A Phylogenetic Estimator of Effective Population Size or Mutation Rate

Yun-Xin Fu

Center for Demographic and Population Genetics, University of Texas, Houston, Texas 77225

Manuscript received July 12, 1993
Accepted for publication October 18, 1993

ABSTRACT

A new estimator of the essential parameter $\theta = 4N_e\mu$ from DNA polymorphism data is developed under the neutral Wright-Fisher model without recombination and population subdivision, where N_e is the effective population size and μ is the mutation rate per locus per generation. The new estimator has a variance only slightly larger than the minimum variance of all possible unbiased estimators of the parameter and is substantially smaller than that of any existing estimator. The high efficiency of the new estimator is achieved by making full use of phylogenetic information in a sample of DNA sequences from a population. An example of estimating θ by the new method is presented using the mitochondrial sequences from an American Indian population.

GENETIC variation at the nucleotide level is a powerful source of information for studying the evolution of a population. The quantity $\theta = 4N_e\mu$, where N_e is the effective size of the population and μ is the mutation rate per sequence (gene, locus) per generation, is an essential parameter because it determines the degree of polymorphism at the locus. The degree of success in our inference about the evolution of a population is measured to some extent by the accuracy of estimation of this essential parameter. The purpose of this paper is to develop an efficient estimator of θ under the neutral Wright-Fisher model without recombination and population subdivision. The estimation of θ becomes, on one hand, the estimation of the mutation rate, μ, when the effective population size, N_e, is known, and, on the other hand, the estimation of the effective population size, N_e, when the mutation rate, μ, is known.

There are two commonly used estimators of θ from a sample of n DNA sequences from a population. One is $\hat{K} = K/a_n$, where K is the number of segregating sites and a_n is given by

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

The other estimator is $\hat{\pi}$, the average number of nucleotide differences in all pairwise comparisons. These two estimates are unbiased under the infinite-site neutral Wright-Fisher model, that is under the assumptions that the population evolves according to the Wright-Fisher model with a constant effective size, that the number of sites in the sequences is very large so that every mutation occurs at a new site and that all mutations are selectively neutral. The variances of $\hat{K}$ and $\hat{\pi}$ were found respectively by WATTERSON (1975) and TAJIMA (1983):

$$\mathrm{Var}(\hat{K}) = \frac{\theta}{a_n} + \frac{\theta^2}{a_n^2}\sum_{k=1}^{n-1}\frac{1}{k^2}, \qquad (2)$$

$$\mathrm{Var}(\hat{\pi}) = \frac{n+1}{3(n-1)}\theta + \frac{2(n^2+n+3)}{9n(n-1)}\theta^2 \qquad (3)$$

where a_n is given by (1).

Estimating θ using either of the two estimators, in particular $\hat{K}$, is rather simple. However, the price for simplicity in computation is a large variance. Although attempts to improve the estimation of θ were made by STROBECK (1983) and AVISE *et al.* (1988), such efforts had been difficult because it was not clear how much improvement can be achieved. Recently, FELSENSTEIN (1992a) and FU and LI (1993a) showed that the efficiencies of these two estimators are actually very low. This is contrary to the common belief that by analogy to the number of alleles in a sample, the number of segregating sites should be a sufficient statistic for the parameter θ. Although this is not entirely wrong, the fact that the sufficiency requires an extremely large sample (FU and LI 1993a) makes it necessary to explore efficient methods for estimating θ. FELSENSTEIN (1992b) proposed one method to do so, but required an impractical amount of computation. The new method developed in this paper is, on the one hand, highly efficient by making full use of the phylogenetic information in a sample of DNA sequences and, on the other hand, practical because the computation can easily be carried out by a desktop computer.

THEORY

Let $m_1, \ldots, m_\gamma$ be random variables such that their expectations are all linear functions of θ, and that their variances and covariances are all quadratic functions of θ. Without loss of generality, we assume that

$$E(m_i) = \theta,$$
$$\mathrm{Cov}(m_i, m_j) = \sigma_{ij}\theta$$

where

$$\sigma_{ij} = \alpha_{ij} + \beta_{ij}\theta \tag{4}$$

and α_{ij} and β_{ij} are constants. We wish to find a linear function of the random variables $m_1, \ldots, m_\gamma$

$$m = \sum_k u_k m_k$$

such that m is an unbiased estimate of θ and that the variance of m is the minimum among all unbiased estimators of θ that are linear functions of the random variables $m_1, \ldots, m_\gamma$. That is, we want to find the linear minimum variance estimate of θ.

The variance of the linear function m is

$$\mathrm{Var}(m) = \theta \sum_{ij} u_i u_j \sigma_{ij}. \tag{5}$$

One can obtain the solutions of u_i's by solving the set of linear equations

$$\frac{\partial\ \mathrm{Var}(m)}{\partial u_k} = 0, \quad k = 1, \ldots, \gamma$$

with the constraint

$$u_1 + \ldots + u_\gamma = 1$$

because we require that $E(m) = \theta$. It can be shown that this set of linear equations is

$$\begin{cases} \sigma_{k\gamma} - \sigma_{\gamma\gamma} + \sum\limits_{i<\gamma} (\sigma_{ki} - \sigma_{\gamma i} - \sigma_{k\gamma} + \gamma_{\gamma\gamma})u_i = 0, \\ \qquad\qquad\qquad k = 1, \ldots, \gamma - 1, \\ u_1 + \ldots + u_\gamma = 1. \end{cases}$$

A more elegant yet equivalent approach is to use the theory of linear models. Let $\mathbf{m} = (m_1, \ldots, m_\gamma)^T$ (superscript T stands for transpose), $\mathbf{x} = (1, \ldots, 1)^T$ and $\boldsymbol{\epsilon} = (m_1 - \theta, \ldots, m_\gamma - \theta)^T$. Then $\mathbf{m}$ can be expressed by the linear model

$$\mathbf{m} = \theta\mathbf{x} + \boldsymbol{\epsilon}$$

and $\mathrm{Var}(\boldsymbol{\epsilon}) = \theta\mathbf{V}_\theta$ where

$$\mathbf{V}_\theta = \{\sigma_{ij}\}, \quad \text{for } i = 1, 2, \ldots, \gamma, \quad \text{and } j = 1, 2, \ldots, \gamma.$$

The estimate of θ by the method of generalized linear square [see, *e.g.*, Equation 38 of SEARLE (1982)] is therefore

$$\theta^* = \mathbf{u}^T\mathbf{m} \tag{6}$$

where the coefficient $\mathbf{u}^T = (u_1, \ldots, u_\gamma)$ is

$$\mathbf{u}^T = (\mathbf{x}^T\mathbf{V}_\theta^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{V}_\theta^{-1}. \tag{7}$$

Note that the scalar $\mathbf{x}^T\mathbf{V}_\theta^{-1}\mathbf{x}$ is simply the sum of all the elements of $\mathbf{V}_\theta^{-1}$ while $\mathbf{x}^T\mathbf{V}_\theta^{-1}$ is the vector with elements being the column-sums of $\mathbf{V}_\theta^{-1}$. Therefore, u_i is the ith column-sum of $\mathbf{V}_\theta^{-1}$ divided by the sum of all column-sums. However, Equation 6 cannot be used to obtain an estimate of θ directly because it requires the value of θ which is unknown. This difficulty can be circumvented by an iterative procedure.

Suppose an initial estimate of θ, denoted by $\theta_{(0)}$, is obtained. Then Equations 6 and 7 suggest that one can obtain a series of $\mathbf{u}_{(k)}$ and a series of $\theta_{(k)}$ by

$$\mathbf{u}_{(k)}^T = (\mathbf{x}^T\mathbf{V}_{\theta_{k-1}}^{-1}\mathbf{x})^{-1}\mathbf{x}^T\mathbf{V}_{\theta_{k-1}}^{-1},$$
$$\theta_{(k)} = \mathbf{u}_{(k)}^T\mathbf{m}.$$

If the series $\mathbf{u}_{(k)}(k = 1, 2, \ldots)$ converges, its limiting value $\mathbf{u}_{(\infty)}$ can be used as an estimate of $\mathbf{u}$ and

$$\tilde{\theta} = \theta_{(\infty)} \tag{8}$$

can be used as an estimate of θ. We shall refer to this as the best linear unbiased estimator (BLUE) procedure and $\tilde{\theta}$ as the BLUE of parameter θ. Although we are not able to prove that the series $\mathbf{u}_{(k)}$, $(k = 1, 2, \ldots)$ must converge, in the application of the BLUE procedure described later, we found that the series $\mathbf{u}_{(k)}(k = 1, 2, \ldots)$ not only always converges, but does so rapidly. When WATTERSON's estimate of θ is taken as the initial estimate of θ ($\theta_{(0)} = \hat{K}$), the iterative process usually needs no more than four cycles. That is, $\tilde{\theta} \approx \theta_{(4)}$.

It should be pointed out that although we intended to find an estimate of θ that is a linear function of $m_1, \ldots, m_\gamma$, strictly speaking, $\tilde{\theta}$ is not a linear function of $m_1, \ldots, m_\gamma$ because $\mathbf{u}_{(\infty)}$ depends on $\theta_{(\infty)}$ which is a function of $m_1, \ldots, m_\gamma$. This makes it difficult to obtain an exact sampling variance of $\tilde{\theta}$ and also implies that $\tilde{\theta}$ may not be unbiased. However, numerical results later will show that treating $\mathbf{u}$ as a vector of constants is appropriate for studying the sampling properties of $\tilde{\theta}$. Doing so, we have

$$\mathrm{Var}(\tilde{\theta}) = \mathbf{u}^T(\mathbf{M}_\alpha\theta + \theta^2\mathbf{M}_\beta)\mathbf{u}$$
$$= a_n'\theta + b_n'\theta^2$$

where $\mathbf{M}_\alpha = \{\alpha_{ij}\}$ for $i, j = 1, \ldots, \gamma$, $\mathbf{M}_\beta = \{\beta_{ij}\}$ for $i, j = 1, \ldots, \gamma$, $a_n' = \mathbf{u}^T\mathbf{M}_\alpha\mathbf{u}$ and $b_n' = \mathbf{u}^T\mathbf{M}_\beta\mathbf{u}$. Since the unbiased estimate of θ^2 is

$$\frac{\tilde{\theta}(\tilde{\theta} - a_n')}{1 + b_n'},$$

an approximately unbiased estimate of the variance of $\tilde{\theta}$ is

$$V_c = a_n'\tilde{\theta} + \frac{\tilde{\theta}(\tilde{\theta} - a_n')b_n'}{1 + b_n'}. \tag{9}$$

However, if we estimate θ^2 directly by $\tilde{\theta}^2$, we can take the advantage of Equation 7 and obtain an estimate of the variance of $\tilde{\theta}$ as

$$V_{nc} = \frac{\tilde{\theta}}{\mathbf{x}^T\mathbf{V}_{\tilde{\theta}}^{-1}\mathbf{x}}. \qquad (10)$$

A third method of estimating the variance of $\tilde{\theta}$ will be introduced later.

In the simplest case in which $\gamma = 2$, we have

$$\mathbf{V}_{\tilde{\theta}}^{-1} = \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}^2}\begin{pmatrix}\sigma_{22} & -\sigma_{12} \\ -\sigma_{12} & \sigma_{11}\end{pmatrix}.$$

Therefore

$$u_1 = \frac{\sigma_{22} - \sigma_{12}}{(\sigma_{11} - \sigma_{12}) + (\sigma_{22} - \sigma_{12})}, \qquad (11)$$

$$u_2 = \frac{\sigma_{11} - \sigma_{12}}{(\sigma_{11} - \sigma_{12}) + (\sigma_{22} - \sigma_{12})}. \qquad (12)$$

Although the exact values of u_1 and u_2 depend on the σ values, it is always true from the above two equations that the random variable with a smaller variance has a larger value for its coefficient u than that with a larger variance. In other words, the variable with a smaller variance has a higher weight on the outcome of estimation, which is naturally what one would expect.

ESTIMATION OF θ WHEN THE GENEALOGY OF A SAMPLE IS KNOWN

We assume in this section that the genealogy of genes in a random sample from a population is known. By genealogy, we mean collectively the topology connecting the genes in a sample to their most recent common ancestor, the order of branchings in the topology and the number of mutations on each branch of the topology. We study the estimation of θ under a known genealogy not only because it provides a method to estimate θ from an estimated genealogy but also because such a study provides the minimum variance of all possible unbiased estimates of θ. FU and LI (1993a) have derived one lower bound of the variances of all possible unbiased estimates, but their derivation was based on information that could never be fully recovered from sequence data. The genealogy of a sample can be estimated and there are chances that the reconstructions are perfect. Therefore, the best estimator under a known genealogy represents the most one could hope to achieve in practice. The variance of the best estimator will be a more realistic lower bound than the one we derived earlier.

Let the branching events be numbered successively so that the 1st branching event is the root and n − 1th is the most recent branching event. For convenience, the time when the sample was taken (the external nodes) is considered as the nth branching event (see Figure 1). Let t_k, commonly referred to as the k-coalescent time, be the time length (in terms of number of generations) between (k − 1)th and kth branching events. Then under the assumptions that the population from which the sample is drawn has a constant effective size N_e and evolves according to the Wright-Fisher model, that all mutations on the DNA sequences are selectively neutral and that there is no recombination, t_k is a random variable with exponential distribution and parameter $k(k-1)/(4N_e)$ (KINGMAN 1982; HUDSON 1982; TAJIMA 1983). Therefore

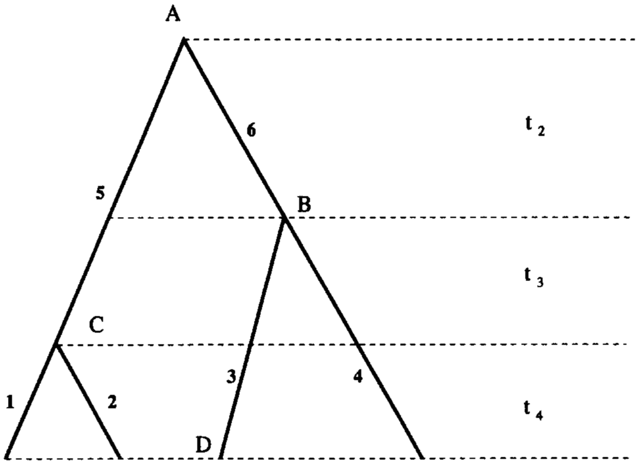

FIGURE 1.—An example of a topology of four sequences. The 1st, 2nd, 3rd and 4th branching events are A, B, C and D respectively.

$$E(t_k) = \frac{4N_e}{k(k-1)},$$

$$E(t_k^2) = 2\left[\frac{4N_e}{k(k-1)}\right]^2$$

where E stands for mathematical expectation. It should be noted that there are several definitions of effective population size (for example, see EWENS 1979) but for many models in population genetics they do not differ much; nevertheless, from the derivations of the distribution of coalescent times (KINGMAN 1982; HUDSON 1982; TAJIMA 1983), inbreeding effective population size seems to be the most suitable definition for N_e in this paper.

The genealogy of n genes has exactly $2(n-1)$ branches. For branch i, we define $n-1$ index variables $s_{ik}(k = 2, \ldots, n)$ such that $s_{ij} = 1$ if the branch has a segment between the $(j-1)$th and the jth branching events and $s_{ij} = 0$ otherwise. The topology of a genealogy of n genes is completely characterized by these $2(n-1)^2$ index variables. For example, the topology in Figure 1 has 18 index variables (Table 1).

Let l_i be the time length of branch i and n_k be the number of mutations occurred on branch i which is

TABLE 1

s_{ij} for the genealogy in Figure 1

	j		
i	2	3	4
1	0	0	1
2	0	0	1
3	0	1	1
4	0	1	1
5	1	1	0
6	1	0	0

assumed to follow a Poisson distribution with parameter μl_i. Then it is easy to see that

$$l_i = \sum_{k=2}^{n} s_{ik} t_k. \tag{13}$$

It follows that

$$\begin{aligned} E(n_i) &= \theta\omega_i, \\ \mathrm{Var}(n_i) &= E(n_i^2) - E^2(n_i) \\ &= \omega_i\theta + \left\{E\left(\sum_{k=2}^{n} s_{ik}t_k\right)^2 - \left[\sum_{k=2}^{n} \frac{s_{ik}}{k(k-1)}\right]^2\right\}\theta^2 \\ &= \omega_i\theta + \phi_{ii}\theta^2 \end{aligned}$$

where

$$\omega_i = \sum_{k=2}^{n} \frac{s_{ik}}{k(k-1)}, \tag{14}$$

$$\phi_{ij} = \sum_{k=2}^{n} \frac{s_{ik}s_{jk}}{k^2(k-1)^2}. \tag{15}$$

Furthermore, we have

$$\begin{aligned} \mathrm{Cov}(n_i, n_j) &= E(n_in_j) - E(n_i)E(n_j) \\ &= E_{l_i,l_j}(\mu^2 l_il_j) - E(n_i)E(n_j) \\ &= \mu^2 E\left(\sum s_{ik}t_k \sum s_{jk}t_k\right) - \omega_i\omega_j\theta^2 \\ &= \mu^2\{[\sum s_{ik}E(t_k)][\sum s_{jk}E(t_k)] + \sum s_{ik}s_{jk}E^2(t_k)\} \\ &\quad - \omega_i\omega_j\sigma^2 \\ &= \phi_{ij}\theta^2. \end{aligned}$$

Define

$$m_i = \frac{n_i}{\omega_i}, \quad i = 1, \ldots, 2(n-1) \tag{16}$$

Then $E(m_i) = \theta$ and the variance of m_i and the covariance between m_i and m_j are all quadratic functions of θ. Therefore, the BLUE procedure developed earlier can be applied to these $2(n-1)$ variables to obtain $\tilde{\theta}$, the BLUE of θ. Note that the α_{ij} and β_{ij} in (4) for these $2(n-1)$ variables are

$$\alpha_{ij} = \begin{cases} \dfrac{1}{\omega_i} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \tag{17}$$

$$\beta_{ij} = \frac{\phi_{ij}}{\omega_i\omega_j}. \tag{18}$$

When there are only two genes in the sample ($n = 2$), it is easy to see that

$$\omega_1 = \omega_2 = 1/2 \quad \text{and} \quad \phi_{11} = \phi_{12} = \phi_{22} = 1/4.$$

Therefore,

$$\sigma_{11} = \sigma_{22} = 2 + \theta, \quad \text{and} \quad \sigma_{12} = \theta$$

Putting these quantities into (11) and (12), we have

$$u_1 = u_2 = 1/2.$$

It follows that $\tilde{\theta} = \hat{K} = \hat{\pi}$ when $n = 2$.

The efficiency of the BLUE procedure for these $2(n-1)$ random variables $m_1, \ldots, m_{2(n-1)}$ can be measured against a lower bound of variances of all possible unbiased estimators of the parameter θ. One such lower bound from Equation 28 of FU and LI (1993a) is

$$V_{\min} = \frac{\theta}{\sum_{k=1}^{n-1} \dfrac{1}{\theta + k}}. \tag{19}$$

We use simulated samples to evaluate the performance of the BLUE procedure. For a combination of the values of θ and n, we simulated a number of genealogies according to the values of the parameters and coalescent theory (KINGMAN 1982; HUDSON 1982; TAJIMA 1983). For each simulated genealogy, we observe the values of n_i, $i = 1, \ldots, 2(n-1)$ and compute the values of s_{ij}'s. From these quantities, the value of m_i, α_{ij} and β_{ij} are computed respectively from (16), (17) and (18). Then the BLUE procedure is applied to these $2(n-1)$ random variables and the BLUE of parameter θ is obtained. Statistics measuring the performance of the BLUE procedure can thus be calculated. One such statistic is the theoretical variance of the BLUE of θ, which is computed from (10) by substituting θ for $\tilde{\theta}$. Results of these simulations are summarized in Table 2.

It is clear from Table 2 that the iterative nature of the BLUE procedure does not reduce the quality of estimation because $\tilde{\theta}$ is unbiased (or at least has no obvious bias) and its sampling variance is indistinguishable from the theoretical variance. It is encouraging to see that the variance of $\tilde{\theta}$, the BLUE of θ, is only slightly larger than the lower bound of variances (19), confirming our prediction (FU and LI 1993b) that a genealogy contains nearly as much information as that

TABLE 2

Properties of BLUE procedures and comparisons to other two estimators

θ	n	Var($\hat{\pi}$)	Var($\hat{K}$)	$V_{\min}$	BLUE	$\bar{\tilde{\theta}}$	s.v.	$\bar{V}_c$
2	5	2.47	2.27	2.11	2.13	2.01	2.15	2.14
	10	1.93	1.48	1.32	1.33	2.00	1.33	1.33
	20	1.73	1.07	0.93	0.94	2.00	0.95	0.94
	50	1.62	0.77	0.66	0.67	1.98	0.65	0.66
	100	1.59	0.63	0.54	0.54	2.01	0.54	0.55
5	5	11.67	10.60	9.16	9.29	4.98	9.21	9.21
	10	9.01	6.58	5.16	5.25	5.00	5.30	5.24
	20	8.03	4.57	3.35	3.41	5.00	3.37	3.39
	50	7.52	3.14	2.18	2.21	5.00	2.27	2.20
	100	7.37	2.49	1.70	1.72	5.02	1.73	1.72
10	5	41.67	37.60	31.00	31.38	9.99	30.96	31.20
	10	31.98	22.77	16.16	16.44	10.02	16.37	16.45
	20	28.42	15.48	9.68	9.86	10.00	10.25	9.85
	50	26.63	10.33	5.77	5.86	9.97	5.78	5.82
	100	26.08	8.03	4.27	4.32	9.94	4.41	4.28
20	5	156.67	140.80	112.22	113.18	19.96	114.55	113.03
	10	119.75	84.03	54.96	55.74	20.00	56.70	55.76
	20	106.32	56.28	30.50	31.03	20.00	30.85	30.97
	50	99.56	36.86	16.38	16.67	20.02	16.23	16.66
	100	97.51	28.26	11.35	11.52	19.99	12.27	11.50
30	5	345.00	309.60	243.46	245.03	30.04	246.04	245.72
	10	263.33	183.76	116.03	117.37	29.95	122.11	117.42
	20	233.68	122.41	61.96	62.93	29.94	62.91	62.66
	50	218.82	79.58	31.32	31.88	29.96	32.31	31.79
	100	214.30	60.69	20.75	21.09	29.94	24.53	21.04

$\bar{\tilde{\theta}}$, the average of $\tilde{\theta}$ over simulated samples; $V_{\min}$, the lower bound of variances (19); BLUE, theoretical variance of $\tilde{\theta}$ (see text for details); s.v., sampling variance of $\tilde{\theta}$ and $\bar{V}_c$: the average of V_c over simulated samples. The results for each combination of θ and n are from 50,000 simulated samples for n = 5, 10 and 20 and 25,000 simulated samples for n = 50 and 100.

used to derive the lower bound (19). Considering the magnitude of the difference between $V_{\min}$ and Var($\tilde{\theta}$), we are confident that $\tilde{\theta}$ has a minimum variance among all practically possible unbiased estimators of θ, although theoretically it is only a minimum variance estimator among all linear unbiased estimators of θ.

Table 2 also shows that the variance of $\tilde{\theta}$ estimated by equation (9) is appropriate. Since the variance of $\tilde{\theta}$ is very close to the lower bound (19), it suggests that Var($\tilde{\theta}$) can be estimated by

$$V_m = \frac{\tilde{\theta}}{\sum_{k=1}^{n-1} \frac{1}{\tilde{\theta} + k}}. \qquad (20)$$

It appears at first glance that V_m may be an underestimate of Var($\tilde{\theta}$) because it is supposed to estimate $V_{\min}$ which is smaller than Var($\tilde{\theta}$). On the contrary, V_m on average overestimates Var($\tilde{\theta}$) slightly. This is because for a non-negative random variable, r, $E(1/r) > 1/E(r)$, therefore, the denominator in V_m is often an underestimate of its true value which leads to an overestimate of $V_{\min}$. Simulation results which are not presented show that the three estimates of the variance of $\tilde{\theta}$ have the relationship Var($\tilde{\theta}$) $\approx V_c \leq V_{nc} \leq V_m$ in general when sample sizes are small ($n \leq 15$), and have the relationship $V_{nc} \leq$ Var($\tilde{\theta}$) $\approx V_c \leq V_m$ when sample sizes are large, which is a little unexpected because without correction for bias, V_{nc} is likely to be larger than V_c. However, the differences among the three estimates of Var($\tilde{\theta}$) are usually quite small, therefore they all seem to be appropriate for estimating the variance of $\tilde{\theta}$. Overall, V_c appears to be the best estimator while, on the other hand, V_m has the advantage of being the simplest to compute.

A PHYLOGENETIC ESTIMATOR OF θ

Since the genealogy of genes in a sample is usually unknown in reality, one has to estimate the genealogy in order to apply the BLUE procedure for estimating θ. The errors in the reconstructed genealogy are expected to cause bias in the estimation of θ and may also increase the variance of estimation as well. The usefulness of the BLUE procedure depends on whether the bias can be corrected. The degree of bias in the estimation of θ is likely to depend on the method used to construct the genealogy of a sample. Among several methods that can provide all the required information for use with the BLUE procedure, the maximum likelihood method with a molecular clock may be the best choice from a theoretical point of view, but is impractical when the sample size is large or when many genealogies have to be studied. The unweighted pair-group method with arithmetic mean (UPGMA) represents another extreme and it will be used to derive a phylogenetic estimator of θ [see, *e.g.*, NEI (1987) for a detailed description of the method]. The UPGMA is not only the simplest in computation but also an efficient method under the assumption of constant rate of evolution, which is true under the neutral Wright-Fisher model.

Computer simulation is an efficient way to study the properties of estimation of θ from UPGMA trees. Given the values of θ and sample size n, we simulated a large number of samples and for each simulated sample, we computed the number of mutations separating each pair of sequences. That is the number of mutations that occurred on the two sequences since their most recent common ancestor. These numbers formed the distance matrix upon which the UPGMA tree of the sample was constructed. We used the infinite-site model for our simulations so that these numbers were obtained accurately. This is because under the infinite-site model, the number of mutations separating a pair of sequences is simply the number of nucleotide differences between the pair of sequences. The length of branch i of an UPGMA tree is taken as the value of n_i although strictly speaking this is not right because the branch lengths of an UPGMA tree may not be integers since UPGMA gives the expected number of mutations on a branch instead of the realized number of mutations. An obvious alternative is to round each branch length to its nearest integer. This second approach, however, does not perform as well as the first. For each UPGMA tree, the quantities m_i, α_{ij} and β_{ij} are, respectively, computed by (16), (17) and (18). The BLUE procedure is then applied to the $2(n-1)$ variables $m_1, \ldots, m_{2(n-1)}$ and the BLUE of θ, denoted by $\tilde{\theta}_U$, is obtained. Some of the simulation results are shown in Figure 2 where each point in the figure is the mean of $\tilde{\theta}_U$ over at least 2000 simulated samples. Figure 2 shows that $\tilde{\theta}_U$ is on average an underestimate of θ. However the extent of underestimation is largely a function of sample size n with some effects from θ. A regression analysis shows that the following regression equation summarizes remarkably well ($R^2 = 99.9\%$) the relationship between θ, n and mean of $\tilde{\theta}_U$ (see Figure 2)

$$\tilde{\theta}_U = (-0.0336\sqrt{n-2} + 1.002\sqrt{\theta})^2. \qquad (21)$$

This regression equation suggests that one can obtain an unbiased (or nearly unbiased) estimate of θ by the following equation:

$$\hat{\theta} = (0.0335\sqrt{n-2} + 0.998\sqrt{\tilde{\theta}_U})^2. \qquad (22)$$

From the above analysis, I propose a procedure for estimating θ, which will be referred to as UPBLUE in the subsequent discussion, as follows: (1) calculate the number of mutations separating each pair of sequences and form a distance matrix, (2) use UPGMA to construct a genealogy, (3) obtain the estimate $\tilde{\theta}_U$ from the UPGMA tree by BLUE procedure, (4) use Equation 22 to obtain a nearly unbiased estimate, $\hat{\theta}$, of θ and (5) compute the variance of $\hat{\theta}$ by Equation 9, substituting $\tilde{\theta}$ by $\hat{\theta}$.

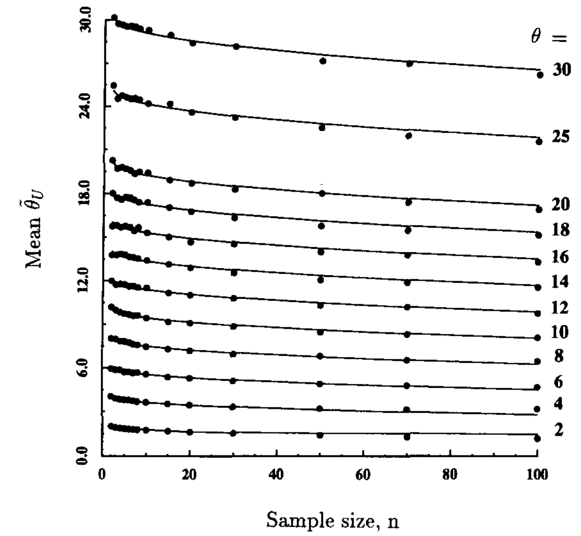

FIGURE 2.—Relationship between θ, sample size n and mean of $\tilde{\theta}_U$. Each dot is the mean of $\tilde{\theta}_U$ over at least 2000 simulations and curves are the regression Equation 21. The number on the right side of each curve is the θ value for simulating the samples upon which the mean $\tilde{\theta}_U$ is based.

The performance of UPBLUE was also investigated by simulations. For a given combination of θ and sample size n, we simulated a large number of samples and UPBLUE was applied to each sample to obtain a value of $\hat{\theta}$. The results which are summarized in Table 3 justify step 5 of UPBLUE for computing the variance of $\hat{\theta}$. Table 3 also shows that $\hat{\theta}$ is a nearly unbiased estimate of θ. The small biases of $\hat{\theta}$ in some combinations of θ and sample size n are unlikely to be significant in practice compared to its variance. What is most encouraging is that the variance of $\hat{\theta}$ is only slightly larger than the lower bound of variances (19) and is almost the same as the variance of $\tilde{\theta}$ obtained assuming a known genealogy (Table 2). This indicates that the correction for bias by Equation 22 is quite effective and UPGMA does not inflate greatly the variance of estimation. It also suggests that V_m and V_{nc} are appropriate estimates of the variance of $\hat{\theta}$. We thus conclude that UPBLUE gives a nearly unbiased estimate of θ with nearly minimum variance. A Fortran program of the UPBLUE procedure is available upon request to the author (E-mail address: fu@gsbs18.gs.uth.tmc.edu). The program takes a distance matrix as its input and outputs the estimate $\hat{\theta}$ and its variance together with other estimates of θ.

TABLE 3

Properties of UPBLUE of θ

θ	n	$\bar{\hat{\theta}}$	s.v.	$\bar{V}_c$	$V_{\min}$
2	5	1.97	2.15	2.09	2.11
	10	1.96	1.36	1.29	1.32
	20	1.95	0.92	0.91	0.93
	50	2.10	0.68	0.71	0.66
	100	2.32	0.54	0.65	0.54
5	5	5.05	9.41	9.42	9.16
	10	4.97	5.48	5.21	5.16
	20	4.92	3.41	3.32	3.35
	50	4.99	2.19	2.20	2.18
	100	5.36	1.71	1.88	1.70
10	5	9.92	32.19	31.15	31.00
	10	9.87	17.14	16.15	16.16
	20	9.85	9.80	9.62	9.68
	50	9.88	5.80	5.75	5.77
	100	10.25	4.29	4.47	4.27
20	5	19.97	115.39	113.26	112.22
	10	20.05	58.62	56.21	54.96
	20	19.82	31.29	30.56	30.50
	50	20.09	16.29	16.76	16.38
	100	20.00	11.52	11.53	11.35
30	5	29.88	244.29	243.32	243.46
	10	30.06	119.12	117.85	116.03
	20	29.79	64.97	62.23	61.96
	50	29.71	31.47	31.35	31.32
	100	29.61	21.34	20.68	20.75

$\bar{\hat{\theta}}$, the average of $\hat{\theta}$ over simulated samples; s.v., sampling variance of $\hat{\theta}$; $\bar{V}_c$, the average of V_c over simulated samples (see step 5 of UPBLUE procedure) and $V_{\min}$, the lower bound of variances (19).

TABLE 4

Comparison of the variances ($\times 10^5$) of estimates of $\theta/360$

	θ				
Estimator	5	10	15	20	25
$\hat{\pi}$[a]	5.75	20.36	43.82	76.14	117.32
$\hat{K}$[b]	2.23	7.30	15.19	25.91	39.47
EWENS[c]	2.30	6.90	13.85	23.40	35.83
UPBLUE[d]	1.54	3.99	7.19	11.08	15.64

[a] Variance calculated by (3).

[b] Variance calculated by (2).

[c] Variance calculated by Equation 15 in CHAKRABORTY and SCHWARTZ (1990).

[d] Variance calculated by (20).

APPLICATION TO THE NUU-CHAH-NULTH DATA

We now apply UPBLUE to a set of mitochondrial sequences. For mitochondrial sequences, θ is defined as $2N_e\mu$ where N_e is the effective female population size and μ is the mutation rate per sequence per generation because mitochondrial sequences are maternally inherited. Sequences of 360 bp of 63 Nuu-Chah-Nulth (Nootka) in the control region of the mitochondrial genome were reported by WARD *et al.* (1991). There are 26 segregating sites in this set of sequences. However, parsimony analyses (WARD *et al.* 1991; LUNDSTROM *et al.* 1992) indicated that many of the segregating sites had experienced multiple mutations because the most parsimonious tree required at least 41 mutations. Therefore, the infinite-site model does not hold for this set of sequences and consequently calculating the number of mutations separating a pair of sequences by the number of nucleotide differences between the pair of sequences will often be an underestimate. To obtain better estimates, we constructed a most parsimonious tree of the sequences. From the parsimony tree the number of mutations separating a pair of sequences was calculated, which was simply the sum of branch lengths of the shortest path between the pair sequences. The resulting distance matrix was then used to construct the UPGMA tree. The value of $\hat{\theta}_U$ from the UPGMA tree is 11.52 and from Equation 22 we obtain $\hat{\theta} = 13.32$ and the variance of $\hat{\theta}$ is estimated to be 7.78. The estimate of θ can be converted into an estimate of nucleotide diversity by $\hat{\theta}/360 = 0.037$ with variance estimated to be 6.0×10^{-5}. In comparison, Watterson's estimate $\hat{K}$ of θ with 41 mutations gives 8.90, $\hat{\pi}$ based on the parsimony tree gives 6.56 and EWENS' estimate based on the number of alleles is 19 (WARD *et al.* 1991). Table 4 shows the variances of these four estimates for several possible values of θ. It is clear that $\hat{\theta}$ has a substantially smaller variance than any of the existing methods. Of course, we should take into account the ambiguity in constructing the distance matrix because the infinite-site model does not hold for this data set. However, all four estimators are affected by the violation of the infinite-site model. The considerable differences among these four estimates of θ may well be a result of the large variances of these estimates but it may also be a consequence that these 63 individuals do not form a truly random sample because they were chosen from 13 of the 14 tribal bands and had to be maternally unrelated for four generations (LUNDSTROM *et al.* 1992). On the other hand, one may argue that our estimate of θ may be unreliable because it was based on a single parsimony tree. In fact, we analyzed one hundred most parsimonious trees and found that $\hat{\theta}$ and Var($\hat{\theta}$) differed little among these trees.

DISCUSSION

The new method of estimating θ, UPBLUE, makes full use of the information from a population sample of sequences, so the resulting estimate is highly efficient. This study shows that the estimated genealogies provide substantially more information for estimating θ than do pairwise differences, the number of segregating sites or the number of haplotypes. The BLUE procedure based on an UPGMA tree is on average an underestimate of the true θ but the bias can be corrected. This method is therefore not very sensitive to minor errors in the reconstruction of a genealogy. This is not surprising because only the number of mutations between successive branching events are needed to achieve the minimum variance (FU and LI 1993a). The present version of the UPBLUE procedure is derived from simulation studies using parameter values in the range $\theta \leqslant 30$ and $n \leqslant 100$. The pattern of $\tilde{\theta}_U$ in Figure 2 suggests that it should be applicable to a wider range of parameter values unless θ is considerably larger than 30 or n is substantially larger than 100. Nevertheless, the procedure can be extended to cover wider range of parameter values when it is necessary.

An important advantage of the BLUE procedure is that it can be applied to partial information in a genealogy. For example, one can obtain BLUE of θ based on only the internal branches or only the external branches of a genealogy. Therefore, the BLUE procedure may be very useful for testing hypotheses about θ. Because of the smaller variance of BLUE of θ, statistical tests of the hypothesis of neutrality of mutations based on BLUE of θ for different parts of a genealogy are likely to be more powerful than those proposed by TAJIMA (1989) and FU and LI (1993b).

The nearly minimum variance of $\hat{\theta}$ suggests that further efforts to improve the estimate of θ will not be fruitful under the neutral infinite-site Wright-Fisher model without recombination and population subdivision. However, any real sample of DNA sequences is likely to violate to certain degree the infinite-site neutral Wright-Fisher model. Therefore it is not necessary that UPGMA is the best treeing method to use with the BLUE procedure. Other more robust distance methods, such as neighbor-joining or minimum evolution, may be better choices when the neutral infinite-site Wright-Fisher model is slightly violated. Investigations on the performances of other treeing methods, including Maximum likelihood method, with the BLUE procedure will be useful. It is clear that the keys for constructing a good phylogenetic estimator of θ are to accurately construct the genealogy of a sample and to know the distributions of the times between consecutive branching events in the genealogy. When the neutral Wright-Fisher model without recombination and population subdivision is severely violated, phylogenetic estimators may be difficult to construct. For example, when recombinations are frequent, a reconstructed genealogy by any existing treeing method may be grossly misleading (HUDSON and KAPLAN 1985), estimation of θ based on such an erroneous genealogy is unlikely to be much better, for example, than WATTERSON's estimator. However, the BLUE procedure developed in this paper does not have to be associated with a genealogy. In a subsequent paper, I shall consider a different use of the BLUE procedure for estimating θ which is applicable under the neutral Wright-Fisher model with or without recombination and population subdivision.

I thank B. GOLDING, W. H. LI and two anonymous reviewers for suggestions for improving the manuscript.

LITERATURE CITED

AVISE, J. C, R. M. BALL and J. ARNOLD, 1988 Current versus historical population sizes in vertebrate species with high gene flow: A comparison based on mitochondrial DNA lineages and inbreeding theory for neutral mutations. Mol. Biol. Evol. **5:** 331–344.

CHAKRABORTY, R., and R. J. SCHWARTZ, 1990 Selective neutrality of surname distribution in an immigrant indian community of Houston, Texas. Am. J. Hum. Biol. **2:** 1–15.

EWENS, W. J., 1979 *Mathematical Population Genetics*. Springer-Verlag, New York.

FELSENSTEIN, J., 1992a Estimating effective population size from samples of sequences: inefficiency of pairwise and segregation sites as compared to phylogenetic estimates. Genet. Res. **56:** 139–147.

FELSENSTEIN, J., 1992b Estimating effective population size from samples of sequences: a bootstrap monte carlo integration method. Genet. Res. **60:** 209–220.

FU, Y. X., and W. H. LI, 1993a Maximum likelihood estimation of population parameters. Genetics **134:** 1261–1270.

FU, Y. X., and W. H. LI, 1993b Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

HUDSON, R., 1982 Testing the constant-rate neutral allele model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R., and N. L. KAPLAN, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genet. **111:** 147–164.

KINGMAN, J., 1982 On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

LUNDSTROM, R., S. TAVARÉ and R. H. WARD, 1992 Estimating substitution rates from molecular data using the coalescent. Proc. Natl. Acad. Sci. USA **89:** 5961–5965.

NEI, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

SEARLE, S. R., 1982 *Matrix Algebra Useful for Statistics*. John Wiley & Sons, New York.

STROBECK, C., 1983 Estimation of the neutral mutation rate in a finite population from DNA sequence data. Theor. Popul. Biol. **24:** 160–172.

TAJIMA, F., 1983 Evolutionary relationship of DNA sequences in finite populations. Genetics **105:** 437–460.

TAJIMA, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

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., 1975 On the number of segregation sites. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: G. B. GOLDING