A Likelihood Approach to Populations Samples of Microsatellite Alleles

Rasmus Nielsen

Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Manuscript received October 30, 1996
Accepted for publication February 21, 1997

ABSTRACT

This paper presents a likelihood approach to population samples of microsatellite alleles. A Markov chain recursion method previously published by GRIFFITHS and TAVARÉ is applied to estimate the likelihood function under different models of microsatellite evolution. The method presented can be applied to estimate a fundamental population genetics parameter θ as well as parameters of the mutational model. The new likelihood estimator provides a better estimator of θ in terms of the mean square error than previous approaches. Furthermore, it is demonstrated how the method may easily be applied to test models of microsatellite evolution. In particular it is shown how to compare a one-step model of microsatellite evolution to a multi-step model by a likelihood ratio test.

MICROSATELLITE loci have received a lot of attention in recent years because of their use as polymorphic DNA markers. The use of microsatellites extend from linkage mapping (TODD *et al.* 1991) to studies concerning population structure and demography (DEKA *et al.* 1991), and reconstruction of evolutionary trees (BOWCOCK *et al.* 1994). The theoretical tools for analysis of microsatellites are based on the work of OHTA and KIMURA (1973), WEHRHAHN (1975), and MORAN (1975) concerning electrophoretic alleles. DEKA *et al.* (1991) suggested that the evolution of microsatellites could be described by the same mathematical tools as applied in the description of electrophoretic alleles since the mutational process of both approximately conforms to a one-step mutation model, *i.e.*, a model including only mutations to neighboring states. For the purpose of modeling a change in the electrophoretic mobility in a model of electrophoretic alleles corresponds to a change in allele size in a model of microsatellite evolution. Assuming this mutational model and a Wright-Fisher model of evolution it is possible to obtain measures of the expected heterozygosity (OHTA and KIMURA 1973), the variance in allele size (MORAN 1975) and the expected distance between alleles within and between populations (GOLDSTEIN *et al.* 1995; SLATKIN 1995). These measures have subsequently been applied to obtain estimators of the divergence between populations (GOLDSTEIN *et al.* (1995), population subdivision (SLATKIN 1995) and $\theta = 4N_e\mu$ (four times the effective population size times the mutation rate) (VALDES *et al.* 1993). The estimators are typically unbiased moment estimators but are not sufficient statistics for the relevant parameters. For example, since the variance in allele size is equal to $\theta/2$ (WEHRHAHN 1975), we estimate θ by two times the estimate of the variance of allele size. However, the variance in allele size does not contain all of the information in the sample regarding θ.

Another theoretical problem has been how to test the one-step mutation model and compare it with other models using data from population samples. DI RIENZO *et al.* (1994) performed extensive simulations to compare the observed heterozygosity in samples from the Sardinian human population with the heterozygosity predicted under the one-step mutation model given the observed variance in allele size. On the basis of these simulations they were able to reject the one-step mutation model. Simulations have also been applied by DEKA *et al.* (1991) and SHRIVER *et al.* (1993) to compare a one-step mutation model with other models. However, there is no rigorous statistical framework in which different models of microsatellite evolution can be directly compared.

The aim of this article is to provide a statistical framework for the analysis of microsatellite loci. It will be shown how to cast the problem in a likelihood framework and how to evaluate the likelihood function based on the pioneering work of GRIFFITHS and TAVARÉ (1994a). Some properties of the maximum likelihood estimator of θ will be evaluated, and it will be demonstrated how to test hypotheses regarding microsatellite evolution in a likelihood framework. Specifically it will be demonstrated how to compare a one-step mutation model with a multistep mutation model.

METHODS AND RESULTS

The aim of this section is to develop a method for estimating θ by maximum likelihood for a population sample of microsatellite alleles. To do this we assume a neutral coalescent model with Wright-Fisher sampling between generations (see for example TAVARÉ 1984). Let $\boldsymbol{\nu} = (n_1, n_2, n_3, \ldots, n_l)$ be a vector containing the observed data from a population

Author e-mail: rasmus@mws4.biol.berkeley.edu

sample with a difference of $l - 1$ repeats between the smallest and the largest allele, and n_1 copies of the smallest allele, n_2 copies of the second smallest allele and so forth. This coding of the data is convenient because only the relative and not the absolute allele size is relevant in the context of most common models of microsatellite evolution.

The sampling probability provides the likelihood function of θ, *i.e.*, $L(\theta) = P(\boldsymbol{\nu}|\theta)$. To evaluate $L(\theta)$, a model of the mutation process must be adopted. For example, the common one-step mutation model is given by a symmetric random walk on the integers, *i.e.*, a mutation decreases the number of repeats by one with probability $^1/_2$ and increases the repeat number by one with probability $^1/_2$. Applying this model the likelihood function can be calculated directly for small samples.

OHTA and KIMURA (1973) and WEHRHAHN (1975) obtained an expression for the probability that a randomly drawn gene is m repeats larger than another randomly drawn gene. I will rederive this result using coalescence theory and, as an example, show how it provides the likelihood function for a sample of size two ($n = 2$). The coalescence time between the two genes is exponentially distributed with mean 1 and the conditional number of mutations on each lineage is Poisson distributed with mean $\theta t/2$. The joint distribution of the number of mutations increasing repeat size (M_1) and the number of mutations decreasing repeat size (M_2) along both branches in the underlying coalescent tree for two genes is found by integrating over the coalescence time (t) in the underlying genealogy.

$$P(M_1 = m_1, M_2 = m_2) = \int_0^\infty e^{-t} e^{-\theta t/2} \frac{(\theta t/2)^{m_1}}{m_1!} e^{-\theta t/2} \frac{(\theta t/2)^{m_2}}{m_2!}\, dt = \binom{m_1 + m_2}{m_1} \frac{(\theta/2)^{m_1+m_2}}{(1+\theta)^{m_1+m_2+1}}. \quad (1)$$

Therefore, the probability that a randomly drawn gene is m repeats larger than another randomly drawn gene is

$$P_2(m) = \sum_{i=m}^{\infty} \binom{2i - m}{i - m} \frac{(\theta/2)^{2i-m}}{(1+\theta)^{2i-m+1}},$$

which is a series of a type often occurring in connection with random walks (see, for example, DWASS 1967). It equals

$$\frac{[(1 + \theta - \sqrt{1 + 2\theta})/\theta]^m}{\sqrt{1 + 2\theta}}. \quad (2)$$

So for $n = 2$, the likelihood function $[L(\theta)_2]$ is given by

$$L(\theta)_2 = 2P_2(m) \quad \text{if } m > 0 \quad \text{and} \quad L(\theta)_2 = 1/\sqrt{1 + 2\theta} \quad \text{if } m = 0. \quad (3)$$

Similar expressions can be obtained for $n > 2$ by integrating over all coalescence times and summing over all topologies of the gene genealogy. However, the resulting expressions become very difficult to evaluate even for moderate sample sizes despite the simplicity of the mutational process.

One simplification that can be made is to limit the values $\boldsymbol{\nu}$ can take. We may assume that the mutational process follows a symmetric random walk with k states and with reflecting barriers. Constraints on the number of repeats have previously been assumed by GOLDSTEIN *et al.* (1995). In effect such a model becomes a k-allele model. The advantage of this model is that the effect of introducing the barriers can be reduced arbitrarily by letting k become large while it allows the model to be treated in the theoretical framework of k-allele models. However, it is still not possible to obtain an analytical result for the likelihood function for the k-allele model but $q(\boldsymbol{\nu})$ can be estimated by a Markov chain recursion approach described by GRIFFITHS and TAVARÉ (1994a). The GRIFFITHS and TAVARÉ method is discussed in APPENDIX A in the context of models of microsatellite evolution. In short, the method evaluates the likelihood function by representing $q(\boldsymbol{\nu})$ as the expected value of a functional of a Markov process run backward in time on coalescent trees. $q(\boldsymbol{\nu})$ is then evaluated as an average in repeated simulations. The entire likelihood function can be evaluated for many values of θ in one series of simulations by running the simulations using a reasonable value of θ, θ_0, while evaluating the likelihood function for many values of θ.

Applying this method, good estimates of $q(\boldsymbol{\nu})$ can be obtained for realistic sample sizes in minutes. Computing time is greatly reduced for this model in comparison with other k-allele models because the potential number of transitions in the Markov chain is considerably reduced under the one-step mutation model.

TABLE 1

Mean squared error

	MSE variance-based estimator	MSE maximum likelihood estimator (θ)
$n = 10$	2.37	2.39
$n = 50$	2.00	1.13
$n = 100$	1.15	1.23

Evaluation of the estimator: The performance of the method for estimating θ when twice the variance in allele size is used for θ_0 is evaluated in Table 1. For each value of n, 1000 data sets were generated by traditional coalescent simulations (HUDSON 1990) and for each data set 100,000 runs through the Markov chain were performed. The true value of θ was set to 1.0. A model with 100 alleles and reflection boundaries was assumed. The boundaries were very rarely hit and increasing the number of alleles to 200 in some sample cases had no detectable effect. The results of the simulations are evaluated in terms of the estimated mean squared error (MSE) and the performance of the method is compared to the MSE of the variance-based estimator.

It would be expected that the maximum likelihood estimator performs best for large samples. This is true for $n = 50$. However, in the case of $n = 100$, the variance-based estimator performs slightly better. In fact, the MSE is higher for $n = 100$ than for $n = 50$ for $\hat{\theta}$. The reason for this high MSE is the increase in variance introduced by the limited number of runs (100,000). Obviously, $\hat{\theta}$ is a superior estimator only if a sufficient number of runs have been performed. More runs through the Markov chain were not performed because of constraints on the computational resources. In estimation on real data the standard deviation in the estimate may be used to assess if the number of runs performed is sufficient (GRIFFITHS and TAVARÉ 1994a).

Hypothesis testing: One of the great advantages of casting the analysis of microsatellite data in a likelihood framework is that it becomes easy to test specified hypotheses regarding microsatellite evolution. For example, one can easily test if the mutation rate differs between loci. Since the likelihood values obtained for each locus are multiplicative (assuming free recombination), the likelihood function for the two loci (with parameters θ_1 and θ_2, respectively) is simply described by

$$L(\theta_1, \theta_2) = L_1(\theta_1) L_2(\theta_2),$$

under the assumption that the rates differ in the two loci and

$$L(\theta) = L_1(\theta) L_2(\theta),$$

if the rates are assumed to be the same. In other words, the likelihood function and the resulting maximum likelihood estimate are obtained simply by multiplying the likelihood values for the two loci. The hypothesis $\theta_1 = \theta_2 = \theta$ may therefore be tested by a likelihood ratio test with test statistic

$$-2 \log \frac{L_1(\hat{\theta}) L_2(\hat{\theta})}{L_1(\hat{\theta}_1, \hat{\theta}_2)}, \quad (5)$$

which under the null hypothesis is χ^2_1 distributed for large samples. This approximation assumes that the estimator is consistent. Consistency implies that the estimate converges to the true value of the parameter as the sample size goes to infinity. However, for k-allele models, the maximum likelihood estimator may not be consistent. Only a finite number of alleles exist and, therefore, there is a natural limit to the amount of information it is possible to obtain about the mutation parameter. However, in infinite allele models, mutations to new alleles at the tip of the genealogy will introduce more information about the mutation parameter as the sample size increases. Estimators based on infinite allele models will therefore be consistent. In our case a model with infinitely many alleles (the stepwise mutation model) is approximated by a k-allele model. However, the effect of using the k-allele approximation is minimal since the boundaries are very rarely hit. Therefore, it is safe to conclude that the likelihood estimator, for all practical purposes, has the same properties as an estimator based directly on the stepwise mutation model.

Above, the likelihood method was applied to estimate θ. However, the same method may be applied to estimate other parameters of the mutation model. By modifying (4) to include other mutational models, such models can be directly compared by evaluation of the likelihood function.

An important question that has been raised in numerous studies is the fit of the one-step model. An obvious solution to this problem is to compare the one-step model to a multistep model by a likelihood ratio test. To gain as much power as possible in such a test, a multistep model with only one parameter more than the one-step model should be applied. For example, multistep mutations may occur with probability q and one step mutations with probability $1 - q$. If the maximum length of multistep mutations is set to 10 and all multistep mutations are assumed to be equally likely, then the mutation matrix **P** has the following entries:

$$p_{ij} = \begin{cases} q/18 & \text{if } i = j \pm 2, 3, \ldots, 10 \\ (1 - q)/2 & \text{if } i = j \pm 1 \\ 0 & \text{else,} \end{cases}$$

when i and j are not close to the boundaries. This model can be regarded as a generic model of multi-step mutations used for hypothesis testing.

Applying this mutation model, one may apply the likelihood ratio

$$-2 \log \frac{L(\hat{\theta}, q = 0)}{L(\hat{\theta}, \hat{q})}, \quad (6)$$

which for large samples is χ^2_1 distributed under the null hypothesis that $q = 0$.

Applications: The method for testing the one-step mutation model discussed above was applied on a data set published in CRAIGHEAD *et al.* (1995) of microsatellites in the arctic grizzly bear. Data from four loci each containing data from 152 individuals (304 gene copies) were applied. The results of the analysis is reported in Table 2. For each locus, the estimates are based on 3,000,000 runs. Combining the likelihood functions for the four loci under the one-step model provides an estimate of $\hat{\theta} = 4.5$ and $l(\hat{\theta}) = -93.7$. The likelihood ratio

$$-2 \log \frac{L_1(\hat{\theta}) L_2(\hat{\theta}) L_3(\hat{\theta}) L_4(\hat{\theta})}{L_1(\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3, \hat{\theta}_4)} = 2.9$$

provides a test statistic for the hypothesis that the rate is the same in all four loci. Under the null hypothesis, this value should be approximately χ^2_3 distributed and we conclude that to the degree a one-step model is an appropriate description of the evolution in these loci, we cannot reject the hypothesis of constant rate. The assumption of a one-step model is the next hypothesis that will be tested.

TABLE 2

Likelihood values

	One-step model		Multistep model		
	$\hat{\theta}$	$l(\hat{\theta})$	$\hat{\theta}$	$\hat{q}$	$l(\hat{\theta}, \hat{q})$
G10C	3.9	−27.8	3.9	0.0	−27.8
G10L	4.6	−32.5	2.3	0.10	−28.3
G10P	7.6	−40.7	4.3	0.08	−37.5
G10M	4.4	−34.8	3.0	0.12	−33.6

For the for loci $-2 \log [L(\hat{\theta}, q = 0)] / [L(\hat{\theta}, \hat{q})]$ is 0, 8.4, 6.4 and 2.4, respectively. The one-step model can therefore be rejected at the 5% level in favor of a multistep model in two out of four cases. This suggest that a simple one-step mutation model is not sufficient to describe the data in these particular loci. However, it does not necessarily imply the acceptance of the multistep mutation model. Other factors may contribute to the high likelihood ratio values such as fluctuating population size, selection and population subdivision. The test shares this feature with many other methods commonly applied in population genetic analysis. However, the present samples are obtained from what apparently is one interbreeding population. Furthermore, there appear to be no evidence for a recent population expansion in the grizzly bear: in fact, the grizzly bear is more likely to have experienced a declining population size (ALBERT *et al.* 1987). The analysis therefore suggests that multistep mutations may be of importance in the evolution of the analyzed microsatellite loci although other explanations for the rejection of the one-step mutation model may be difficult to rule out. The rejection of a one-step mutation model is also in accordance with the results of simulation studies on human loci (DI RIENZO *et al.* 1994).

DISCUSSION

The analysis of microsatellites in a likelihood framework has many advantages. θ and parameters of the mutational model may easily be estimated and the estimation of these parameters leads directly to a framework for testing hypotheses regarding microsatellite evolution. In this paper it has been shown how several hypotheses that previously were difficult or impossible

to test can now easily be tested in a likelihood framework. In particular, the presented method was used to compare models of the mutational process. This type of analysis can easily be extended to include effects of the boundaries, more categories of mutations, and biased mutation rates toward smaller or larger allele sizes. However, the method may also be generalized to include different demographic processes and multiple loci. The most obvious applications include estimation of rates of recombination between several loci, tests of changing population size and analyses of population subdivision. The case of varying population size has already been implemented for the general k-allele model by GRIFFITHS and TAVARÉ (1994c). This new likelihood approach provides a general framework for analysis of microsatellite population samples that should be applicable in most studies involving population samples of microsatellites. A computer program performing the estimation procedures mentioned in the article is available from the author at http//:mw511.biol.berkeley.edu/software.html.

I thank M. SLATKIN and W. J. EWENS for comments and discussion. I thank S. TAVARÉ for discussion and many helpful comments on the manuscript and for pointing out the problem of consistency in k-allele models. This work was supported in part by National Institutes of Health grant GM-40282 to M. SLATKIN and by a personal grant to R.N. from the Danish Research Council.

LITERATURE CITED

ALBERT, H. L., J. CANFIELD, R. D. MACE, K. A. PATNODE, M. N. LEFRANC *et al.*, 1987 *Grizzly Bear Compendium. Sponsored by the interagency grizzly bear committee.* National Wildlife Federation.

BOWCOCK, A. M., A. RUIZ-LINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 455–457.

CRAIGHEAD, L., D. H. V. PAETKAU, H. V. REYNOLDS, E. R. VYSE and C. STROBECK, 1995 Microsatellite analysis of paternity and reproduction in arctic grizzly bears. J. Hered. **86:** 255–261.

DWASS, M., 1967 Simple random walk and rank order statistics. Ann. Math. Stat. **38:** 1042–1053.

DEKA, R., R. CHAKRABORTY and R. E. FERRELL, 1991 A population genetic study of six VNTR loci in three ethnically defined populations. Genomics **11:** 83–92.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

GOLDSTEIN, D. B., A. R. LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995 An evaluation of genetic distance for use with microsatellite loci. Genetics **139:** 463–471.

GRIFFITHS, R. C., and S. TAVARÉ, 1994a Simulating probability distributions. Theor. Pop. Biol. **46:** 131–159.

GRIFFITHS, R. C., and S. TAVARÉ, 1994b Ancestral inference in population genetics. Stat. Sci. **3:** 307–319.

GRIFFITHS, R. C., and S. TAVARÉ, 1994c Sampling theory for neutral alleles in a varying environment. Phil. Trans. R. Soc. **310:** 403–410.

HUDSON, R., 1990 Gene genealogies and the coalescent process. Oxford Surveys Evol. Biol. **7:** 1–44.

HUDSON, R., and N. L. KAPLAN, 1986 On the divergence of alleles in nested subsamples from finite populations. Genetics **113:** 1057–1076.

MORAN, P. A. P., 1975 Wandering distributions and the electrophoretic profile. Theor. Popul. Biol. **8:** 318–330.

O'BRIEN, P., 1982 Allele frequencies in a multidimensional Wright-Fisher model with general mutation. J. Math. Biol. **15:** 227–237.

OHTA, T., and M. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretic detectable alleles in a finite population. Genet. Res. **22:** 201–204.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY and E. BOERWINKLE, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

SLATKIN, M., 1995 A measure of population subdivision based on microsatellite allele frequencies. Genetics: **139:** 457–462.

TAVARÉ, S., 1984 Lines of descent and genealogical processes, and their application in population genetic models. Theor. Pop. Biol. **26:** 119–164.

TODD, J. A., T. J. AITMAN, R. J. CORNALL, S. GHOSH, J. R. S. HALL *et al.*, 1991 Genetic analysis of autoimmune type 1 diabetes mellitus in mice. Nature **351:** 542–547.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

WEHRHAHN, C., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

Communicating editor: W. J. EWENS

APPENDIX A

It is possible to write down a recursion for the probability of observing a given sample by considering the underlying genealogy of the sample and summing over all possible previous states looking back in time. This type of recursion has previously been applied by HUDSON and KAPLAN (1986) and GRIFFITHS and TAVARÉ (1994b) among others. It is obtained by conditioning on the last event occurring before the present either being a mutation or a coalescence. Then, the sample probability $q(\boldsymbol{\nu})$ is given by the sum over all possible previous states of the probability of being in that state multiplied by the transition probability from that state to the present state (GRIFFITHS and TAVARÉ 1994a). Under the one step mutation model we get

$$q(\boldsymbol{\nu}) = \frac{\theta}{n+\theta-1} \sum_{i,j=i\pm1,n_j>0} \frac{n_i+1}{n} \frac{1}{2} q(\boldsymbol{\nu}+\mathbf{e}_i-\mathbf{e}_j) + \frac{n-1}{n+\theta-1} \sum_{n_j>0} \frac{n_j-1}{n-1} q(\boldsymbol{\nu}-\mathbf{e}_j), \quad \text{(A1)}$$

where $\mathbf{e}_i$ is a unit vector that adds 1 from entry i in $\boldsymbol{\nu}$. This recursion is easy to derive by realizing that $\theta/(n+\theta-1)$ is the probability that the last event before the present is a mutation given that either a mutation or a coalescent event happened, $(n-1)/(n+\theta-1)$ is the probability that a coalescent event happened given that either a mutation or coalescent happened, $(n_i+1)/n$ is the probability that a mutation occurred from allele i given that a mutation occurred, and $(n_j-1)/(n-1)$ is the probability that two genes of state j coalesced given that a coalescent occurred. One-half is the probability that a mutation occurred from state i to state j given that a mutation from state i occurred, *e.g.*, it is ijth entry of the mutation matrix (P) with entries $^1/_2$ if $j = i \pm 1$ and 0 in all other cases.

GRIFFITHS and TAVARÉ (1994b) have devised a method for evaluating likelihood functions based on recursions such as (A1). First note that (A1) has the form

$$q(y) = \sum_{x \in A} q(x)\, r_{xy}, \tag{A2}$$

where r_{xy} is the kernel in the recursion and A is the set of all states from which y can be obtained in one step. In principal a recursion such as (A2) can be evaluated by successively iterating the recursion until an expression is obtained that only contains terms for which $q(x)$ can be directly evaluated. In this manner

$$q(y) = \sum_{x \in V} q(x)\, r_{xy} + \sum_{x_1 \in V} r_{x_1 y} \left(\sum_{x \in A} q(x)\, r_{x x_1} \right) + \sum_{x_2 \in V} r_{x_2 y} \left(\sum_{x_1 \in A} r_{x_1 x_2} \left(\sum_{x \in A} q(x)\, r_{x x_1} \right) \right) + \cdots, \tag{A3}$$

where V is the set of values of x for which $q(x)$ can be directly evaluated. For example, in the case of microsatellites under the one-step mutation model, we could imagine iterating (A2) and hope that after many iterations only terms of $L(\theta)_2$ would be left. This approach would be feasible under an infinite sites model under which one may only loose alleles when going back in time. However, under a k-allele model or a stepwise mutation model, new alleles may actually be gained when going back in time, because mutations to previously existing alleles are allowed. Consequently, the recursion cannot in practice be evaluated directly by iteration. Instead, the method of GRIFFITHS and TAVARÉ (1994a) and GRIFFITHS and TAVARÉ (1994b) can be applied. In this method only some of the paths through the recursion are evaluated. These paths are chosen by defining a Markov chain going back in time with state space given by (A2) and with transition probabilities P_{yx} that are positive when r_{xy} in (A2) is positive. Next, we define a new function $f(x, y) = r_{xy}/P_{yx}$. Now (A3) can be rewritten as

$$q(y) = \sum_{x \in V} q(x) P_{yx} f(x, y) + \sum_{x_1 \in V} P_{y x_1} f(x_1, y) \left(\sum_{x \in A} q(x) P_{x_1 x} f(x, x_1) \right) + \sum_{x_2 \in V} P_{y x_2} f(x_2, y) \left(\sum_{x_1 \in A} P_{x_2 x_1} f(x_1, x_2) \left(\sum_{x \in A} q(x) P_{x_1 x} f(x, x_1) \right) \right) + \cdots. \tag{A4}$$

Therefore, if τ is the random number of states the chain passes through before hitting a state in V, and x_j is the jth state the chain passes through, we see that

$$q(y) = E\left[q(x_\tau) \prod_{j=1}^{\tau} f(x_{j-1}, x_j) \right]. \tag{A5}$$

Therefore, by simulating the Markov chain Ntimes $q(y)$ can be estimated as

$$\frac{1}{N} \sum_{i=1}^{N} \left[q(x_\tau) \prod_{j=1}^{\tau} f_i(x_{j-1}, x_j) \right]. \tag{A6}$$

The last step is simply to choose some appropriate transition probabilities to define the Markov chain. Following the general case described for the k-allele models by GRIFFITHS and TAVARÉ (1994a), we can establish the following chain for the one-step mutation model: Let

$$\frac{\theta(n_i + 1)}{2n(n + \theta + 1) f(\boldsymbol{\nu})}$$

be the probability of a transitions from $\boldsymbol{\nu}$ to $\boldsymbol{\nu} + \mathbf{e}_i - \mathbf{e}_j$ and

$$\frac{(n_j - 1)}{(n + \theta - 1) f(\boldsymbol{\nu})}$$

from $\boldsymbol{\nu}$ to $\boldsymbol{\nu} - \mathbf{e}_j$. Then,

$$f(x, y) = f(\boldsymbol{\nu}) = \frac{\theta}{2n(n + \theta - 1)} \sum_{i,j=i\pm 1} (n_i + 1) + \frac{1}{n + \theta - 1} \sum_{n_j > 0} (n_j - 1). \tag{A7}$$

We can now evaluate the likelihood function by simply performing simulations of this Markov chain while evaluating $f(\boldsymbol{\nu})$, until only two copies of the gene is left in the sample. Then (A6) provides an estimate of the likelihood for a particular value of θ.

The computational time may be large when applying this approach. One of the reasons for this is that many runs through the Markov chain effectively do not contribute anything to the likelihood value (some genealogies with mutation have very small probability). These are runs where very many mutations have occurred because the Markov chain has followed a less probable path. Computational time can be greatly reduced by truncating such runs. It was found empirically that runs could be stopped safely when more than the expectation plus four times the standard deviation in the number of mutations have occurred, that is, more than

$$\theta \sum_{i=1}^{n-1} \frac{1}{i} + 4 \sqrt{\theta \sum_{i=1}^{n-1} \frac{1}{i} + \theta^2 \sum_{i=1}^{n-1} \frac{1}{i^2}}$$

for all explored values of θ and n.

As shown by GRIFFITHS and TAVARÉ (1994a), computational time can be further reduced when evaluating the entire likelihood surface because $q(\boldsymbol{\nu})$ can be estimated for multiple values of θ simultaneously. This can be done by using a plausible value of θ to drive the simulations (θ_0) and then for each value of θ evaluate

$$f(\boldsymbol{\nu}) = f_{\theta_0}(\boldsymbol{\nu}) \frac{(n + \theta_0 - 1)}{(n + \theta - 1)},$$

if the last event when simulating the Markov chain was a coalescence and

$$f(\boldsymbol{\nu}) = f_{\theta_0}(\boldsymbol{\nu}) \frac{\theta(n + \theta_0 - 1)}{\theta_0(n + \theta - 1)},$$

if the last event was a mutation. However, a reasonable value of θ (θ_0) is needed to drive the simulations. When the difference between θ_0 and the evaluated value of θ increases so does the variance in the estimate. Fortunately, a reasonable value of θ_0 can be obtained by using the variance-based estimator of θ. A sample likelihood surface is shown in Figure A1. In these simulations the size of the state space was set to 100 and the distribution of allele classes was centered around 50. Changing the size of the state space from 100 to 200 did not have any observable effect on the estimation of $L(\theta)$ for the values of θ and n explored in this paper. Notice how the performance of the method depends on the choice of θ_0. However, also notice how well the true likelihood function is estimated in this particular example when the value obtained by the variance-based estimator is used for θ_0.

The method can also easily be applied to other mutational models such as (6) by modifying the recursion (A1) appropriately. Likewise, just as the full likelihood surface for θ can be evaluated in one set of simulations driven by a specific value of θ (θ_0), so can many values of the parameters in the mutational model be evaluated simultaneously by choosing one mutation matrix (P_0) to run the simulations and evaluating $(P_{(ij)}/P_{0(ij)})f(\boldsymbol{\nu})$ instead of $f(\boldsymbol{\nu})$ for each value of θ, if the last event was a mutation from state i to state j ($P_{(ij)}$ is ijth entry of P). This greatly reduces computational time. Expressions for $L(\theta)_2$ for other mutational models may be obtained by the method of O'BRIEN (1982).

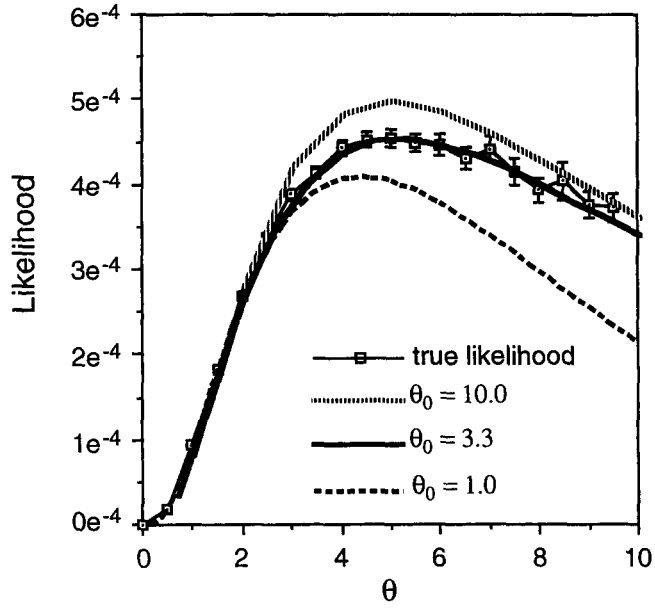

FIGURE A1.—The likelihood surface for a random simulated sample ($\theta = 5.0$) for a sample size of $n = 10$. The values for the true likelihood are obtained by evaluating the likelihood for each value of θ independently. The confidence intervals are obtained as ± twice the standard deviation in the estimate. The variance based estimate of θ is 3.3. Notice that the likelihood function obtained for $\theta_0 = 3.3$ in all cases is within the confidence values obtained when the likelihood value has been estimated independently for each point.