The Effects of Variable Mutation Rates Across Sites on the Phylogenetic Estimation of Effective Population Size or Mutation Rate of DNA Sequences

Hong-Wen Deng and Yun-Xin Fu

Human Genetics Center, School of Public Health, University of Texas, Houston, Texas 77225

Manuscript received March 29, 1996
Accepted for publication July 29, 1996

ABSTRACT

Multiple hits at some sites of human mitochondrial DNA sequences suggest that the commonly assumed infinite-sites model can be violated. Under the neutral Wright-Fisher model without recombination and population subdivision, we investigated, by computer simulations, the effect of multiple hits on the estimation of the essential parameter $\theta = 4N\mu$ by Fu’s UPBLUE procedure. We found that with moderate mutation rate heterogeneity, UPBLUE performs very well in terms of unbiasedness and efficiency. Under extreme mutation rate heterogeneity, if sample size is reasonably large (e.g., $>60$), UPBLUE is still very satisfactory; otherwise we developed a new correction equation. Given knowledge of the degree of mutation rate heterogeneity, the performance of UPBLUE with the new correction equation was tested to be fairly satisfactory: there is almost no bias and the sampling variance is only slightly higher than the theoretical minimum variance. Thus, with an appropriate correction, UPBLUE is relatively robust to the multiple hits. In genealogies reconstructed by UPGMA, we found that the total length of branches directly linked to the tips is underestimated, and those far away tend to be overestimated, while the total length of all branches is not biased.

POLYMORPHISM at the DNA nucleotide level is the ultimate resolution of information on genetic variation to study population evolution. A prominent parameter in stochastic population genetics theory of evolution is $\theta$, which is equal to $4N\mu$ for diploid and $2N\mu$ for haploid genomes, where $N$ is the effective population size and $\mu$ is the mutation rate per DNA sequence (gene, locus) per generation. Given $\theta$, $N$ can be estimated if $\mu$ is known or vice versa. Inferences about $\theta$ are commonly based on two quantities easily obtainable from a sample of $n$ DNA sequences from a population.

One is the average number of nucleotide differences in all pairwise comparisons ($\hat{\theta}$, known as the TAJIMA’S estimator), the other is the segregating (polymorphic) sites $K$ ($\hat{\theta} = K/a_n$, known as the WATTENSON’S estimator), where

$$a_n = 1 + \frac{1}{2} + \cdots + \frac{1}{(n-1)}.$$

Although easy to compute, these two estimators are not efficient, having large variances (WATTERSON 1975; TAJIMA 1983): especially, the variance of $\hat{\theta}$ does not diminish with an increasing sample size.

Recently, Fu (1994a,b; also see Li and Fu 1994) developed two more efficient estimation procedures for $\theta$, by making better use of the information in a sample. One is the phylogenetic estimator (Fu 1994a), which is based on the reconstructed genealogy for the sampled DNA sequences. The other is the frequency estimator (Fu 1994b), which is based on classifying segregating nucleotide sites into mutations of various classes, and making use of their frequencies. It is developed under the framework of a combination of coalescent theory, general linear model and Monte-Carlo integration, and can be used under a variety of population genetics models. However, under the neutral infinite-sites Wright-Fisher model without recombination and population subdivision, it is slightly less efficient than the phylogenetic estimator. The phylogenetic estimator uses more information from a sample and is nearly the most efficient estimator in practice (Fu 1994a,b). We thus study its statistical properties when some ideal assumptions necessary for its development are not fulfilled.

As is common to many estimators of $\theta$ (WATTERSON 1975; TAJIMA 1983; GRIFFITHS and TAVARE 1993, 1994), Fu’s phylogenetic estimator UPBLUE is based on the infinite-sites model (KIMURA 1969, 1971), which assumes that each mutation occurs at a site that has not been mutated before in the population. Thus by counting segregating sites, the number of mutations separating each pair of DNA sequences is known without error (but see DISCUSSION). However, this model is clearly violated by some DNA sequences such as those from human mitochondrial control region (e.g., VIGILANT et al. 1991; WARD et al. 1991; WAKELEY 1993), in which mutation rate varies across sites such that some sites are nearly immutable while others are highly mutable, and multiple hits are likely and frequent. Although KUHNER et al. (1995) does not use the infinite-sites model di-
rectly, multiple hits under mutation rate heterogeneity were not considered by them.

At the interspecific level, mutation rate heterogeneity is not considered by many procedures in molecular evolution, which may have serious consequences (Olsen 1987; Palumbi 1989; Jin and Nei 1990; Li et al. 1990; Hasegawa and Fujiwara 1993; Kuhner and Felsenstein 1994; Tateno et al. 1994). Therefore, their effects on molecular clock calibration, phylogenetic inference, distance calculation and transition bias estimation have attracted a growing number of research efforts (e.g., Van de Peer et al. 1993; Yang 1993; Wakeley 1994; Gu et al. 1995; Rzhetsky 1995). At the intraspecific level, multiple hits for a DNA sequence are less common unless there exists mutation rate heterogeneity. Although the effect of mutation rate heterogeneity is expected to be less serious compared with that at the interspecific level, it has not been studied for the estimation procedures of $\theta$.

Under the neutral Wright-Fisher model without recombination and population subdivision, we study the effects of multiple hits due to mutation rate heterogeneity among sites on Fu's (1994a) UPBLUE procedure of phylogenetic estimation of $\theta$. Then we develop a new correction procedure for the adverse effects of multiple hits. Finally, we test the performance of the new correction procedure.

MODEL OF MUTATION RATE HETEROGENETITY AMONG SITES

Gamma distribution has been used widely to describe substitution rate variation among different sites (Uzzell and Corbin 1971; Nei 1987; Kocher and Wilson 1991; Tamura and Nei 1993; Gu et al. 1995; Rzhetsky 1995; Aris-Brosou and Excoffier 1996). Letting $x$ be a random variable for the mutation rate at a nucleotide site, we use the following gamma distribution, which is determined by parameters $\alpha$ and $\beta$, to model $x$ across sites:

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \quad (x > 0),$$

(1)

where $\Gamma(\cdot)$ is the gamma function defined by $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$. The mean ($\lambda$) and the variance ($\sigma^2$) of the gamma distribution are respectively, $\lambda = \alpha \beta$ and $\sigma^2 = \alpha \beta^2$, and the relative magnitude of the rate heterogeneity can be measured by the coefficient of variation of $x$

$$CV_x = \left(\frac{\sigma^2}{\lambda^2}\right) = \frac{1}{\sqrt{\alpha}}.$$  

The smaller the $\alpha$, the larger the mutation rate heterogeneity. $\alpha = 0.1, 0.5$ and $1.0$ may represent extreme, moderate and small rate heterogeneity, respectively (Tamura and Nei 1993; Wakeley 1993; Tateno et al. 1994; Gu et al. 1995; Sullivan et al. 1995). In particular, when $\alpha = \infty$, the substitution rates are uniform among sites. The whole sequence mutation rate ($\mu$) is the sum of the mutation rates of all its $L$ nucleotide sites; i.e., $\mu = \sum_{i=1}^L x_i$, where $x_i$ is the mutation rate of the $i$th site. It is well known that the sum (such as $\mu$ here) of $L$ gamma-distributed random variables of parameters $\alpha$ and $\beta$, also has a gamma distribution of parameters $L\alpha$ and $\beta$, with a coefficient of variation $CV = (1/\sqrt{L\alpha})$.

ESTIMATION OF $\theta$ WHEN THE GENEALOGY OF A SAMPLE IS KNOWN

By genealogy, we mean collectively the topology connecting the sequences in a sample to their most recent common ancestor, the branching orders in the topology and the number of mutations on each branch of the topology. The genealogy of a sample can be reconstructed by the existing phylogenetic reconstruction methods, and it is possible that the reconstruction is of high resolution. The estimator of $\theta$ under a known genealogy thus represents the best that one can aim to achieve in practice. Therefore, studying such an estimator can serve as a starting point to estimate $\theta$ from an estimated genealogy as in Fu (1994a).

Theoretical background: Let $t_i$ be the time duration (the number of generations) required for $k$ sequences to coalesce into $k - 1$ sequences ($k > 1$) (Figure 1). Under the neutral Wright-Fisher model ($N_e$ remains constant, all mutations are selectively neutral), without recombination and population subdivision, $t_i$ is a random variable following an exponential distribution (Hudson 1982; Kingman 1982; Tajima 1983)

$$f(t_i) = \frac{k(k-1)}{4N_e} \exp \left(-\frac{k(k-1)}{4N_e} t_i \right).$$

(2)

Among the several definitions of $N_e$ (Ewens 1979), inbreeding effective population size seems to be the most suitable one in Equation 2 (Fu 1994a).

There are exactly $2(n - 1)$ branches in a genealogy of a DNA sequences, whose topology can be completely characterized by the following $2(n - 1)^2$ index variables (Fu 1994a). For branch $i$, we have $n - 1$ index variables $s_{ak}$ ($k = 2, \ldots, n$) such that $s_{a1} = 1$ if the branch has a segment during the time $t_i$ and $s_{a1} = 0$ otherwise. For illustration, the topology in Figure 1 is completely characterized by the 18 index variables (Table 1).

Assume the number of mutations on branch $i$ ($n_i$) follows a Poisson distribution with parameter $\mu t_i$, where $t_i$ is the time length of branch $i$, and

$$l_i = \sum_{k=2}^n s_{ak},$$

(3)

The mean and the variance of $n_i$ are, respectively (Fu 1994a),

$$E(n_i) = \theta \omega_i,$$

(4)

$$\sigma^2_{n_i} = \omega \beta + \phi \beta^2.$$  

(5)
and the covariance of the numbers of mutations on branch \(i\) and \(j\) is \((\text{Fu 1994a})\)

\[
\text{Cov}(n_i, n_j) = \phi \theta^2,
\]

(6)

where

\[
\omega_i = \sum_{k=2}^{n} \frac{s_k}{k(k-1)}, \quad \phi = \sum_{k=2}^{n} \frac{s_k \theta^2}{(k-1)^2}.
\]

(7)

The means of \(n_i\)’s are linear functions of \(\theta\), the variance among them are quadratic functions of \(\theta\). Thus Fu’s (1994a) iterative BLUE (best linear unbiased estimator) procedure can be applied to these \(2(n - 1)\) variables to obtain \(\theta\), the BLUE of \(\theta\). The lower bound of the variances of all possible unbiased estimator of \(\theta\) is \((\text{Fu and Li 1993})\)

\[
V_{\min} = \frac{\theta}{\sum_{k=1}^{n-1} \frac{1}{\theta + k}}.
\]

(8)

The efficiency of the BLUE procedure under multiple hits can be measured against \(V_{\min}\).

**Simulations:** Genealogies were simulated for DNA sequences with multiple hits due to mutation rate heterogeneity across sites, and DNA sequence evolution was monitored along branches on each genealogy.

1. A DNA sequence of length \(L\) was generated randomly as the most recent common ancestral sequence of the \(n\) sequences sampled from a population.

2. Mutation rate at each site \(\mu_i (l = 1, 2, \ldots, L)\) was determined by random sampling from a gamma distribution of parameters \(\alpha\) and \(\beta\). The mutation rate for the whole sequence \(\mu\) was equal to \(\Sigma_{l=1}^{L} \mu_i\).

3. For a given sample size \(n\), \(\mu\), and \(N_s\), a number of genealogies were randomly constructed according to the coalescent theory (Hudson 1982; Kingman 1982; Tajima 1983). Briefly, the topology of the genealogy was constructed, from the \(n\) nodes (representing \(n\) sequences sampled) backward in time, by randomly joining two nodes at a time, until only one (representing the most recent common ancestor) was left. The time interval between two successive node joining events \((t_k, k = 2, 3, \ldots, n)\) was determined by a random variable from the exponential distribution of Equation 2.

4. For each genealogy, \(s_k\)’s were inferred. The values of the time length of all the branches \(l_i\)’s were computed by Equation 3, and the \(\omega\)’s and \(\phi\)’s by Equation 7.

5. For branch \(i\) \((i = 1, 2, \ldots, 2n - 2)\), the mutations at each site were determined by a Poisson process, with its parameter being the product of \(l_i\) and \(\mu_i\) \((l = 1, 2, \ldots, L)\). To focus our study on the effects of multiple hits, we chose to use the simplest [Jukes-Cantor’s (1969) one-parameter] mutation model, in which, given a mutation event, the original nucleotide has an equal probability to mutate to the other three.

6. The iterative BLUE procedure of Fu (1994a) was applied to these \(2(n - 1)\) random variables \((n_i\)’s) to obtain \(\theta\), together with different variance estimates (Table 2).

**TABLE 1**

| Index variable \(s_k\)'s for the genealogy in Figure 1 |
|-----------------|---|---|---|
| \(i\)          | \(2\) | \(3\) | \(4\) |
| 1              | 0   | 1   | 1   |
| 2              | 0   | 0   | 1   |
| 3              | 0   | 0   | 1   |
| 4              | 1   | 1   | 1   |
| 5              | 0   | 1   | 0   |
| 6              | 1   | 0   | 0   |

**FIGURE 1**—An example of the topology of a genealogy of four DNA sequences. The branches are numbered consecutively from 1 through 6. The total time duration is divided into three intervals \(t_6, t_3\) and \(t_4\), according to the branching events and the existing number of sequences during the time intervals.
assuming a known genealogy, the mutation events are each pair of DNA sequences are known accurately by considering genetic distance between interspecific sequences. analogies, the BLUE performs as well as under the infinite-sites model. There is no bias in estimating the theoretical variance derived under the infinite-sites model, the mutation events separating sites model. 6' is the average of 2000 simulated genealogies. For simulations in this table, sequence length \( L = 1000 \), \( N_s = 1000 \), \( \theta = 4 N_s \mu \), where \( \mu \) is the sequence mutation rate, determined as described in the text. Multiple hits were observed in simulations for all the parameter sets in this and the following tables.

\[
\alpha = 1.0, 5.0, 0.5, 0.1, 0.01, 1.0, 5.0, 0.5, 0.1, 0.01
\]
\[
\beta = 1.3 \times 10^{-6}, 2.6 \times 10^{-6}, 1.0 \times 10^{-5}, 1.5 \times 10^{-5}, 6.0 \times 10^{-5}
\]

\( n \) is the sample size of DNA sequences. For simulations in this table, sequence length \( L = 1000 \), \( N_s = 1000 \), \( \theta = 4 N_s \mu \), where \( \mu \) is the sequence mutation rate, determined as described in the text. Multiple hits were observed in simulations for all the parameter sets in this and the following tables.

### Results
From Table 2, it is evident that under multiple hits due to the rate heterogeneity, with known genealogies, the BLUE performs as well as under the infinite-sites model. There is no bias in \( \theta \) estimation, and the sampling variance is not different from \( V_{\text{min}} \) and the theoretical variance derived under the infinite-sites model (\( V \)). These results should not be surprising. In infinite-sites model, the mutation events separating each pair of DNA sequences are known accurately by counting the number of segregating sites. However, assuming a known genealogy, the mutation events are known exactly, thus multiple hits per se do not pose any problem in \( \theta \) estimation.

### THE PHYLLOGENETIC ESTIMATOR OF \( \theta \)
In reality, the genealogy of sampled DNA sequences is unknown and must be estimated to apply the BLUE to estimate \( \theta \). Under the infinite-sites model, we can obtain an accurate distance matrix, upon which treeing methods such as UPGMA (unweighted pair-group method with arithmetic mean, Nei 1987) can be applied. The genealogy reconstruction by UPGMA will cause an under estimation of \( \theta \), which can be corrected by (Fu 1994a)

\[
\hat{\theta} = \left( 0.0335v_n - 2 + 0.998v_{\hat{\theta}_U} \right)^2,
\]

where \( \hat{\theta}_U \) is the estimator obtained by applying BLUE to the genealogy reconstructed by UPGMA. \( \hat{\theta} \) is shown to be nearly unbiased with a variance close to \( V_{\text{min}} \) (Fu 1994a).

With multiple hits, obtaining an accurate distance matrix is not easy. There are several studies on recovering genetic distance between interspecific sequences (Golding 1983; Tamura and Nei 1993; Gu et al. 1995; Rzhetsky 1995) in the presence of multiple hits under rate heterogeneity. They either depend on the inferred parameters of the underlying distribution of mutation rates across sites, or are limited to a few sequences due to the demanding computational requirement. Unbiased estimation of the parameters of the underlying distribution of mutation rates across sites is difficult. For studies of within-population evolution, the number of sequences sampled is often on the order of dozens, or even one hundred (e.g., Vigilant et al. 1991) and the divergence among them is often relatively small (compared to the interspecific data). In the absence of a simple method to compute distance with multiple hits under rate heterogeneity, Jukes-Cantor's (1969) method (which assumes constant mutation rates across sites) is chosen. Hence, the bias in estimating distance, thus likely \( \theta \) is expected. The question is how serious the bias is going to be, i.e., its direction and magnitude. Furthermore, can the bias be satisfactorily corrected?

### Simulations
Random determination of mutation rates across sites from a gamma distribution, thus the \( \mu \), is likely to affect the degree of bias of distance and \( \theta \) estimations if the sequence length \( (L) \) is not long enough. Therefore, for a given gamma distribution, a number of \( \theta \)'s \( (N_\theta) \) were simulated for the DNA sequence. Each \( \theta \) was simulated as described in the previous section. For a given sample size \( n \), \( \theta \) and \( N_\theta \), a number of genealogies \( (N_s) \) were randomly constructed according to the coalescent theory (Hudson, 1982; Kingman 1982; Tajima 1983) unless otherwise specified, \( L = 600 \), \( N_s = 20 \), and \( N_\theta = 400 \) in all the following simulations. The \( n \) sequences at the tips of the genealogy (representing those sampled from the population)
were determined as described in the previous section. We applied JUKES-CANTOR’S (1969) formulae to these sequences to reconstruct the distance matrix. JUKES-CANTOR’S (1969) multiple mutation correction formula is

\[ K = -\lambda \ln (1 - \frac{K}{p}), \]

where \( K \) is the number of substitutions per site since the divergence of the two sequences, and \( p \) is the proportion of different nucleotides between them.

There are a number of treeing methods for reconstructing phylogenetic relationship among different species (NEI 1987). Their relative performance for reconstructing genealogy from within-population DNA sequence data is not clear. The maximum likelihood method with a molecular clock may be the best choice from a theoretical point of view, but is computationally impractical when \( n \) is large or many genealogies have to be studied. The UPGMA is computationally simple and only requires that the evolution rates of different lineages are constant, which is met in this study and most likely by DNA sequences within populations or species. Thus, it was applied to the distance matrix to reconstruct the genealogy, from which \( s_k \)'s were inferred, \( l_k \)'s were thus computed by Equation 3, and \( \omega_i \)'s and \( d_j \)'s by Equation 7. The branch lengths of the reconstructed genealogy \( (n_i)'s \) were taken to be the expected number of mutations on the branches. Fu’s (1994a) BLUE procedure was applied to these 2\((n - 1)\) random variables \( (n_i)'s \) to obtain the BLUE estimator of \( \theta \) with the UPGMA-reconstructed genealogy \( (\hat{\theta}_i) \), and the corrected estimator \( \hat{\theta} \) by Equation 9. By one-way ANOVA, the variation of \( \hat{\theta}_i \) and \( \hat{\theta} \) was partitioned into that due to different \( \theta_i(V_i) \), and that due to different genealogy \( (V_0) \).

The distance: Genealogy reconstruction by UPGMA per se does not result in an underestimation of the total branch length of the genealogy \( (d) \). However, it underestimates the total length of branches that are directly linked to the tips of the genealogy and overestimates those that are far away from the tips (Figure 2). This is because UPGMA, at each step, joins two nodes of minimum distance that are not necessarily the two nodes that coalesced, due to stochastic mutational occurrences. Therefore, UPGMA tends to underestimate the lengths of branches closer to the tips and overestimate those closer to the root. Since lengths of branches closer to the tips of the genealogy tend to have smaller variances (Equations 5 and 7), they have higher weight on the outcome of the BLUE procedure. This explains why the UPGMA does not cause a biased estimation of \( d \), but when combined with BLUE procedure (UPBLUE), results in an underestimate of \( \theta \), even under the infinite-sites model (Fu 1994a).

The distance computed by JUKES-CANTOR’S formulae underestimate the true distance, as reflected by the underestimate of the total branch length of the reconstructed genealogy \( (\hat{d}) \) relative to \( d \) (Table 3). The degree of bias of \( \hat{d} \) depends on the \( \alpha \), \( \theta \), and \( n \). Everything else being about equal, the smaller the \( \alpha \), the larger the underestimation. It is easy to understand, since the smaller the \( \alpha \), the more heterogeneous are the mutation rates across sites. It means that, for the same \( \theta \), more sites will have higher mutation rates and thus more frequently will the multiple hits happen. A similar argument can explain that everything else being the same, the higher the \( \theta \), the larger the underestimation of \( d \). Everything else being the same, the larger the sample size \( n \) is, the larger the underestimation of \( d \). This is because with an increasing \( n \), more sequences with multiple hits will likely be included in the sample, resulting in a larger degree of underestimation of mutation events. Also we note, for a sequence of 600 nucleotides, the variance of the ratio \( (\hat{d}/d) \) is largely due to random genealogy, suggesting that random determination of \( \theta \) from a gamma distribution has minor effects on \( \hat{d}/d \), if the sequence length \( (L) \) is of moderate size.

The uncorrected estimator \( \hat{\theta}_{22} \): There is bias (underestimation) for every simulation parameter set (Table 4). This is expected, because even under the infinite-sites model, genealogy reconstruction alone will result in underestimation of \( \theta \) (Fu 1994a); with multiple hits, underestimation of \( d \) by \( \hat{d} \) will only cause further under-

\[ \text{FIGURE 2.—The relationship of the ratio of the total number of } \mu \text{ of mutations of the UPGMA reconstructed genealogy to that of the true genealogy (Y-axis) and the mutation size } i \text{ (X-axis). A mutation is of size } i \text{ (Fu 1994b) if it is has exactly } i \text{ descendents in the sample. The larger the } i \text{ of a mutation size, the further the branch where it occurs tends to be away from the tips of the genealogy, vice versa. Each dot in the plots is the mean of 10,000 simulations, in which the distance between each pair of sequences is known without error. The deviation of the ratio from 1.0 is totally due to genealogy reconstruction by UPGMA. The total branch length of the reconstructed genealogy equals that of the true genealogy in the above 10,000 simulations. For simulations in this plot, } n \text{ is 10, } \theta \text{ is } \sim 5.0, L \text{ is 600, and } N \text{ is 1000.} \]
TABLE 3

<table>
<thead>
<tr>
<th>α</th>
<th>β</th>
<th>d̄</th>
<th>d̄/d</th>
<th>V̄(d̄/d)</th>
<th>V̄(d̄/d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.1</td>
<td>4.0E-5</td>
<td>9.35</td>
<td>0.88</td>
<td>3.5E-4</td>
<td>6.2E-3</td>
</tr>
<tr>
<td>0.5</td>
<td>8.0E-6</td>
<td>9.58</td>
<td>0.95</td>
<td>0.0E-4</td>
<td>2.5E-3</td>
</tr>
<tr>
<td>1.0</td>
<td>4.0E-6</td>
<td>9.45</td>
<td>0.97</td>
<td>0.0E-4</td>
<td>2.3E-3</td>
</tr>
<tr>
<td>2.0</td>
<td>2.0E-6</td>
<td>9.65</td>
<td>0.98</td>
<td>0.0E-4</td>
<td>1.9E-3</td>
</tr>
</tbody>
</table>

B. The dependency of distance estimation bias on β

<table>
<thead>
<tr>
<th>α, β</th>
<th>d̄</th>
<th>d̄/d</th>
<th>V̄(d̄/d)</th>
<th>V̄(d̄/d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1E-5</td>
<td>5.27</td>
<td>0.91</td>
<td>0.0E-3</td>
<td>6.0E-3</td>
</tr>
<tr>
<td>4.2E-5</td>
<td>9.77</td>
<td>0.88</td>
<td>0.2E-3</td>
<td>6.2E-3</td>
</tr>
<tr>
<td>8.5E-5</td>
<td>19.90</td>
<td>0.80</td>
<td>0.4E-3</td>
<td>8.2E-3</td>
</tr>
<tr>
<td>1.3E-4</td>
<td>31.55</td>
<td>0.75</td>
<td>0.9E-3</td>
<td>9.1E-3</td>
</tr>
<tr>
<td>1.7E-4</td>
<td>41.64</td>
<td>0.71</td>
<td>0.7E-3</td>
<td>9.7E-3</td>
</tr>
</tbody>
</table>

C. The dependency of distance estimation bias on n

<table>
<thead>
<tr>
<th>α, β</th>
<th>d̄</th>
<th>d̄/d</th>
<th>V̄(d̄/d)</th>
<th>V̄(d̄/d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.71</td>
<td>5</td>
<td>0.88</td>
<td>0.5E-3</td>
<td>11.5E-3</td>
</tr>
<tr>
<td>10.65</td>
<td>10</td>
<td>0.87</td>
<td>0.4E-4</td>
<td>8.3E-3</td>
</tr>
<tr>
<td>10.34</td>
<td>20</td>
<td>0.87</td>
<td>0.2E-3</td>
<td>6.4E-3</td>
</tr>
<tr>
<td>10.65</td>
<td>30</td>
<td>0.86</td>
<td>0.5E-3</td>
<td>5.7E-3</td>
</tr>
<tr>
<td>9.88</td>
<td>40</td>
<td>0.86</td>
<td>0.2E-3</td>
<td>5.3E-3</td>
</tr>
<tr>
<td>10.75</td>
<td>60</td>
<td>0.86</td>
<td>0.4E-3</td>
<td>4.4E-3</td>
</tr>
<tr>
<td>10.70</td>
<td>80</td>
<td>0.86</td>
<td>0.1E-3</td>
<td>3.9E-3</td>
</tr>
<tr>
<td>10.76</td>
<td>100</td>
<td>0.84</td>
<td>0.2E-3</td>
<td>3.9E-3</td>
</tr>
</tbody>
</table>

d̄ is the total length of the true genealogy, d̄ is that of the reconstructed genealogy. For a given gamma distribution, 20
θ̂'s were simulated with mean θ̂. For each simulated θ̂, 400
genealogies were simulated. d̄/d is the mean of the d̄/d over
all simulated samples. The total variation of d̄/d is partitioned,
by one-way ANOVA, to the components due to θ̂ (V̄(d̄/d))
and different genealogy (V̄(d̄/d)). For simulations in this
table, L = 600, and N = 1000.

The corrected estimator θ̂ with Equation 9: It is generally
underestimated, but not always (Table 4). For the simulated parameter set in Table
4, the trends of the degree of underestimation of θ by θ̄V are similar to those of the distance estimation (Table 3). Everything else being about the same, the smaller the α, or the higher the θ, or the larger the n, the larger the underestimation of θ by θ̄V.

Correlation for the mean and variance of estimation of θ

Correction for the mean: Simulations were performed for different parameter combinations of α, θ, and n, with other conditions being the same as described in the last section. Some results are shown in Figure 3. Clearly, θ̄V is, on average, an underestimate of θ. The degree of underestimation is a function of n, the mean of θ̂'s (θ̂) and α of the gamma distribution. Since an analytical approach is difficult and not clear to us as to how to proceed at present, we chose to construct a correction based on the simulated data as in Fu (1994a). A regression analysis found that the following regression equation summarizes remarkably well
(H2 = 99.99%) the relationship among n, θ̄, α and the
mean of θ̂V (Figure 3):
TABLE 4

Estimation of \( \theta \) with the UPGMA reconstructed genealogy

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \beta )</th>
<th>( \bar{\theta} )</th>
<th>( \bar{\theta}_U )</th>
<th>( V_e(\bar{\theta}_U) )</th>
<th>( V_s(\bar{\theta}_U) )</th>
<th>( \bar{\theta} )</th>
<th>( V_e(\bar{\theta}) )</th>
<th>( V_s(\bar{\theta}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>4.0E-5</td>
<td>9.35</td>
<td>7.95</td>
<td>1.18</td>
<td>7.00</td>
<td>8.73</td>
<td>1.29</td>
<td>7.67</td>
</tr>
<tr>
<td>0.5</td>
<td>8.0E-6</td>
<td>9.58</td>
<td>8.60</td>
<td>0.34</td>
<td>8.03</td>
<td>9.40</td>
<td>0.37</td>
<td>8.76</td>
</tr>
<tr>
<td>1.0</td>
<td>4.0E-6</td>
<td>9.45</td>
<td>8.57</td>
<td>0.09</td>
<td>8.24</td>
<td>9.37</td>
<td>0.10</td>
<td>8.99</td>
</tr>
<tr>
<td>2.0</td>
<td>2.0E-6</td>
<td>9.65</td>
<td>8.86</td>
<td>0.03</td>
<td>8.71</td>
<td>9.68</td>
<td>0.03</td>
<td>9.49</td>
</tr>
</tbody>
</table>

A. The dependency of \( \theta \) estimation on \( \alpha \) (\( n = 20 \))

\[ \bar{\theta}_U = \left( -0.0113\sqrt{n} - 2 + 0.971\sqrt{\bar{\theta}} ight) + \frac{0.0168}{\sqrt{n}\alpha} - \frac{0.00583}{n^*\alpha} \cdot \sqrt{\bar{\theta}} \quad (10) \]

Thus, one can obtain an unbiased (or nearly unbiased) estimate of \( \theta \) by the following equation:

\[ \hat{\theta} = \left( \frac{0.0113\sqrt{n} - 2 + \sqrt{\bar{\theta}_U} - 0.0168}{0.971 - \frac{0.00583}{\sqrt{n}\alpha}} \right)^2. \quad (11) \]

Note that in the above equations, the larger the \( \alpha \), the smaller are the terms involving \( \alpha \), which is consistent with the previous simulation results that the larger the \( \alpha \), the smaller the bias introduced by multiple hits under the rate heterogeneity in \( \theta \) estimation. It is noted that even if \( \alpha \) tends to infinity, the coefficients of \( \sqrt{n} - 2 \) and \( \sqrt{\bar{\theta}_U} \) in Equations 9 and 11 do not agree with each other exactly. This is largely because Equation 11 was constructed for the distance corrected for multiple hits by \textsc{Jukes-Cantor}'s (1969) method when \( n \leq 60 \), while Equation 9 was for the true distance when \( n \leq 100 \).

Theoretical variance of the estimation of \( \theta \): In the simulations, the sampling variance of \( \bar{\theta} \) comes from two sources: one is the random determination of \( \theta \) from...
the gamma distribution \((V_g(\theta))\), the other is the random genealogy determination \((V_d(\theta))\). Phylogeny reconstruction by UPGMA alone does not inflate the sampling variance of \(\hat{\theta}_u\) (Fu 1994a). The approximate theoretical variance of \(\hat{\theta}\) computed by Equation 9 of Fu (1994a), using the variance-covariance structure of Equations 5 and 6, is for constant \(p\), thus only approximates \(V_g(\theta)\). However, the total sampling variance of \(\hat{\theta}\) in the simulation may be approximated by taking into consideration that \(p\) is variable. Under the variable \(p\), the corresponding Equations 4–6 are as follows:

\[
E(n_i) = E_p[E(n_i/p)] = \omega_i4N\sigma(E(\mu)),
\]

\[
\sigma^2_{n_i} = E_p[E(n_i^2/p)] - E^2(n_i) = 4N\sigma(E(\mu))\omega_i + \phi_p(4N\sigma)^2
\]

\[
\times E(\mu^2) + \omega_i^2(4N\sigma)^2(E(\mu^2) - E^2(\mu)) = \omega_i 4N\sigma(\mu) + \phi_p(4N\sigma)^2 E(\mu^2) + \omega_i^2(4N\sigma)^2 \sigma^2(\mu),
\]

\[
\text{Cov}(n_i, n_j) = E_p[E(n_i n_j/p)] - E(n_i)E(n_j) = \phi_p(4N\sigma)^2
\]

\[
\times E(\mu^2) + \omega_i\omega_j(4N\sigma)^2(E(\mu^2) - E^2(\mu)) = \phi_p(4N\sigma)^2
\]

\[
\times E(\mu^2) + \omega_i\omega_j(4N\sigma)^2 \sigma^2(\mu),
\]

where \(\omega_i\)'s and \(\phi_p\)'s are determined by Equation 7. Note, if \(\mu\) is constant, \(\sigma^2(\mu) = 0, E(\mu^2) = E^2(\mu)\), Equations 12–14 recover Equations 4–6. If the mutation rates across sites follow a gamma distribution with parameters \(\alpha\) and \(\beta\) and the sequence length is \(L\), \(\mu\) has a gamma distribution of parameters \(L\alpha\) and \(\beta\). Then in the above equations, \(E(\mu) = L\alpha\beta, \sigma^2(\mu) = L^2\alpha\beta^2\).

**Testing the new correction equations**: For a given gamma distribution and sample size \(n\), 20 \(\hat{\theta}\)'s were simulated, and 400 genealogies were simulated for each \(\theta\). The results are summarized in Table 5. It can be seen that after correction by Equation 11, \(\hat{\theta}\) is nearly unbi-
Mutation Rate and Population Size

TABLE 5
Properties of the BLUE estimates of \( \theta \) with the new correction Equation 11

<table>
<thead>
<tr>
<th>( \alpha )</th>
<th>( \bar{\theta} )</th>
<th>( n )</th>
<th>( \bar{\theta} )</th>
<th>( V_s(\bar{\theta}) )</th>
<th>( V_e(\bar{\theta}) )</th>
<th>( V_v )</th>
<th>( V_s )</th>
<th>( V_{min} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>5.47</td>
<td>20</td>
<td>5.61</td>
<td>0.512</td>
<td>4.590</td>
<td>4.994</td>
<td>4.346</td>
<td>3.851</td>
</tr>
<tr>
<td>5.02</td>
<td>40</td>
<td>5.13</td>
<td>0.405</td>
<td>2.778</td>
<td>3.137</td>
<td>2.627</td>
<td>2.415</td>
<td></td>
</tr>
<tr>
<td>9.77</td>
<td>40</td>
<td>10.00</td>
<td>1.003</td>
<td>7.296</td>
<td>8.688</td>
<td>6.822</td>
<td>6.263</td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td>5.12</td>
<td>20</td>
<td>5.14</td>
<td>0.264</td>
<td>3.873</td>
<td>4.072</td>
<td>3.800</td>
<td>3.481</td>
</tr>
<tr>
<td>4.91</td>
<td>40</td>
<td>4.85</td>
<td>0.130</td>
<td>2.517</td>
<td>2.644</td>
<td>2.418</td>
<td>2.335</td>
<td></td>
</tr>
<tr>
<td>9.95</td>
<td>20</td>
<td>10.13</td>
<td>0.757</td>
<td>10.628</td>
<td>11.702</td>
<td>10.697</td>
<td>9.640</td>
<td></td>
</tr>
<tr>
<td>10.09</td>
<td>40</td>
<td>10.11</td>
<td>0.776</td>
<td>7.178</td>
<td>7.854</td>
<td>6.907</td>
<td>6.560</td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td>4.94</td>
<td>20</td>
<td>5.04</td>
<td>0.140</td>
<td>5.656</td>
<td>5.376</td>
<td>5.222</td>
<td>5.081</td>
</tr>
<tr>
<td>4.85</td>
<td>40</td>
<td>4.76</td>
<td>0.184</td>
<td>2.407</td>
<td>2.560</td>
<td>2.355</td>
<td>2.290</td>
<td></td>
</tr>
<tr>
<td>9.95</td>
<td>20</td>
<td>10.07</td>
<td>0.934</td>
<td>10.690</td>
<td>11.286</td>
<td>10.621</td>
<td>9.642</td>
<td></td>
</tr>
<tr>
<td>10.14</td>
<td>40</td>
<td>10.02</td>
<td>0.363</td>
<td>7.076</td>
<td>7.425</td>
<td>6.806</td>
<td>6.592</td>
<td></td>
</tr>
<tr>
<td>0.4</td>
<td>5.17</td>
<td>20</td>
<td>5.14</td>
<td>0.127</td>
<td>3.813</td>
<td>3.913</td>
<td>3.778</td>
<td>3.528</td>
</tr>
<tr>
<td>5.03</td>
<td>40</td>
<td>4.95</td>
<td>0.119</td>
<td>2.560</td>
<td>2.601</td>
<td>2.484</td>
<td>2.418</td>
<td></td>
</tr>
<tr>
<td>10.10</td>
<td>20</td>
<td>10.13</td>
<td>0.404</td>
<td>10.639</td>
<td>11.185</td>
<td>10.684</td>
<td>9.864</td>
<td></td>
</tr>
<tr>
<td>10.14</td>
<td>40</td>
<td>10.01</td>
<td>0.332</td>
<td>7.154</td>
<td>7.261</td>
<td>6.798</td>
<td>6.598</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>5.10</td>
<td>20</td>
<td>5.07</td>
<td>0.072</td>
<td>3.805</td>
<td>3.810</td>
<td>3.705</td>
<td>3.463</td>
</tr>
<tr>
<td>4.95</td>
<td>40</td>
<td>4.85</td>
<td>0.055</td>
<td>2.493</td>
<td>2.482</td>
<td>2.393</td>
<td>2.366</td>
<td></td>
</tr>
<tr>
<td>9.98</td>
<td>20</td>
<td>10.01</td>
<td>0.146</td>
<td>10.624</td>
<td>10.878</td>
<td>10.487</td>
<td>9.653</td>
<td></td>
</tr>
</tbody>
</table>

The simulation conditions are as described in the legend to Table 4. \( V_v \) is the total variance of \( \theta \) when sequence mutation rate varies as in the simulation. It is computed by Equation 9 of Fu (1994a), using the variance-covariance structure determined by Equations 5 and 6, substituting \( \bar{\theta} \) for \( \theta \). \( V_v \) is the variance of \( \theta \) from constant mutation rate. It is similarly obtained, except that it uses the variance-covariance structure determined by Equations 5 and 6. \( \theta \) is the mean of \( \theta \) estimates using correction Equation 11. The other notations are the same as in Table 4.

The fairly small bias of \( \bar{\theta} \) is unlikely to be significant compared to its variance. The theoretical variance computed by using the variance-covariance matrix of Equations 13 and 14 (\( V_v \)) closely approximates the total sampling variance of \( \bar{\theta} \) (\( V_s(\bar{\theta}) + V_e(\bar{\theta}) \)), though slightly smaller. Similarly, the theoretical variance computed by using the variance-covariance matrix of Equations 5 and 6 (\( V_s \)) approximates \( V_e(\bar{\theta}) \), though smaller. However, \( V_e(\bar{\theta}) \) is always between \( V_v \) and \( V_s \) thus the latter two quantities may provide an interval estimate for \( V_e(\bar{\theta}) \). In practice, the mutation rate of a DNA sequence is normally constant within a population or a species. Thus in using UPBLUE, the sampling variance of \( \theta \) estimation only comes from random genealogy determination and may be approximated by \( V_v \); or at least, its upper and lower bounds may be provided by \( V_v \) and \( V_v \), respectively. Encouragingly, \( V_v(\bar{\theta}) \) is not much larger than the theoretical minimum variance (\( V_{min} \)), especially when \( n \) is not so small (\( n = 40 \)).

Therefore, when \( n \) is not large (\( n \approx 60 \)), and the rate heterogeneity is extreme (\( \alpha \approx 0.5 \)), the UPBLUE procedure with correction Equation 11 can provide a nearly unbiased estimation of \( \theta \), with its sampling variance very close to \( V_{min} \).

**DISCUSSION**

The present study revealed that, with multiple hits under mutation rate heterogeneity across sites, Fu's (1994a) UPBLUE procedure will sometimes result in a conservative (under-) estimation of the true \( \theta \). However, only with extreme mutation rate heterogeneity, plus a relatively small sample size \( n \), will the bias be significant. Even then, new correction equation can be constructed to be used in UPBLUE to yield a very efficient and nearly unbiased estimate of \( \theta \).

In practice, if \( n \) is large (e.g., \( n \approx 60 \)), Fu's (1994a) UPBLUE can be used directly to yield a nearly unbiased and efficient estimate of \( \theta \); otherwise, it is necessary to check if there are multiple hits due to the rate heterogeneity across sites. With multiple hits, if the distribution of the mutation changes per site follows a Poisson distribution, the rate heterogeneity is not suggested and Jukes-Cantor's method may be used to reconstruct the distance matrix, upon which Fu's (1994a) UPBLUE may be used directly. If mutation rates across sites conform to a gamma distribution, the number of the mutational changes per site should follow a negative binomial distribution (a well known result). Data analysis for within- and between-specific hypervariable DNA sequences often revealed such a pattern (e.g., Kocher and Wilson 1991; Tamura and Nei 1993; Wakeley 1993; Tateno et al. 1994; Sullivan et al. 1995). There have been several methods proposed to estimate \( \alpha \) of the gamma distribution (see Sullivan et al. 1995 for references). Once \( \alpha \) is estimated, UPBLUE with the correction Equation 11 and the variance-covariance structure determined by Equations 5 and 6 can be em...
ployed to obtain a nearly unbiased \( \theta \) estimate and its approximate sampling variance.

Since our main purpose is to investigate the qualitative effects of multiple hits under the rate heterogeneity on the phylogenetic estimation of \( \theta \), we used the simplest (Jukes-Cantor’s one-parameter) mutation model. However, the same logic and procedure should be easily extended to incorporate more complex mutation models in the study, such as Kimura’s (1980) two-parameter mutation model etc. If Jukes-Cantor’s one-parameter mutation model is seriously violated, the application of the correction Equations 11 with UPBLUE will likely result in a biased estimation. The investigation of the performance of UPBLUE of the phylogenetic estimator of \( \theta \) with more complex mutation models should be of some interest.

This study was performed under the neutral Wright-Fisher model without population subdivision and recombination, which some real samples of DNA sequences are likely to violate to a certain degree. When there is apparent population subdivision and recombination, the phylogenetic estimator of \( \theta \) may be misleading due to the difficulty of reconstructing genealogy (Hudson and Kaplan 1985; Fu 1994a). In this situation, the effects of multiple hits under the rate heterogeneity on \( \theta \) estimation can be investigated with Fu’s (1994b) mutation frequency estimator.

We investigated here the frequently used model of the mutation rate heterogeneity across sites, i.e., the gamma distribution. Other models of among-site rate variation were used by some other workers. One is the multiclass model, which classifies sites into several classes, within each of which the mutation rate is constant (Fitch and Margoliash 1967; Shoemaker and Fitch 1989; Hasegawa et al. 1993). Another one is the log-normal distribution (Olson 1987). Although these models differ from each other, the qualitative effects of multiple hits under the rate heterogeneity in estimating \( \theta \) should be similar. As demonstrated, mutation rate heterogeneity per se does not add an additional source of underestimation of \( \theta \) by Fu’s UPBLUE (section “Estimation of \( \theta \) when the genealogy of a sample is known”). It is the underestimation of genetic distance due to multiple hits under the rate heterogeneity that caused the additional bias. Thus, the correction Equation 11 may differ if different mutation rate variation models are used, but the qualitative conclusion of the present study should remain valid. That is, multiple hits will result in an underestimation (if any) of \( \theta \) with Fu’s (1994a) UPBLUE procedure.

As in Fu (1994a), under multiple hits, UPBLUE may be used in conjunction with the parsimony treeing method to yield a better estimate of \( \theta \). Although parsimony method partly recovers the number of mutations on sites under multiple hits, it generally underestimates it (Wakeley 1993), thus likely the \( \theta \) estimate. The substantial increase of computational time of the parsimony method relative to the UPGMA method also precludes it from the computer simulations in the present and Fu’s (1994a) studies, in which many genealogies and large samples are simulated. However, the use of the correction Equation 11 requires that the mutation rates across sites follow a gamma distribution and its parameter \( \alpha \) be estimated, which may not always be true or easy in practice. Thus, without the knowledge of the distribution of mutation rates across sites, Fu’s (1994a) approach of using UPBLUE in conjunction with the parsimony treeing method may be adopted.

Due to mutation rate heterogeneity, multiple hits are common in the human mitochondrial control region, in which the distribution of mutation rates across sites is fit adequately by a gamma distribution (Kocher and Wilson 1991; Tamura and Nei 1993; Wakeley 1993). For the sample of 63 American Indian sequences from mitochondrial control region (hypervariable region 1) (Ward et al. 1991), Fu’s (1994a) estimated \( \theta \) to be 13.32 by using UPBLUE in conjunction with parsimony. For this region, \( \alpha \) is estimated (by the largest data set) to be 0.47 (Wakeley 1993). By using UPBLUE with correction Equation 11 and assuming \( \alpha = 0.47 \), \( \theta \) is estimated to be 13.84, slightly higher than Fu’s (1994a) estimate. Interestingly, direct use of Fu’s (1994a) original UPBLUE procedure (i.e., using the number of nucleotide difference between two sequences as the distance between them directly), \( \theta \) is estimated to be 12.88, which is close to the above two estimates. Their differences are unlikely to be significant in comparison with their variances, which is \( \sim 7.70 \). This appears at a glance surprising because there are 26 segregating sites, while parsimony analysis shows that there are at least 41 mutations (Ward et al. 1991; Fu 1994a), so that the infinite-sites model is apparently violated. However, the truth is that the real requirement of Fu’s UPBLUE is to know the number of mutations between each pair of sequences, which strictly speaking does not require the assumption of the infinite-sites model. Since the sample size is reasonably large and \( \alpha \) is not too small, Fu’s UPBLUE needs little correction as demonstrated in this paper. This example demonstrates the robustness of Fu’s UPBLUE procedure in the presence of multiple hits under mutation rate heterogeneity.

We thank Drs. X. Gu, R. Adkins and two anonymous reviewers for helpful comments. The work was supported by First Award from National Institutes of Health to Y-X.F.

LITERATURE CITED


Mutation Rate and Population Size