $I(X, G, \eta)$ is an indicator function that returns 1 if η on G is compatible

with X and θ otherwise. η consists of a vector of mutations for each edge in G , in which mutations are labeled with respect to type (e.g. $G \rightarrow T$) and time. $P_{G,\theta}(\eta)$ can then easily be calculated for any model such as (3).

A Markov chain with stationary distribution proportional to $L(\Theta | X)$ can be established by proposing updates to η and G according to some proposal kernel $h[(\eta, G), (\eta', G')]$ such that $I(X, G, \eta) = 1$. A proposed update is then accepted with probability

$$\min(1, r), \quad r = \frac{P_{\Theta}(G')P_{G',\theta}(\eta')h[(\eta', G'), (\eta, G)]}{P_{\Theta}(G)P_{G,\theta}(\eta)h[(\eta, G), (\eta', G')]} . \quad (11)$$

The proposal algorithm used here is based on updating each edge of the genealogy one at a time. In each update, the length of the edge, and potentially also the topology of the tree is updated simultaneously with the assignment of mutations to the updated edge. Mutations on the new edge are simulated under the condition $I(X, G, \eta) = 1$ using a fast approximation

based on a Poisson process that enables easy evaluation of $\frac{P_{G',\theta}(\eta')h[(\eta', G'), (\eta, G)]}{P_{G,\theta}(\eta)h[(\eta, G), (\eta', G')]}$. It is

designed to be efficient when the expected number of mutations per site in an edge is small and the mutation process is approximately Poisson. It can be described by the following algorithm:

- (1) Choose an edge uniformly among all edges in the tree.
- (2) Move the time at which the edge connects to its parent edge a Gaussian distributed distance with mean 0 and variance σ^2 . Truncate such that the new length of the edge (t) is positive. When a vertex is encountered, continue to move the edge along either of the two other edges connecting to the vertex, each with probability 0.5.

- (3) For each site in the new edge, determine the ancestral state (A) and the endstate (E) under the condition $I(X, G, \eta) = 1$. If $A = E$ simulate a Poisson distributed number of mutations with rate $-q_{AA}t$ conditional on not observing exactly one mutation. If $A \neq E$ simulate a Poisson distributed number of mutations with rate $-q_{AA}t$ conditional on observing more than zero mutations.
- (4) Distribute the mutations on the edge according to the relative rates given by q , but set $q_{iE} = 0$ for all i for the second to last mutation in the edge, and $q_{ij} = 0$ for all $j \neq E$ for the last mutation in the site.

Notice, that updates to the topology are achieved at step 2 at the same time the length of the edge is updated. Also notice, that in the present representation, A and E will be given for any site in an edge because of the condition $I(X, G, \eta) = 1$. Then,

$$\frac{P_{G', \Theta}(\eta')}{h[(\eta, G), (\eta', G')]} = \prod_s z_s f_s \exp(u_s), \quad (12)$$

where z_s, f_s and u_s are the functions z, f and u evaluated in site s and

$$u = \begin{cases} q_A t_1 + \sum_{i=2}^k q_{b_i} (t_i - t_{i-1}) + q_{b_k} (t - t_k) & \text{if } k > 0 \\ q_A t & \text{if } k = 0 \end{cases},$$

$$f = \begin{cases} \exp(q_A t) - q_A t & \text{if } A = E \\ \exp(q_A t) - 1 & \text{if } A \neq E \end{cases},$$

$$z = \begin{cases} 1 & \text{if } k = 0 \\ q_{b_k E} / q_A & \text{if } k = 1 \\ q_{b_k E} \sum_{j: j \neq E \wedge j \neq b_{k-1}} q_{b_{k-1} j} / q_A^2 & \text{if } k > 1 \end{cases}, \quad (13)$$

where k is the number of mutations assigned to the site, t is the length of the edge, t_i is the time of the i th mutation on the edge since the origin of the edge and b_i is the nucleotide

resulting from the i th mutation in the site. $\frac{P_{G,\theta}(\eta)}{h[(\eta', G'), (\eta, G)]}$ can be found similarly and since

$\frac{P_{\theta}(G')}{P_{\theta}(G)}$ easily can be evaluated from (2) under the assumption of a standard neutral

coalescence model, the Metropolis-Hastings ratio can easily be evaluated under this proposal distribution. The value of σ^2 will determine the magnitude of the proposed updates and can be adjusted to guarantee an appropriate intermediate ratio of accepted to rejected updates (e.g. Gelman *et al.* 1995).

4. Evaluation of the methods

To check the computer implementation of the new methods, results were compared to the results obtained using numerical integration for very small data sets and to the results obtained using the program described in Kuhner *et al.* (1995) for data sets of moderate size. In all cases, close agreement between the methods was found.

The performance of the new methods was evaluated by repeated analysis of a real and a simulated data set was performed. The real data set consists of a 360 nucleotide long region of 63 chromosomes from the Nuu-Chah-Nulth tribe. It was published by Ward *et al.* (1991) and was also used in Kuhner *et al.* (1995) to evaluate their method. The simulated data is a large data set consisting of a 500 nucleotide long region from 200 chromosomes.

The real data set was analyzed assuming a prior distribution of genealogies given by (2) and a model of nucleotide mutation given by (3) assuming $\kappa = 113.036352$ and using the empirical nucleotide frequencies as estimates of π_A , π_C , π_G and π_T . This corresponds to the value chosen in Kuhner *et al.* (1995) when evaluating their method. 10 likelihood surfaces for θ were generated based on (9) and (10) using 200,000 steps in the Markov chain and a burn-in time of 20,000 steps. The starting value of θ_0 was chosen uniformly in the interval $[0, 0.1]$

and a value of $s = 0.01$ was used. The estimation of each likelihood surface takes approximately 5 minutes on a 450 Mhz Pentium II machine.

Notice (Figure 2), that the likelihood surfaces obtained from different runs are roughly similar. The inferences made regarding θ from each run of the Markov chain would not vary too much and the approximate maximum likelihood estimates would be almost identical. However, there is some variability among chains, which could be decreased by increasing the run-time.

Likelihood surfaces were also estimated using (7) and (10), i.e. the method using a fixed value of θ_0 . Three value of θ_0 was used: 0.01, 0.04 and 0.4. Five likelihood surfaces were generated for each value. Notice (Figure 3) that the inferences made differ strongly between runs. In particular, the approximate MLE is strongly dependent on θ_0 . The runs for which $\theta_0 = 0.01$ provide approximate MLEs of θ less than 0.02. The runs for which $\theta_0 = 0.4$ all provide MLEs larger than 0.06. As expected, the method underestimates θ when $\theta_0 < \theta$ and overestimates θ when $\theta_0 > \theta$. Also notice that the different runs using the same value of θ_0 are disturbingly similar. It may therefore be very deceptive to run multiple chains using the same value of θ_0 as a method for evaluating the performance of the method.

The large data set (200 chromosomes, 5000 nucleotides) was simulated under model (2) and (3) assuming $\theta = 0.02$, $\pi_A = \pi_T = \pi_G = \pi_C = 0.25$ and $\kappa = 0.0$. This corresponds to the familiar J-C model of nucleotide mutation (Jukes and Cantor 1969). Ten likelihood surfaces based on (9) and (10) were obtained using 2,000,000 steps of the chain and a burn-in time of 200,000. Notice (Figure 4), that the likelihood surfaces again are similar. Roughly the same inferences would be done from each run of the chain. However, there is still some variability among chains, demonstrating that not even 2,000,000 runs of the chain are quite sufficient for convergence for such a large data set.

Estimates of the likelihood surface based on (7) and (10), i.e. the method using a fixed value of θ_0 , were also obtained (Figure 5). Notice again that the likelihood is strongly dependent on the value of θ_0 .

5. Discussion

The new method for estimating likelihood surfaces provides a viable alternative to previous methods, e.g. Thompson and Guo (1991), Kuhner *et al.*(1995). Likelihood surfaces based on (7) are in general unreliable. However, it should be noted that the practice of running multiple chains used by Kuhner *et al.*(1995) alleviates much of the problems with the method discussed here. Nonetheless, the current method should be more efficient and easy to implement because it allows inferences to be done using a single run of the Markov chain.

The MCMC method based on data augmentation (10) will find applications when inferences regarding ages or identities of specific mutations are of importance. Also, in models where the transition probabilities of the nucleotide mutation model cannot be calculated analytically, it may provide a marked improvement in computational speed. These issues will be addressed in future studies.

References

- Beaumont, M. A. (1999). Detecting Population Expansion and Decline Using Microsatellites. *Genetics* **153**, 2013-2029.
- Berli, P. and Felsenstein, J. (1999) Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. *Genetics* **152**, 763-773.
- Felsenstein, J. (1984). DNAML, computer program. Distributed from <http://evolution.genetics.washington.edu>.

- Felsenstein, J. (1992). Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genetical Research* **60**, 209-220.
- Fisher, R. A. (1930). *The genetical theory of natural selection*, 1st Ed. Clarendon press, Oxford.
- Gelman, A., Roberts, G. O. and Gilks, W. R. (1995). Efficient Metropolis jumping rules. In *Bayesian Statistics 5* (eds. Bernardo, J. M., Berger, J. O., Dawid, A. P. and Smith, A. F. M.). Oxford University Press, Oxford, UK.
- Geyer, C. J. (1991). Reweighting Monte Carlo mixtures. Technical Report 568, School of Statistics, Univ. Minnesota.
- Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with discussion). *Journal of the Royal Statistical Society series B.* **54**, 567-699.
- Goldman, N., and Yang, Z. (1994). A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Molecular Biology and Evolution.* **11**, 725-736.
- Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97-109.
- Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. Pp 21-123 in Munro, H. N. ed. *Mammalian protein metabolism*, Academic Press, New York.
- Kingman, J. F. C. (1982a). The coalescent. *Stochastic Processes Applications.* **13**, 235-248.
- Kingman, J.F.C. (1982b). On the genealogy of large populations. *Journal of Applied Probability* **19**, 27-43.
- Kishino, H. and Hasegawa, M. (1989.) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of Hominoidea. *Journal of Molecular Evolution* **31**, 151-160.

- Kuhner, M. K., Yamato, J. and Felsenstein, F. (1995). Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. *Genetics* **140**, 1421-1430.
- Kuhner, M. H., Yamato, J. and Felsenstein, J. (1998). Maximum likelihood estimation of population growth rates based on the coalescent *Genetics* **149**, 429-434.
- Larget B. and Simon D. L. (1999). Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Molecular Biology and Evolution*. **16**,750-759.
- Mau B., M. A. Newton and Larget, B. (1999). Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*. **55**,1-12
- Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller. E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **21**, 1087-1091.
- Moran, P. A. P. (1962). *The statistical processes of evolutionary theory*. Clarendon press, Oxford.
- Nielsen, R. (2000). Estimation of Population Parameters and Recombination Rates from Single Nucleotide Polymorphisms (SNPs). In press, *Genetics*.
- Nielsen R. and Weinreich, D. M. (1999). The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics*. **153**, 497-506
- Nielsen, R. and Yang, Z. (1998). Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* **148**, 929-936.
- Ripley, B. (1987). *Stochastic simulation*. New York. Wiley.
- Slatkin, M and Rannala, B. (1997). Estimating the age of alleles by use of intraallelic variability. *American Journal of Human Genetics* **60**, 447-458.

- Stephens, M. (1999). Problems with computational methods in population genetics.
Contribution to the 52nd session of the International Statistical Institute, August 1999.
Available from <http://www.stats.ox.ac.uk/~stephens/group/publications.html>.
- Tanner, M. A. and Wong, W. H.. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*. **83**, 528-540.
- Tierney, L. (1996). Introduction to general state space Markov chain theory. In *Markov Chain Monte Carlo in Practice* (eds. W. R. Gilks, S. Richardson and D. J. Spiegelhalter). Chapman and Hall, London.
- Thompson, E. A. (1975). *Evolutionary trees*. Cambridge University Press, Cambridge, UK.
- Thompson, E. A. and Guo, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA Journal of Mathematical Applications in Medicine and Biology*. **8**, 149-169.
- Ward, R. H., Frazier, B. L., Dew-Jager, K. and Pääbo, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proceeding of the National Academy of Sciences USA* **88**, 8720-8724.
- Wilson, I. J. and Balding, D. J. (1998). Genealogical inference from microsatellite data. *Genetics* **150**, 499-510.
- Wright, S. (1931). Evolution in Mendelian populations. *Genetics* **16**:97-159.
- Yang, Z. H. and Rannala, B. (1997). Bayesian phylogenetic inference using DNA sequences - a Markov Chain Monte Carlo method. *Molecular Biology and Evolution*. **14**, 717-724.

Figure Legends

Figure 1. An example of a gene genealogy for five chromosomes.

Figure 2. Ten estimates of the likelihood surface for the Ward et al (1991) data set based on (9) and (10).

Figure 3. Fifteen estimates of the likelihood surface for the Ward *et al.* (1991) data set based on (7) and (10) using $\theta_0 = 0.01$ (solid lines), $\theta_0 = 0.04$ (dotted lines) and $\theta_0 = 0.4$ (striped lines).

Figure 4. Ten estimates of the likelihood surface for the large simulated data set based on (9) and (10).

Figure 5. Fifteen estimates of the likelihood surface for the large simulated data set based on (7) and (10) using $\theta_0 = 0.005$ (solid lines), $\theta_0 = 0.02$ (dotted lines) and $\theta_0 = 0.2$ (striped lines).

Figure 1

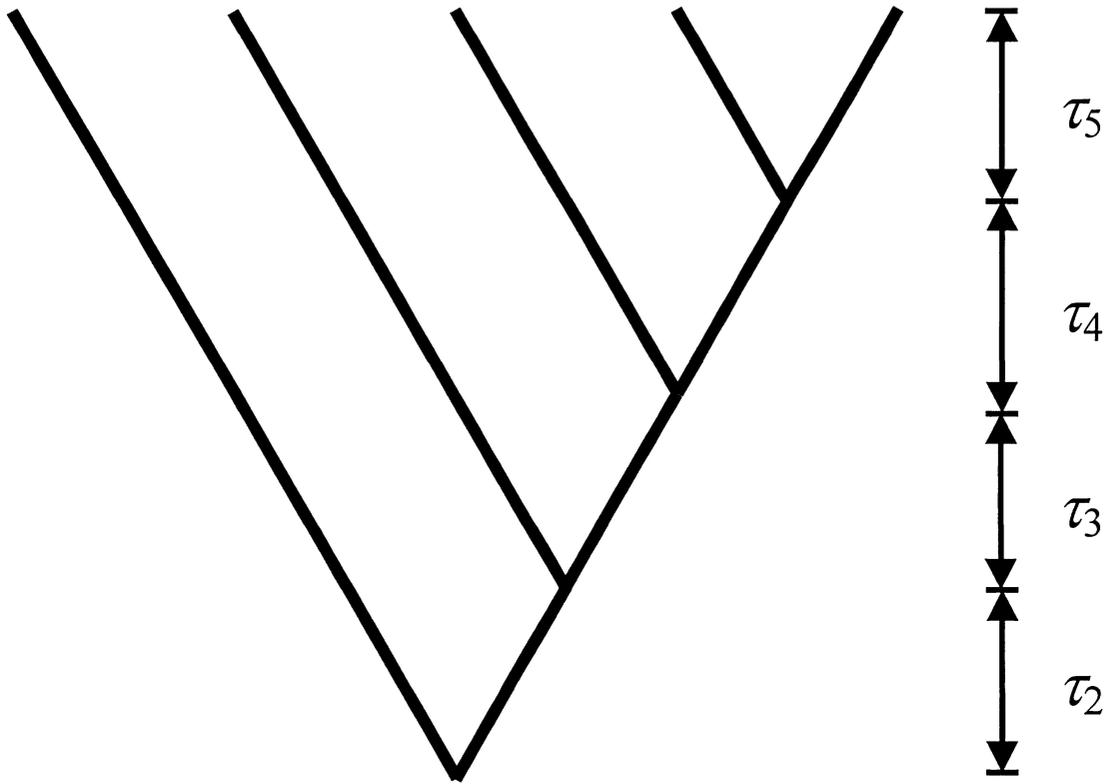


Figure 2

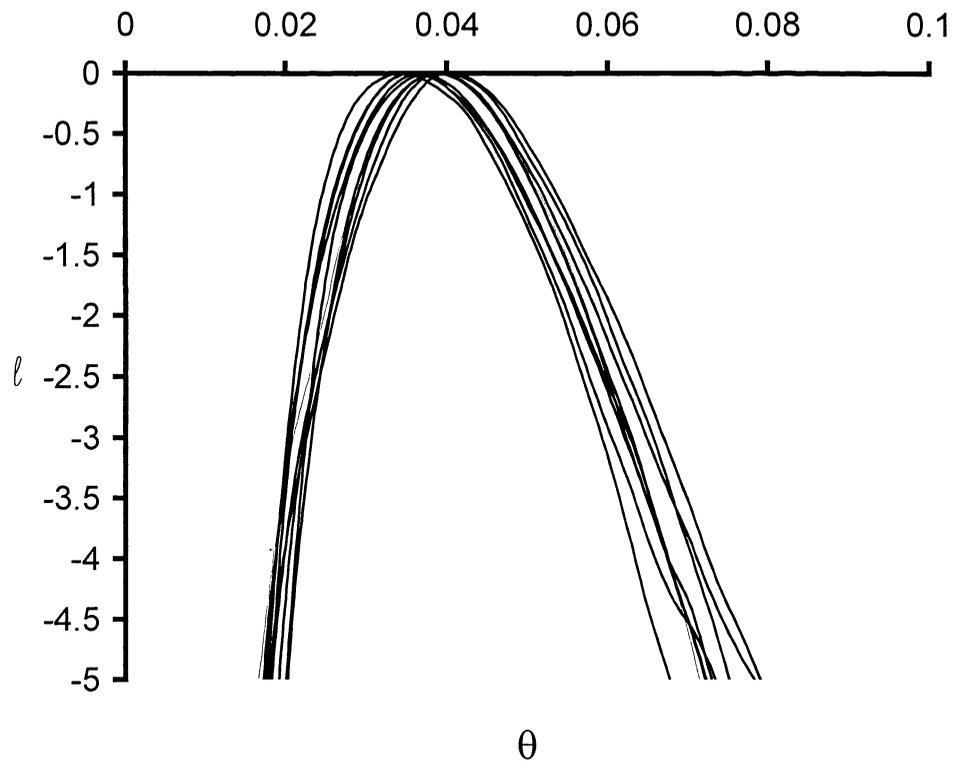


Figure 3

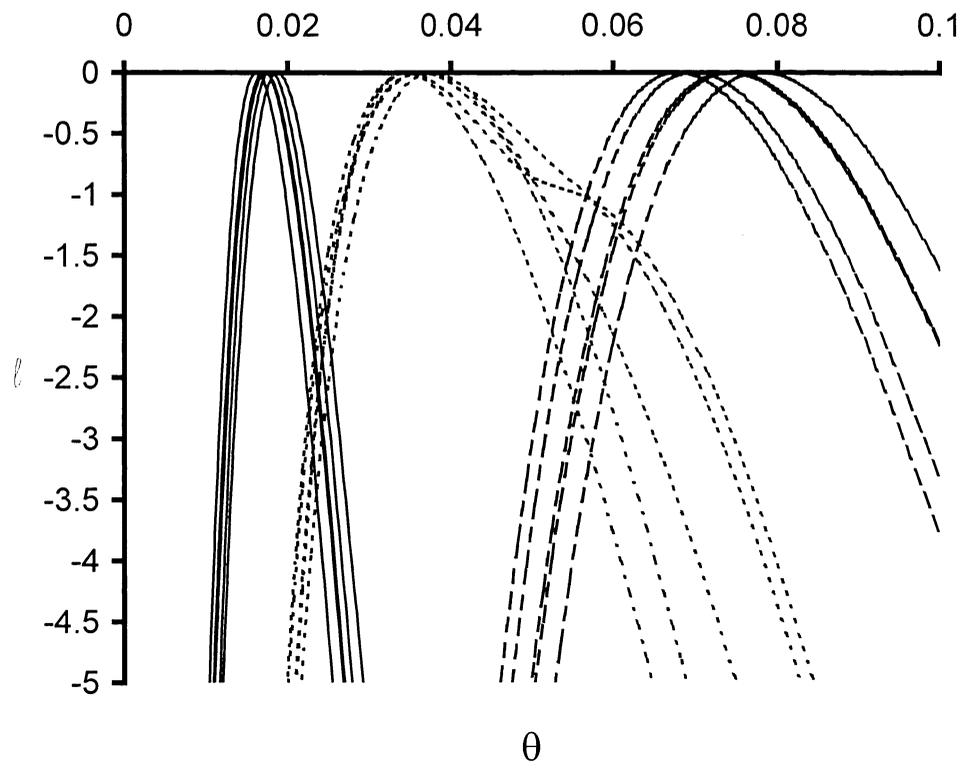


Figure 4

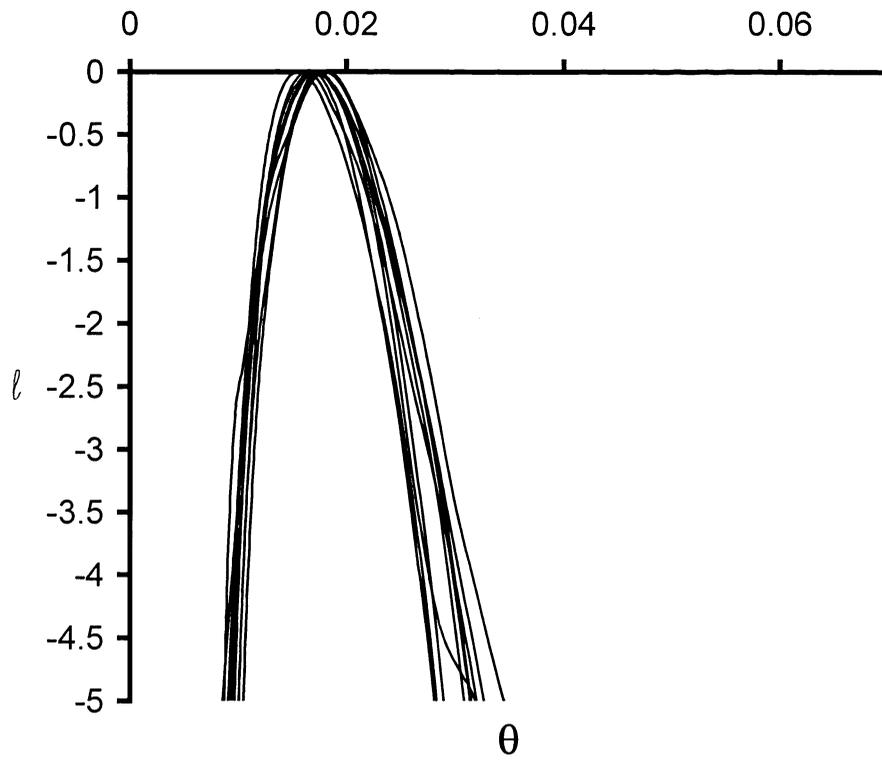


Figure 5

