

MUTATIONS AS MISSING DATA: INFERENCES ON THE AGES AND DISTRIBUTIONS OF NONSYNONYMOUS AND SYNONYMOUS MUTATIONS

M-1555

December 2000

Rasmus Nielsen

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

Abstract:

We present a new Markov Chain Monte Carlo (MCMC) method applicable to DNA sequence data, which treats mutations in the genealogy 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. We demonstrate the utility of the method in three applications. First, we demonstrate how the method can be used to make inferences regarding population genetical parameters such as θ (the effective population size times the mutation rate). Second, we show how the method can be used to estimate the ages of mutations in finite sites models and for making inferences regarding the distribution and ages of nonsynonymous and synonymous mutations. The method is applied to two previously published data sets and we demonstrate that in one of the data sets the age of nonsynonymous mutations, suggesting the presence of slightly deleterious mutations. Third, we demonstrate how the method in general can be used to evaluate the posterior distribution of a function of a mapping of mutations on a gene genealogy. This application is useful for evaluating the uncertainty associated with methods that rely on mapping mutations on a phylogeny or a gene genealogy.

Running head: Mutations as Missing Data

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

Maximum Likelihood, Data Augmentation.

Corresponding author:

Rasmus Nielsen

Department of Biometrics, Cornell University

439 Warren Hall

Ithaca, NY 14853-7801

Phone: 607-255-1643

Fax: 607-255-4698 *USA*

e-mail: rn28@cornell.edu

ABSTRACT

We present a new Markov Chain Monte Carlo (MCMC) method applicable to DNA sequence data, which treats mutations in the genealogy 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. We demonstrate the utility of the method in three applications. First, we demonstrate how the method can be used to make inferences regarding population genetical parameters such as θ (the effective population size times the mutation rate). Second, we show how the method can be used to estimate the ages of mutations in finite sites models and for making inferences regarding the distribution and ages of nonsynonymous and synonymous mutations. The method is applied to two previously published data set and we demonstrate that in one of the data sets the age of nonsynonymous mutations is significantly lower than the age of synonymous mutations, suggesting the presence of slightly deleterious mutations. Third, we demonstrate how the method in general can be used to evaluate the posterior distribution of a function of a mapping of mutations on a gene genealogy. This application is useful for evaluating the uncertainty associated with methods that rely on mapping mutations on a phylogeny or a gene genealogy.

Mapping of character changes on a phylogeny using parsimony or other methods is one of the most important and fundamental tools in evolutionary biology. In molecular evolution, it is common to map the evolution of nucleotides on a phylogeny or an intraspecific gene genealogy. Inferences regarding the evolution of the molecular characters then proceed by treating the estimated mutational events as the true data. The power of this approach is that it converts sequence data to pseudo-data of mutations on a phylogeny/genealogy. It can be used to track the evolution of specific characters, to estimate the ages and distribution of mutations and in general to examine hypotheses regarding molecular evolution. For example, TEMPLETON (1996) maps mutations on a tree to test if internal branches (edges in the tree that are not connected to leaves) have relative less nonsynonymous changes than external branches (edges connected to leaves). BUSH *et al.* (1999) map the evolution of nucleotide characters on a tree in an attempt to determine which branches in the genealogy have had an excess of nonsynonymous mutations and which codons that have had the most nonsynonymous mutations. Mapping of mutations on trees using parsimony has, for example, also been used to investigate hypotheses regarding rate variation among sites and mutational biases (e.g. WAKELEY 1993, 1994).

Although the mapping of mutations on trees has proven an incredibly powerful tool in molecular evolution, the methodology may be criticized on statistical grounds for at least three reasons.

1. These methods do not take into account the uncertainty in the estimation of tree topology.

2. Parsimony mapping assigns the smallest possible number of mutations on the tree, which may lead to serious biases in the parameter estimates in some cases.
3. The uncertainty in the assignment of mutations on the tree is usually ignored.

For example, even under the parsimony criterion, there are multiple ways of mapping character changes on a tree, e.g. the accelerated transformation (ACCTRAN) and the delayed transformation (DELTRAN). Usually one of these algorithms will be used and the other possible mappings under the parsimony criteria will be ignored.

In many cases, these problems can be alleviated by using one of the available maximum likelihood methods. For example, if we are interested in testing the hypothesis that more nonsynonymous mutations relative to synonymous mutations occur in the internal branches of the tree, this can easily be done using maximum likelihood (YANG 1998, YANG and NIELSEN 1998). However, in some cases we are interested in properties of the mutations, such as the ages of mutations, which cannot be directly estimated in finite sites models of DNA sequence evolution using likelihood methods. Likelihood methods also often suffer from problem 1 raised above, because optimization over all parameters including the tree the topology is impractical, although there is some hope that recently developed Bayesian methods (RANNALA and YANG 1996, YANG and RANNALA 1997, LARGET and SIMON 1999, MAU *et al.* 1999, HUELSENBECK *et al.* 2000) will help alleviate this concern.

In this paper we will develop a statistical method for investigating the distribution of mutations on trees. We will be especially interested in estimating ages of mutations in population genetical data.

The general approach we will take to this problem is to treat mutations as missing data. We will devise a Markov Chain Monte Carlo (MCMC) algorithm that effectively integrates over the set of possible mappings of mutations on a gene tree and the set of possible gene trees. In this way, it is possible to make inferences regarding mutations, while taking into account both the uncertainty in the mapping of mutations of the genealogy and the uncertainty associated with the estimation of the tree topology. We will focus on population genetical data and first develop methods for estimating population genetical parameters such as $\theta = 4N_e\mu$, (N_e is the effective population size and μ is the per generation mutation rate). Thereafter, methods for estimating the age of nonsynonymous and synonymous mutations in the genealogy will be presented. Finally, we will discuss how any functional of the mapping of mutation can be estimated with associated measures of uncertainty using this method. Throughout, the new methods are applied to several previously published data sets of DNA sequences.

MODELS

There has recently been a lot of interest in developing statistical methods for estimating population genetical parameters 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 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 using the demarginalization:

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

where Θ is the vector of parameters, X is the observed genetic data, G denotes the gene genealogy, Ω is the set of all possible genealogies 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 rooted, strongly binary tree where each leaf is associated with one of the sampled haplotypes and with lengths associated with the edges. It consists of a topology (of which there are $(n!)/[n2^{n-1}]$) and a vector 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 of 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 under very general conditions under any other exchangeable model, i.e. models that assumes that all individuals in the population have the same constant variance in off-spring numbers. By relaxing this assumption it is also possible to analyze models that include population structure, changes in population size and selection.

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, usually developed for statistical inference in phylogenetics. 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 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 (e.g. FELSENSTEIN 1981). The likelihood for multiple sites is usually calculated by assuming sites are i.i.d., but this assumption can also be relaxed (e.g. YANG 1993, FELSENSTEIN and CHURCHILL 1996). Because the length of the edges in the gene genealogy are scaled by the mutation rate, $\Pr(X | G, \Theta)$ does not depend on θ .

MUTATIONS AS MISSING DATA

In the following we will describe an algorithm for simulating genealogies and mappings of mutations from a distribution proportional to the likelihood function. The method is based on the METROPOLIS-HASTINGS algorithm (METROPOLIS *et al.* 1953,

HASTINGS 1970) and it is similar to the algorithms by BEERLI and FELSENSTEIN (1999), KUHNER *et al.* (1995, 1998), LARGET and SIMON (1999), MAU *et al.* (1998), and HUELSENBECK *et al.* (2000), except that it treats mutations as missing data.

Notice that the likelihood can be written as

$$L(\Theta | X) = \int \int I(X, G, \eta) dP_{G, \Theta}(\eta) dP_{\Theta}(G), \quad (5)$$

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

$$\min(1, r), \quad r = \frac{P_{\Theta}(G') P_{G', \Theta}(\eta') h[(\eta', G') \rightarrow (\eta, G)]}{P_{\Theta}(G) P_{G, \Theta}(\eta) h[(\eta, G) \rightarrow (\eta', G')]} . \quad (6)$$

Assuming that the proposal kernel is constructed such that the chain is ergodic, the chain will have stationary distribution of (η, G) proportional to $\Pr(X | \Theta, \eta, G)$.

Algorithmics: The proposal algorithm presented 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

$\frac{P_{G', \theta}(\eta')h[(\eta', G') \rightarrow (\eta, G)]}{P_{G, \theta}(\eta)h[(\eta, G) \rightarrow (\eta', G)]}$. 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 node is encountered, continue to move the edge along either of the two other edges connecting to the node, 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,

$$\frac{P_{G', \Theta}(\eta')}{h[(\eta, G), (\eta', G')]} = \prod_s z_s f_s \exp(u_s), \quad (7)$$

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

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

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

$$z = \begin{cases} 1 & \text{if } k = 0 \\ q_{b_k E} / q_A & \text{if } k = 1 \\ q_{b_k E} \sum_{j: j \neq E \wedge j \neq b_{k-1}} q_{b_{k-1} j} / q_A^2 & \text{if } k > 1 \end{cases}, \quad (8)$$

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).

Example: To illustrate how the method works, we will in the following give an example of a single update for a data set containing 4 sequences and a single site. The initial state is depicted in FIGURE 2a. We will assume a rate matrix of the form

$$\begin{bmatrix} -1 & 1/3 & 1/3 & 1/3 \\ 1/3 & -1 & 1/3 & 1/3 \\ 1/3 & 1/3 & -1 & 1/3 \\ 1/3 & 1/3 & 1/3 & -1 \end{bmatrix}$$

and $\theta = 1$. We choose a random edge, in this case the striped edge (see FIGURE 2a), and slide it a random distance to a new point in the genealogy. The mutations on the edge are erased and a new set of mutations are simulated. For the site depicted in FIGURE 2, the new edge must start with state **c** and terminate in state **g**. We simulate a random number of mutations (k) from a Poisson with rate $q_{\text{cct}}t = 0.2$, conditional on $k \neq 0$. In this incident $k = 1$, and since $E = \mathbf{g}$, we determine that the mutation is a **c** \rightarrow **g** mutation. The time of the mutation is chosen uniformly in the interval $[0, 0.2]$, and is in this case chosen to be

0.1 (FIGURE 2b). We then have $\frac{P_{\Theta}(G')}{P_{\Theta}(G)} = 9.025$, $\frac{P_{G,\Theta}(\eta)}{h[(\eta', G'), (\eta, G)]} = 0.044$,

and $\frac{P_{G',\theta}(\eta')}{h[(\eta, G), (\eta', G')]} = 0.005$. The acceptance probability for the new update is then

$$\min\{1, 0.937\} = 0.937.$$

Improving mixing: One potential concern is that the chain may not mix fast between different mappings with similar posterior probabilities. Especially, if the number of mutations per site is low, the chain might get trapped in a particular mapping corresponding to a parsimony mapping that does not communicate well with other possible parsimony mappings. To alleviate this problem, an update algorithm is added to the chain, which for all sites in the sequence in which the current state (η) is a parsimony mapping, will choose a new parsimony mapping (η') uniformly among all possible parsimony mappings. The new mapping is then accepted with probability

$$\min(1, r), \quad r = \frac{P_{G',\theta}(\eta')}{P_{G,\theta}(\eta)}. \quad (9).$$

Such updates are added to chain every 50 cycles.

Since the chain is aperiodic, the probability of moving an edge to any other position in the genealogy in any particular cycle of the chain is positive, and the probability of changing the distributions of a mutation on an edge to any other supported distribution of mutations on the edge in a single cycle is positive, the chain is ergodic as desired.

However, this does obviously not guarantee convergence of the ergodic averages in finite time.

ESTIMATION OF POPULATION GENETIC PARAMETERS

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). In the following we will investigate how well the new method can be used to estimate the likelihood function for θ , i.e. we set $\Theta = \{\theta\}$. We calculate the data probability according to the model described in Equation (3) assuming independence among sites and we use Equation (2) to assign probabilities to genealogies. We will assume a uniform prior for θ in the interval $(0, \theta_{max})$, where θ_{max} is chosen to be sufficiently large to provide estimates of the likelihood/posterior distribution for all values of interest. Equation 1 is then evaluated stochastically by simulation of a Markov chain which has stationary distribution of (G, θ) equal to $P(G, \theta | X) \propto \Pr(X | G)P_{\theta}(G)P(\theta)$. We then sample values of θ from this chain. Since $P(\theta | X) \propto \Pr(X | \theta)$ when assuming a uniform prior for θ , the likelihood function for θ can be approximated by the empirical distribution of values of θ sampled from the chain, .i.e.

$$\Pr(\theta \in (a, b) | X) \approx \frac{1}{k} \sum_{i=1}^k I_{\theta_i \in (a, b)} \quad (10)$$

where $I_{\theta_i \in (a, b)}$ is the indicator for the event that the i th sampled value of θ is in the interval (a, b) .

The method used for simulating the Markov chain is as described in the Mutation as Missing Data section, however, the state space of the chain is expanded to also include

θ . Updates to θ are proposed by choosing a new value of θ (θ_1) uniformly in the interval $(\theta_0 - \delta, \theta_0 + \delta)$, where θ_0 is the current value of θ and δ is a value that can be adjusted to provide an appropriate intermediate acceptance rate (e.g. GELMAN *et al.* 1995). If $\theta_1 < 0$ we set $\theta_1 = -\theta_1$, and if $\theta_1 > \theta_{max}$ we set $\theta_1 = 2\theta_{max} - \theta_1$. The acceptance probability for this type of update is then

$$\min(1, r), \quad r = \frac{P_{\theta_1}(G, \eta)}{P_{\theta_0}(G, \eta)}. \quad (11)$$

To check the computer implementation of the new method, 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 if a sufficient number of cycles of the Markov chain were simulated.

Applications: The performance of the new method was evaluated by repeated analysis of a real data set. It consists of a 360 nucleotide long region of 63 mtDNA sequences 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. In the simulations a uniform prior on the interval $(0, 1000)$ was assumed for the parameter κ . The state space of the chain was augmented with this parameter and updates to the κ was performed similarly to the updates for θ . θ_{max} was set to $10\theta_w$, and δ was set to θ_w , where θ_w is the WATTERSON (1975) estimate of θ . The results of the 10 replicates using different initial

values of the parameters are in FIGURE 3a,b. Notice that the likelihood surfaces are roughly similar. We would make essentially the same statistical inferences from each replicate. This suggests that convergence was achieved and demonstrates that the method in fact can be used for population genetical inferences.

ESTIMATION OF THE AGES OF MUTATIONS

There has recently been considerable interest in making inferences about ages of mutations in DNA sequence data. There are several reasons for this. First, information regarding the age of mutations may be used in linkage disequilibrium studies (e.g. SLATKIN and RANNALA 1997, RANNALA and SLATKIN 1998). Second, in many studies the age of a mutation may be of some relevance in itself because it may answer questions regarding the origin and spread of genetic diseases or other heritable traits. Third, information about ages of mutations may be used for testing models of molecular evolution. For example, NIELSEN and WEINREICH (1999) used estimates of the age of nonsynonymous and synonymous mutations to test if nonsynonymous mutations were selectively neutral in mitochondrial DNA.

Several estimators of the age of a mutation have been proposed (e.g. GRIFFITHS and TAVARÉ 1998, 1999, SLATKIN and RANNALA 1997). For example, GRIFFITHS and TAVARÉ (1999) provided an estimator of the age of an allele using the neutral classical neutral coalescence model (KINGMAN 1980a, b). They assumed an infinite sites model and provided a Bayesian estimator of the age of a mutation in a single site. Another example is the likelihood estimator by SLATKIN and RANNALA (1997) applicable to a low frequency, nonrecurring disease mutation.

For a set of aligned DNA sequences, the most efficient estimators of the age of a mutation should employ all of the genealogical information contained in the sample. The estimator of GRIFFITHS and TAVARÉ (1999) achieves this under the infinite sites model. Also, MARKOVTSOVA *et al.* (2000) provided an estimator of the age of mutation in which the polymorphism of interest is assumed to be a unique event polymorphism (UEP), i.e. the polymorphism is caused by a single unique mutation. However, they allow the surrounding sites to follow a finite sites model of DNA sequence evolution.

However, in much of the commonly analyzed data, such as most mammalian mitochondrial data, polymorphisms cannot with certainty be determined to be unique. In much of the available data, especially mitochondrial DNA data, there is a significant probability that more than one mutation is segregating in a particular variable site. The problem of estimating the age of a mutation, therefore, becomes a problem of estimating the age, or average age, of a polymorphism.

The solution to the problem is to integrate over all the possible ways mutations could be distributed on a genealogy to get the average age of a mutation in a site. The posterior expectation of the age can be written as

$$E(A_i | X) = \frac{\int \int E(A_i | G, \eta) I(X, G, \eta) dP_{G, \ominus}(\eta) dP_{\ominus}(G)}{\int \int I(X, G, \eta) dP_{G, \ominus}(\eta) dP_{\ominus}(G)}. \quad (12)$$

Here A_i is the age of a mutation in site i and $E(A_i | G, \eta)$ is the average age of a mutation in site i given a particular gene genealogy and a distribution of mutations on the gene genealogy. $E(A_i | X)$ is the estimator of the age of a mutation and can be considered a

Bayesian estimator, since it is given by the posterior expectation of A_i . The integrals in Equation (3) can easily be solved stochastically by the MCMC method previously discussed because $E(A_i | G, \eta)$ can be observed directly from the genealogy. A simulation consistent estimator of $E(A_i | X)$ is given by

$$A_i = \frac{\sum_{k,j} A_i^{(j,k)}}{\sum_k n(i,k)} \quad (13)$$

where $A_i^{(j,k)}$ is the age of the j th mutation in site i in the k th step of a Markov chain with stationary distribution of G and η given by $P_{\circ}(G, \eta | X)$ and where $n(i,k)$ is the number of mutations in site i in the k th cycle of this chain. A Markov chain with this stationary distribution can be established using the methods described in the section regarding the estimation of population genetic parameters. It is exactly the same Markov chain as in the previous case, except now we are interested in the distribution of genealogies themselves for the purpose of making Bayesian inferences regarding mutations and genealogy.

Applications: The method was applied to two different previously published data sets. The distribution of genealogies and the mutational process was modeled as in the previous section. The first data set is a mtDNA data set (*Cytochrome b*) from *Mesomys hispidus* (spiny tree rat), obtained from GenBank and originally sequenced by DA SILVA and PATTON (1993) and LARA *et al.* (1996). It contains 29 sequences of length 798 bp (Accession Numbers: L23365-81, L23384-86, L23390-91, L23393-98). The second data set is a data set of the *Influenza A hemagglutinin* gene. It contains 28 sequences of length

987 bp. It was previously analyzed in YANG *et al.* (2000) and is a subset of a data set analyzed by FITCH *et al.* (1997).

To ensure convergence 5.500.000 cycles of the Markov chain was simulated for each gene. The first 500.000 cycles were used as burn-in time. Three runs were completed for each data set using different starting trees, all runs giving essentially identical results. Only the results from the first run are shown for each data set.

The ratio of the average age of a nonsynonymous to mutation to the average age of a synonymous mutation was 1.034 for the Influenza data set and 0.682 for the mtDNA data set. To test if the ratios of ages were different than one, a two-sample bootstrap test for difference in the mean age was performed on each data set using 1,000,000 simulations. Only sites in which the posterior probability of at least one mutation was larger than 0.2 were included. The *p*-value for the mtDNA data set was approx. 0.000028 and the *p*-value for the Influenza data set was approx. 0.836. In the *Mesomys Cytochrome b* data set, 25% of sites in which the posterior probability of at least one nonsynonymous mutation is larger than 0.2 have an average age of nonsynonymous mutations less than $0.05N_e$. Only 2.59% of sites in which the posterior probability of at least one synonymous mutation is larger than 0.2, has an average age of synonymous mutations less than $0.05N_e$.

The difference between the two data sets is not surprising. It has previously been suggested that the average age of mutations in nonsynonymous sites is less than the average age of synonymous mutations in a smaller data set of *Cytochrome b* sequences from *Mesomys hispidus* (NIELSEN and WEINREICH 1999). In contrast, BUSH *et al.* (1999) and YANG *et al.* (2000) suggest that positive selection may be occurring in the in

hemagglutinin gene of the Influenza virus. If anything, recurrent positively selected mutations would increase the ratio of the age of nonsynonymous to the age of synonymous mutation (e.g. NIELSEN and WEINREICH 1999). The most obvious explanation for these results is the presence of slightly deleterious mutations in the evolution of *Cytochrome b* in *Mesomys hispidus*. The posterior expectation of the age of a nonsynonymous and a synonymous mutation is plotted along the sequence in FIGURE 4. There appear to be two regions in which the average age of a nonsynonymous mutation is especially low, a region around site 200 and a region around site 550. These areas are located largely outside known reaction sites and outside transmembrane regions of the molecule, although the second region extends into transmembrane domain V. Also, notice that, if anything, there might be negative correlation between the rate of substitution and the age of a mutation in a region. One possible explanation is that slightly deleterious mutations are more frequent in regions of less functional importance.

EVALUATING FUNCTIONS OF THE DISTRIBUTION OF MUTATIONS

In the previous two sections we demonstrated how the method can be used to make inferences regarding population genetical parameters (θ) and to estimate the ages of mutations/polymorphisms. However, the method can be applied quite generally to evaluate the expectation of any function of a mapping of mutations. As previously discussed, mappings of mutations on trees has been a powerful tool in molecular evolution despite some scepticism regarding the robustness of these methods to violations of assumptions regarding tree topology and mapping algorithm. However, by stochastically integrating over the set of possible topologies, edge lengths and mappings, it is possible to

alleviate these valid statistical concerns. The posterior distribution of a discrete function of the genealogy and distribution of mutations $[h(G, \eta)]$ is given by

$$\Pr(h(G, \eta) = y | X) = \int \int \delta_{[h(G, \eta), y]} I(X, G, \eta) dP_{G, \Theta}(\eta) dP_{\Theta}(G), \quad (14)$$

where $\delta_{[h(G, \eta), y]}$ is Kronecker's delta function returning 1 if $h(G, \eta) = y$ and 0 otherwise. It is estimated by

$$\Pr(h(G, \eta) = y | X) \approx \frac{1}{n} \sum_{i=1}^n \delta_{[h(G_i, \eta_i), y]} \quad (15)$$

where G_i and η_i are the i th of n samples of G and η , respectively, from a Markov chain with stationary distribution of G and η equal to the posterior distribution of G and η .

Continuous densities can similarly be evaluated using density estimation methods such as kernel smoothing.

Applications: As an example, we consider tests of the distribution of the number of mutations among branches. For example, TEMPLETON (1996) has suggested that selective neutrality can be tested by mapping mutations on trees and testing if internal branches in the genealogy have significantly more or less nonsynonymous substitutions relative to synonymous mutations than external branches. The test can, as in TEMPLETON (1996) be performed as a test of homogeneity in a 2×2 table with columns containing nonsynonymous and synonymous mutations and rows containing mutations on internal and external branches. We note that the hypotheses of equal rates of nonsynonymous and

synonymous substitutions in internal versus external branches also can be performed directly as a likelihood ratio test using codon based models, although in such a framework it may still be difficult to avoid making simplistic assumptions regarding the tree topology. However, the present framework provides a methods for examining how robust the tests based on mapping of mutations on genealogies are to assumptions regarding tree topology and mapping algorithm. In this case the function of the gene genealogy and mutational mapping we are interested in is the posterior distribution of the test statistic or the posterior distribution of the estimated p -value in the test of homogeneity. We estimated the distribution of p -values for the three previously discussed data sets using the same runs of the Markov chain as in the section regarding the age of mutations.

The results are shown in FIGURE 5. As expected from the previous analysis, the p -values for the *Influenza* data set tend to be rather large whereas the p -values from the *Mesomys* data set are smaller with a majority of p -values less than 0.001. However, also notice that there is quite a spread in the p -values for the *Mesomys* data set. This illustrates that the p -values obtained from a single mapping of mutations is highly dependent on the particular mapping of mutations and emphasizes that results based on mappings on mutation that only considers one of the possible mappings should be interpreted with caution.

DISCUSSION

The new method shows promise as an estimator of population genetical parameters, for Bayesian estimation of properties of the genealogy and the distribution of mutations and in general as a method for evaluating any function of the distribution of

mutations on a gene genealogy. There are already several methods available for estimating population genetical parameters using MCMC and related simulation methods (e.g. KUHNER *et al.* 1995, 1998, BEAUMONT, 1999, BEERLI and FELSENSTEIN 1999, GRIFFITHS and TAVARÉ 1994a,b, WILSON and BALDING 1998, NIELSEN 2000). No effort was done to compare the efficiency of this new method with previous methods, but representing the mutations as missing data may in many cases not lead to more efficient methods because of the decrease in mixing caused by the increase in the size of state space. Some exceptions may include models in which inferences are done separately on nonsynonymous and synonymous mutations. In likelihood methods that do not involve representing the mutations directly on a tree, the treatment of nonsynonymous and synonymous mutations dictates that the state space of the Markov chain is given by the 61 possible codons in the universal genetic code (e.g. GOLDMAN and YANG 1994, MUSE and GAUT 1994). Calculation of the data likelihood directly under these models is extremely computationally intensive and it is therefore likely that the new MCMC method for these models will be faster than methods that do not treat mutations as missing data.

Representing the mutations on the tree as missing data is an attractive alternative because it avoids the computational problem of calculating the data likelihood directly by summing transition probabilities over the unobserved states at the nodes of the genealogy.

However, the most important application of the method appears to be for Bayesian inferences regarding the distribution of mutations on the gene genealogy. Here, we especially focused on inferences regarding the age of nonsynonymous and synonymous mutations. Such inferences are of interest because the ages of mutations are intimately related to models of selection at the DNA level.

One potential problem of concern is the assumption of a neutral coalescent prior. This distribution of genealogies may obviously not be correct under selection and is also likely to be violated under more realistic demographic models. However, if nonsynonymous and synonymous mutations are distributed identically on the tree, the effect of assuming a wrong prior is same for the two types of mutations. Therefore, test of equal average ages of nonsynonymous and synonymous mutations, as implemented here, will not give an excess of falsely significant results beyond the chosen significance level, even if a wrong prior has been assumed for the distribution of genealogies. Also, for large samples the distribution of genealogies will be determined primarily by the data and the prior distribution will have a diminishing effect on any inferences made.

We note here that there are many other potential applications of the method, for example in the analysis of the degree of correlation in the mutation process and in the analysis of the substitution process in interspecific data, i.e. data from multiple different species. In this paper the focus was on population genetical models, population genetical data and population genetical problems. However, analyzing interspecific data would only require a change in the prior distribution of genealogies. For example, RANNALA and YANG (1996) and YANG and RANNALA (1997) assumed a prior derived from birth-death processes and LARGET and SIMON (1999) and MAU *et al.* (1999) assumed a uninformative (uniform) prior. The question arises if treating mutations as missing data in MCMC methods is sufficiently efficient for interspecific data. Part of the problem is that mixing of the chain may decrease as the degree of divergence increases. A few runs were performed on a heavily saturated simulated data set. In the particular case examined, the proportion of updates that lead to changes in the topology was very low, leading to concerns that the

method may be more easily applied to intraspecific population data than to data from highly diverged species. However, it may in the future be possible to devise more efficient update algorithms that will also allow the efficient analysis of interspecific data in a Bayesian framework using mutations as missing data.

A final limitation of the method is the assumption of a single shared genealogy for all sites. This assumption implies that there can be no recombination, and the method is therefore only applicable to mtDNA data, Y-chromosome data and data from viral strains with little or no recombination when considering population genetical problems. In theory, however, the method could also be implemented for models that include recombination as in NIELSEN (2000).

LITERATURE CITED

- BEAUMONT, M. A., 1999 Detecting Population Expansion and Decline Using Microsatellites. *Genetics* **153**: 2013-2029.
- BEERLI, P. and J. FELSENSTEIN, 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>.
- BUSH R. M., FITCH WM, BENDER C. A. and N. J. COX., 1999 Positive selection on the H3 hemagglutinin gene of human influenza virus A. *Mol Biol Evol* **16**: 1457-1465.
- DA SILVA, M. N. F. and J. L. PATTON, 1993 Amazonian phylogeography: mtDNA sequence variation in arboreal echimyid rodents (Caviomorpha) *Mol. Phyl. Evol.* **2**: 243-255.
- FELSENSTEIN, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**: 368-76.
- FELSENSTEIN, J., 1992 Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *Genet. Res.* **60**: 209-220.
- FELSENSTEIN, J. and G. A. CHURCHILL, 1996 A hidden Markov model approach to variation among sites in rate of evolution. *Mol. Biol. and Evol.* **93**: 93-104.
- FISHER, R. A., 1930 *The genetical theory of natural selection*, 1st Ed. Clarendon press, Oxford.

- FITCH, W. M., R. BUSH, M., C. A. BENDER, and N. J. COX, 1997 Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Nat. Acad. Sci.* **94**: 7712-7718.
- GELMAN, A., G. O. ROBERTS, and W. R. GILKS, 1995 Efficient Metropolis jumping rules, in *Bayesian Statistics 5*, edited by BERNADO, J. M., BERGER, J. O., DAVID, A. P. and SMITH, A. F. M. Oxford University Press, Oxford, UK.
- GOLDMAN, N., AND Z. YANG., 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**: 725-736.
- GRIFFITHS, R. C. and S. TAVARÉ, 1994a Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46**: 131-159.
- GRIFFITHS, R. C. and S. TAVARÉ, 1994b, Ancestral inference in population genetics. *Stat. Sci.* **9**: 307-319.
- GRIFFITHS R. C. and S. TAVARÉ, 1998 The age of a mutation in a general coalescent tree. *Stoch. Models* **14**: 273-295.
- GRIFFITHS R. C. and S. TAVARÉ, 1999 The ages of mutations in gene trees. *Ann. Appl. Prob.* **9**: 567-590.
- HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**: 97-109.
- HUELSENBECK, J. P., B. RANNALA, and B. LARGET, 2000 A Bayesian framework for the analysis of cospeciation. *Evolution* **54**: 353-364.
- KINGMAN, J. F. C., 1982a The coalescent. *Stochast. Proc. Appl.* **13**: 235-248.
- KINGMAN, J.F.C., 1982b On the genealogy of large populations. *J. Appl. Prob.* **19A**: 27-43.

- KISHINO, H. and M. HASEGAWA, 1989 Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of Hominoidea. *J. Mol. Evol.* **31**: 151-160.
- KUHNER, M. K., J. YAMATO, and J. FELSENSTEIN, 1995 Estimating effective population size and mutation rate from sequence data using METROPOLIS-HASTINGS sampling. *Genetics* **140**: 1421-1430.
- KUHNER, M. H., J. YAMATO, and J. FELSENSTEIN, 1998 Maximum likelihood estimation of population growth rates based on the coalescent. *Genetics* **149**: 429-434.
- LARA, M. C., J. L. PATTON, and M. N. F. DA SILVA, 1996 The simultaneous diversification of South American echimyid rodents (*Hystricognathi*) based on complete *cytochrome b* sequences. *Mol. Phyl. Evol.* **5**: 403-413.
- LARGET, B. and D. SIMON, 1999 Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* **16**: 750-759.
- MARKOVITSOVA, L, MARJORAM, P, and S. TAVARÉ, 2000 The age of a unique event polymorphism. *Genetics* **156**: 401-409.
- MAU, B., M.A. NEWTON, and B. LARGET, 1999 Bayesian phylogenetic inference via Markov Chain Monte Carlo methods. *Biometrics* **55**:1-12.
- METROPOLIS, N., A. W. ROSENBLUTH., M. N. ROSENBLUTH, A. H. TELLER, and E. TELLER, 1953 Equations of state calculations by fast computing machines. *J. Chem. Phys.* **21**: 1087-1091.
- MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to chloroplast genome. *Mol. Biol. Evol.* **11**: 715-724.

- 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). *Genetics* **154**: 931-942.
- NIELSEN R. and D. M. WEINREICH, 1999 The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics* **153**: 497-506.
- RANNALA, B. and M. SLATKIN, 1998 Likelihood analysis of disequilibrium mapping, and related problems. *Am. J. Hum. Genet.* **62**: 459-473.
- RANNALA, B., and Z. YANG, 1996 Probability distribution of molecular evolutionary trees: a new method of phylogenetic inference. *J. Mol. Evol.* **43**: 304-311
- SLATKIN M and B. RANNALA, 1997 Estimating the age of alleles by use of intraallelic variability. *Am. J. Hum. Genet.* **60**: 447-458.
- TEMPLETON, A. R., 1996 Contingency tests of neutrality using intra/interspecific gene trees: The rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. *Genetics* **144**: 1263-1270.
- THOMPSON, E. A., 1975 *Evolutionary trees*. Cambridge University Press, Cambridge, UK.
- WAKELEY J., 1993 Substitution rate variation among sites in hypervariable region-1 of human mitochondrial DNA. *J. Mol. Evol.* **37**: 613-623.
- WAKELEY J., 1994 Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* **11**: 436-442.

- WARD, R. H., B. L. FRAZIER, K. DEW-JAGER, and S. PÄÄBO, 1991 Extensive mitochondrial diversity within a single Amerindian tribe. *Proc. Natl. Acad. Sci. USA* **88**: 8720-8724
- WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. *Theor. Pop. Biol.* **7**: 256-276.
- WILSON, I. J. and D. J. BALDING, 1998 Genealogical inference from microsatellite data. *Genetics* **150**: 499-510.
- YANG, Z., 1993 Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**: 1396-1401.
- YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* **15**: 568-573.
- YANG, Z. and R. NIELSEN, 1998 Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* **46**: 409-418.
- YANG, Z., R. NIELSEN, N. GOLDMAN, and A.-M. K. PEDERSEN, 2000 Codon-Substitution Models for Variable Selection Pressure at Amino Acid Sites. *Genetics* **155**: 431-449.
- YANG, Z., and B. RANNALA., 1997 Bayesian phylogenetic inference using DNA sequences: a Markov chain Monte Carlo method. *Mol. Biol. Evol.* **14**: 717-724.

FIGURE LEGENDS

FIGURE 1.

Title: An example of a gene genealogy for five gene copies.

FIGURE 2.

Title: Example of an update of the genealogy.

Legend: The genealogy with mutations for a single site before and after the update discussed in the text. The times of coalescence and mutation events are measured backwards and are scaled by the mutation rate, using the convention that the sequences are sampled at time 0. The arrow in (a) indicates the move of an edge proposed in the update leading to the genealogy depicted in (b).

FIGURE 3.

Title: Likelihood function for θ .

Legend: The likelihood function for θ for the mtDNA data set by WARD *et al* (1991). Each likelihood surface was estimated independently with varying initial conditions using 1,000,000 cycles of the Markov chain.

FIGURE 4.

Title: Ages of mutations.

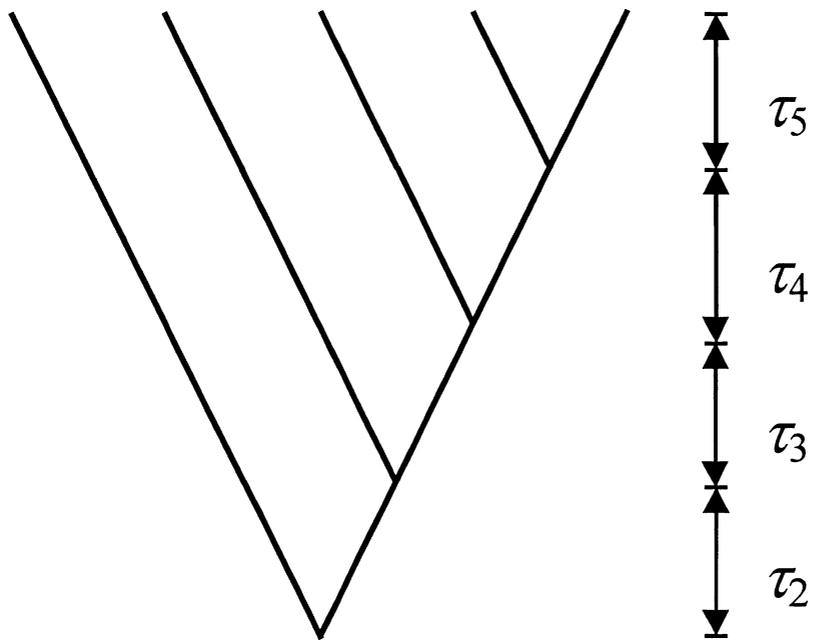
Legend: The distribution of the posterior age of a mutation and the variability (number of nonsynonymous mutations per site) in the *Cytochrome b* data set by DA SILVA and PATTON (1993) and LARA *et al.* (1996), estimated using 5,000,000 cycles of the Markov chain.

FIGURE 5.

Title: The posterior distribution of p -values.

Legend: The posterior distribution of the p -value of the test of homogeneity of the hypothesis that the ratio of the number of synonymous mutations on internal to external branches equals the ratio of the number of nonsynonymous mutations on internal to external branches.

FIGURE
1



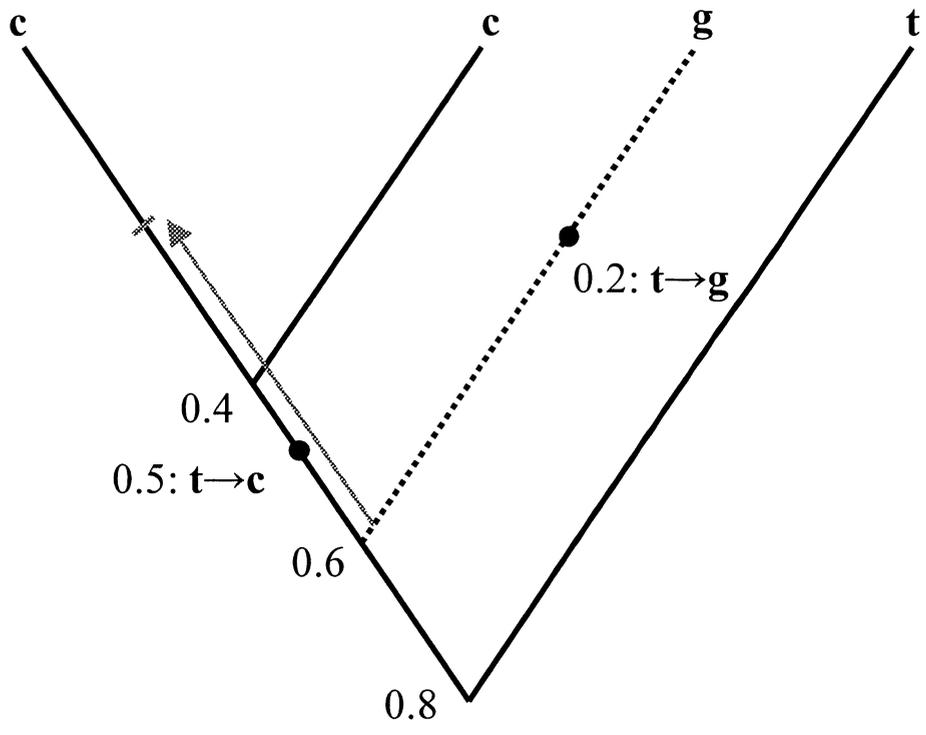


FIGURE 2a

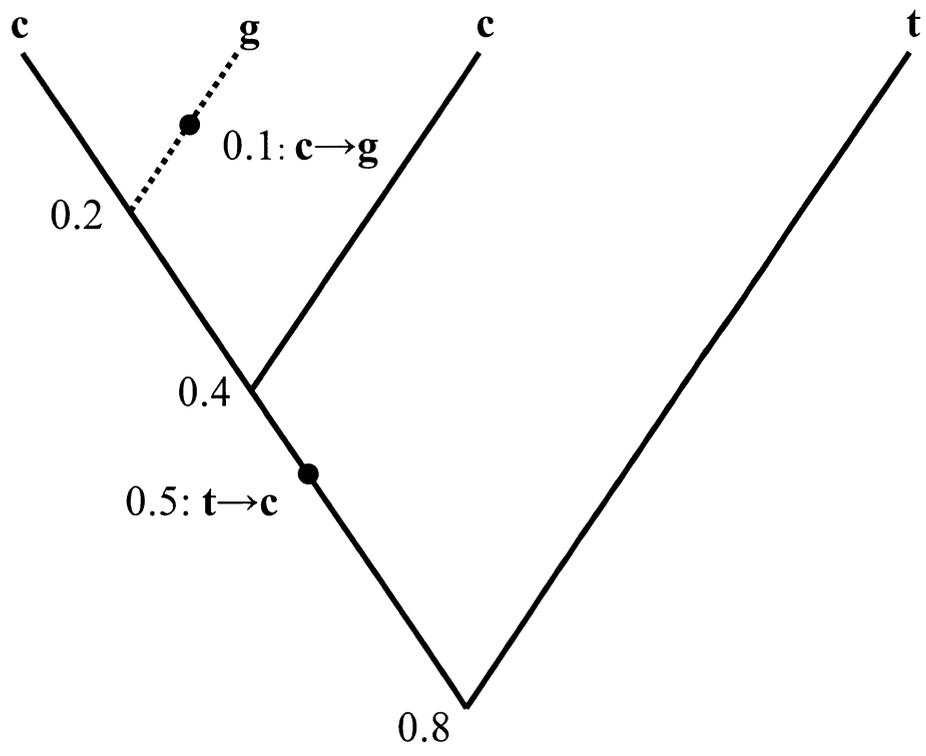


FIGURE 2b

FIGURE 3

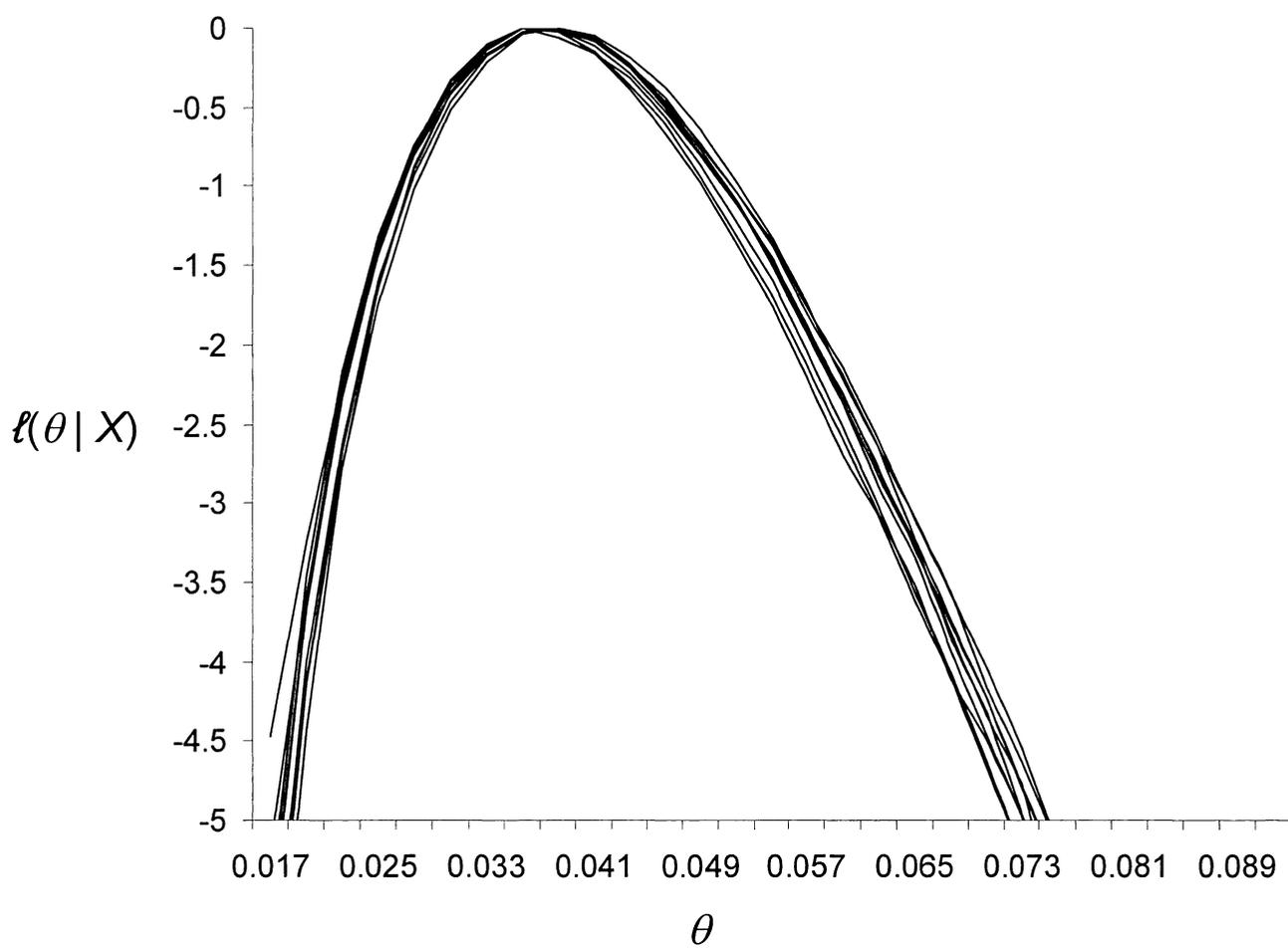


FIGURE 4

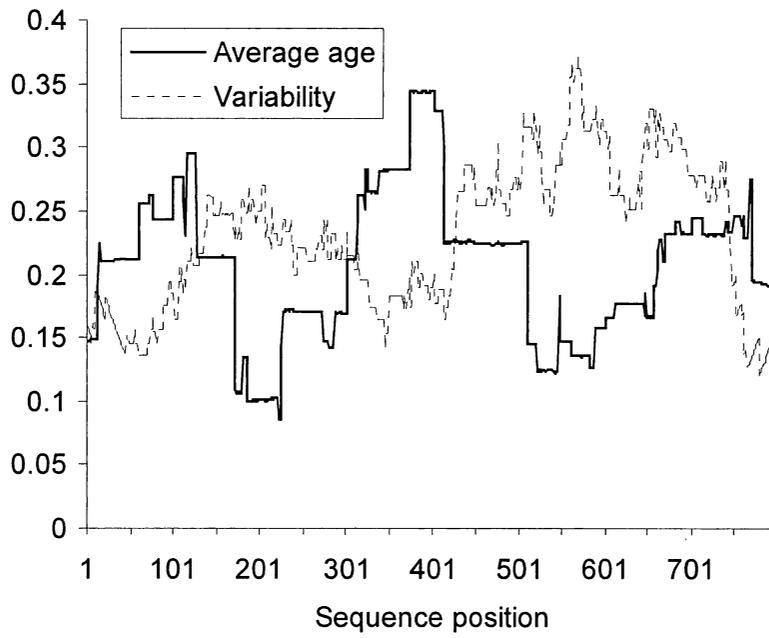


FIGURE 5

