Estimating the Age of the Common Ancestor of a Sample of DNA Sequences

Yun-Xin Fu and Wen-Hsiung Li

Human Genetics Center, University of Texas at Houston

We present a simple Monte Carlo method for estimating the age of the most recent common ancestor (MRCA) of a sample of DNA sequences. We show that Templeton's (1993) estimator of the age of the MRCA based on the maximum number of nucleotide differences between two sequences in a sample is inaccurate, and we demonstrate the new method by reanalyzing a sample of DNA sequences from human Y chromosomes and a sample of human *Alu* sequences.

Introduction

The age of the most recent common ancestor (MRCA) of modern humans is of great interest in the study of human evolution. The availability of human mitochondrial DNA sequences allowed the estimation of the age of the common ancestor of maternal lineages, while the recent DNA samples from human Y chromosomes provide information about the evolutionary history of humans. Essential to the inference of the common ancestry of humans are proper estimators of the age of the MRCA of a sample. Templeton (1993) proposed an estimator based on the maximum number (k_{max}) of nucleotide differences between two sequences in the sample. Templeton's method has been used to estimate the age of the MRCA of mitochondria (Ruvolo et al. 1993; Ruvolo 1996) and that of the Y chromosomes (Hammer 1995). In this report, we will present a simple and flexible Monte Carlo method for obtaining point and interval estimates of the age of the MRCA of a sample and demonstrate that Templeton's method is inaccurate. In our approach, one can consider the number of segregating sites in the sample as well as k_{max}. We shall use both methods to reanalyze the sample of Y chromosome sequences obtained by Hammer (1995) and show that better estimates can be obtained. We shall also reanalyze the *Alu* sequence data by Knight et al. (1996), which have not been properly analyzed.

Templeton's Estimate

Let T be the age (in generations) of the MRCA of a sample. Tajima (1983) showed that when the number (n) of DNA sequences in a sample is 2, the mean and variance of T conditional on the number (π) of nucleotide differences between the two sequences are

$$E(T|\pi) = \frac{\theta(1 + \pi)}{2\mu(1 + \theta)} \quad (1)$$

and

$$V(T|\pi) = \frac{\theta^2(1 + \pi)}{4\mu^2(1 + \theta)^2}, \quad (2)$$

where $\theta = 2N\mu$, N is the effective population size, and μ is the mutation rate per sequence per generation. Define $T' = T/(2N)$, so that one unit corresponds to $2N$ generations. We then have the following simpler formulas

$$E(T'|\pi) = \frac{1 + \pi}{2(1 + \theta)} \quad (3)$$

and

$$V(T'|\pi) = \frac{1 + \pi}{4(1 + \theta)^2}. \quad (4)$$

$E(T|\pi)$ and $E(T'|\pi)$ can thus be used to estimate T and T', respectively, for a sample of two sequences. For $n > 2$, Templeton (1993) proposed to estimate T by replacing π in equation (1) with k_{max}, the maximum number of nucleotide differences between two sequences in the sample, i.e., the number of nucleotide differences between the two most divergent sequences in the sample. He also suggested using a gamma distribution to obtain the confidence interval of T.

As the derivation was heuristic, Templeton's method is inaccurate when $n > 2$. This can be shown analytically for the case $k_{max} = 0$ (i.e., no nucleotide variation in the sample) by using the results of Fu and Li (1996) or Donnelly et al. (1996). For example, from Donnelly et al. (1996), we have

$$E(T|k_{max} = 0) = (2N)\sum_{i=2}^{n} \frac{1}{i(i - 1 + \theta)} \quad (5)$$

$$V(T|k_{max} = 0) = (2N)^2\sum_{i=2}^{n} \frac{1}{i^2(i - 1 + \theta)^2}. \quad (6)$$

These two formulas are different from equations (1) and (2) except for the case of $n = 2$. The Monte Carlo method described in the next section enables us to examine other situations and to show numerically that Templeton's method is inaccurate.

A Monte Carlo Approach

Essential to the estimation of T from k_{max} is the probability $p(T|k_{max})$ of T conditional on the value of k_{max}, because both $E(T|k_{max})$ and $V(T|k_{max})$ can be computed from this probability density. Although it is not easy to derive $p(T|k_{max})$ analytically for a sample of more than two sequences, an adequate estimate of this

Key words: common ancestor, age estimation, coalescent approach, Y chromosome, *Alu* sequence.

Address for correspondence and reprints: Dr. Yun-Xin Fu, Human Genetics Center, School of Public Health, University of Texas at Houston, P.O. Box 20334, 6901 Bertner Avenue, Houston, Texas 77225. E-mail: fu@hgc.sph.uth.tmc.edu.

Mol. Biol. Evol. 14(2):195–199. 1997

probability can be obtained by a simple Monte Carlo approach as follows.

Suppose we observe in a given sample of size n that $k_{max} = k_m$. To estimate the age T of the MRCA of the sample, we start by choosing a population genetics model for the population from which the sample was taken. For the model, a fast algorithm, such as a coalescent algorithm, for simulating samples from the model should be available. For example, one may choose the Wright-Fisher model with constant population size, or a model with expanding population size. In addition to the population genetics model, one may assume the infinite-site model for mutations or use a finite-sites model. Once the model and the associated parameters (e.g., θ) are given, we can simulate samples of size n. For each simulated sample, we obtain a value of T and a value of k_{max}. The range of T is divided into intervals $I_j = (j\delta, (j + 1)\delta]$, $(j = 0,\ldots)$ and $\delta > 0$. Suppose that s samples are simulated, among which there are $\#(k_m)$ samples with $k_{max} = k_m$ and there are $\#(I_j, k_m)$ samples with the age of the MRCA falling into the interval I_j and $k_{max} = k_m$. Note that

$$p(T \in I_j | k_{max}) = \frac{p(T \in I_j, k_{max})}{p(k_{max})}, \quad (7)$$

where $p(T \in I_j, k_{max})$ is the joint probability of $T \in I_j$ and k_{max}, and $p(k_{max})$ is the probability of k_{max}. The former can be estimated by $\#(I_j, k_{max})/s$ and the latter by $\#(k_{max})/s$. We can thus estimate the probability $p(T \in I_j | k_{max} = k_m)$ by

$$\hat{p}(T \in I_j | k_{max} = k_m) = \frac{\#(I_j, k_m)}{\#(k_m)}. \quad (8)$$

Because $\lim_{s\to\infty} \hat{p}\,(T \in I_j | k_{max}) = p(T \in I_j | k_{max})$, adequate accuracy can be achieved by choosing a sufficiently large value for s. In general, the smaller the probability $p(k_{max})$ and the smaller the value of δ are, the larger the value of s should be. Since coalescent algorithms are fast, simulating a large number of samples can be carried out in a reasonable amount of computer time. After comparing the effects of different values of δ and s on the estimates of T in our simulations, we found that there is usually no need to choose δ smaller than 10^{-3} and that $s = 100{,}000$ is usually sufficient. Our approach, although more general, is similar to that of Weiss and von Haeseler (1996), who considered the special case of $k_{max} = 0$.

Figure 1 shows the conditional mean and variance of T for $\theta = 1$ and 3, and for sample sizes 2, 16, and 50 under the neutral Wright-Fisher model with constant population size. It is clear that the conditional mean and variance in the case of $n = 2$ agree with equations (1) and (2) as they should, whereas for $n = 16$ and $n = 50$, they are different from equations (1) and (2). It should be noted that T' is in units of $2N$ generations, so that a deviation may appear to be small. For both values of θ, the mean given by equation (1) is an overestimate except when k_{max} is very close to zero and the variance given by equation (2) is almost always an overestimate of the true variance. These results suggest that most of the time Templeton's estimate will be an overestimate with a wider confidence interval.

In addition to the mean estimate $T_{mean} = E(T|k_{max})$ of T, the Monte Carlo approach allows us to obtain another point estimate T_{mode} of T, which is the value of T that maximizes the posterior probability $p(T|k_{max})$, and an interval estimate of T (for example, see Fu 1996). Furthermore, the Monte Carlo method can also be used to obtain estimates of T from other quantities, such as the number (K) of segregating sites in a sample; Fu (1996) has obtained an analytical solution of $p(T|K)$ under the neutral Wright-Fisher model, but it is computationally less convenient.

The virtue of the Monte Carlo approach presented here is its simplicity and flexibility. It is simple because there is no need to develop an analytical formula for the quantity under study; it is flexible because it can be used under the simple Wright-Fisher model or a more complicated population genetics model, such as population growth, population subdivision, or natural selection, as long as a fast simulation algorithm is available. The disadvantage is that only the simulated samples with a specific k_{max} value are used in the inference; therefore its use of computer resources is less efficient than the algorithm of Griffiths and Tavaré (1994). However, since coalescent algorithms are usually fast, this Monte Carlo approach is an inexpensive and convenient method for making inference about the age of the MRCA of a sample.

The Age of the MRCA of Human Y Chromosomes

We now use a sample of Y sequences to illustrate the method described above. Hammer (1995) examined a sample of 16 Y sequences of 2,654 bp and found $k_{max} = 3$ and $K = 3$. Hammer estimated the mutation rate to be 1.9×10^{-9} per site per year. Taking 20 years as one human generation yields $\mu = 1.9 \times 10^{-9} \times 2{,}654 \times 20 = 1.0 \times 10^{-4}$. Using Templeton's method and taking $N = 4{,}900$, Hammer (1995) obtained 188,000 years as the estimate of the age of the MRCA of the sample, and the 95% confidence interval of the age was from 51,000 to 411,000 years.

Table 1 shows the estimates of T using either k_{max} or K for a number of effective population sizes N. We present the estimates based on both k_{max} and K because currently it is not clear which quantity is better in general. For comparison, we also include the corresponding estimates from K using the analytical distribution (Fu 1996). It is clear that the estimates based on K from the Monte Carlo method are excellent approximations to the exact results, suggesting that the Monte Carlo method is reliable. Figure 2 shows the distributions of $p(T|K = 3)$.

Takahata (1993) estimated that the long-term effective population size of the human population is about 10,000. Assuming an equal sex ratio, the effective male population size is about 5,000. For $N = 5{,}000$, we find from table 1 that $T_{mode} = 134{,}000$ and $T_{mean} = 183{,}000$ and the 95% interval estimate of T is from 66,000 to 390,000 years when the estimation is based on k_{max}, and

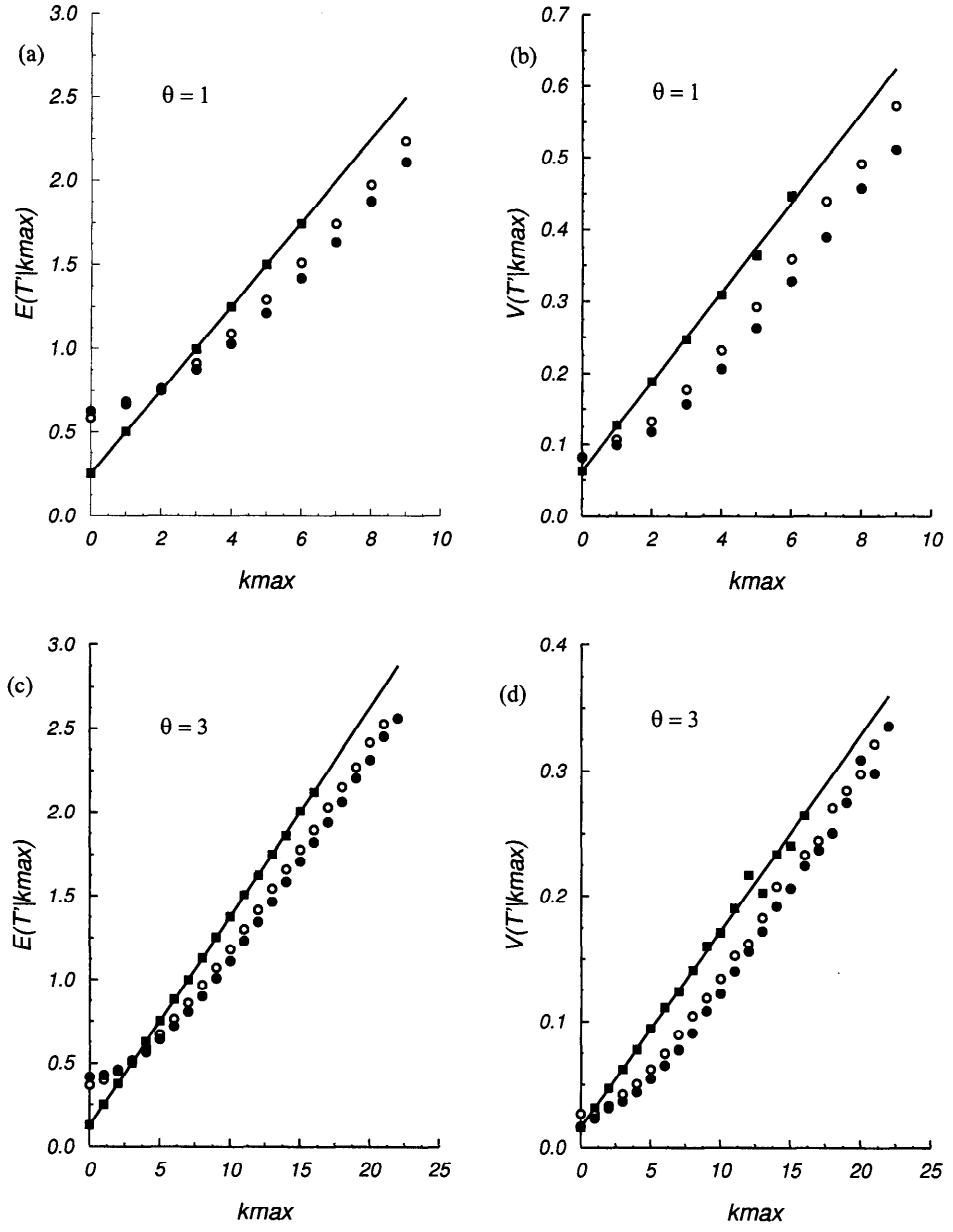

FIG. 1.—The values of $E(T'|k_{max})$ and $V(T'|k_{max})$ for k_{max}'s up to $k99$, where $k99$ is the maximum value so that $p(k_{max} \leq k99) \leq 0.99$. In each panel, solid squares, open circles, and solid circles represent, respectively, the simulation results for sample sizes 2, 16, and 50, and straight lines are Templeton's mean (panels *a* and *c*) and variance (panels *b* and *d*). Five hundred thousand samples were simulated for each value of θ. The times are in units of $2N$ generations.

T_{mode} = 126,000, T_{mean} = 174,000, and the 95% interval estimate of T is from 63,000 to 381,000 years when the estimation is based on K. These two sets of estimates are close to each other. The mean estimates are slightly smaller than Hammer's; a smaller T_{mean} is expected since we can see from Fig. 1*a* that when k_{max} = 3, Templeton's estimate is a slight overestimate. The major difference between our estimates and Hammer's is the 95% confidence interval estimate of T. Our interval estimate based on k_{max} is 38,000 years narrower than Hammer's (39,000 years narrower if we assume N = 4,900 as Hammer did), and the interval estimate based on K is 42,000 years narrower than Hammer's (45,000 years narrower if we assume N = 4,900). All these interval estimates are also better than our previous estimate (Fu and Li 1996) from Dorit, Akashi, and Gilbert's (1995) data, which has no segregating site. This suggests that Hammer's sample is more informative than Dorit, Akashi, and Gilbert's sample.

Between the two point estimates, T_{mode} is preferred over T_{mean} because the former is the most likely value of T. So the age of the MRCA of human Y chromosomes may be as young as about 125,000 years, which is similar to our previous estimate (Fu and Li 1996) from Dorit, Akashi, and Gilbert's data.

In the above analysis, we have relied much on the single estimate of N = 5,000 to draw our main conclusions about the age of the common ancestor of the hu-

Table 1
Estimates of T Based on k_{max} and K for the Data of Hammer (1995) Based the Wright-Fisher Model with Constant Population Size

N	T_{mode}	T_{mean}	T_{95}	95% Interval Estimate
		From k_{max}		
2,500	83	118	229	40 to 261
4,000	118	160	305	58 to 348
4,900	126	180	339	66 to 385
5,000	128	183	344	68 to 390
7,500	170	226	415	88 to 467
10,000	205	259	466	106 to 526
		From K		
2,500	76	112	221	37 to 252
4,000	115	152	295	53 to 337
4,900	122	172	330	62 to 377
5,000	126	174	334	63 to 381
7,500	178	216	403	82 to 458
10,000	181	247	454	97 to 516
		Exact results from K		
2,500	77	112	221	38 to 254
4,000	107	153	296	54 to 338
4,900	123	172	330	62 to 377
5,000	124	174	333	63 to 381
7,500	159	215	404	82 to 460
10,000	186	246	453	97 to 515

NOTE.—T_{95} is the maximum value of T such that $p(T \leq T_{95}|k_{max} = 3) \leq 0.95$ (when based on k_{max}) or $p(T \leq T_{95}|K = 3) \leq 0.95$ (when based on K). The exact results are based on the algorithm by Fu (1996). Two hundred thousand samples were simulated for each N and all estimates are in units of 1,000 years and rounded to the nearest thousand.

man Y chromosomes. One should allow some uncertainty about N (and μ as well) and so we include other N's in table 1. However, although N = 5,000 may be questionable, we probably can assume $N \leq$ 10,000 in light of Takahata's (1993) analysis and can conclude from the results based on K in table 1 that the most likely value of the age of the common ancestor of human Y chromosomes is less than 186,000 years and with 95% probability that it is less than 453,000 years.

The Age of the MRCA of a Sample of Autosomal *Alu* Sequences

Knight et al. (1996) studied three loci of *Alu* elements in autosomal regions. The α-globin 2 *Alu* locus

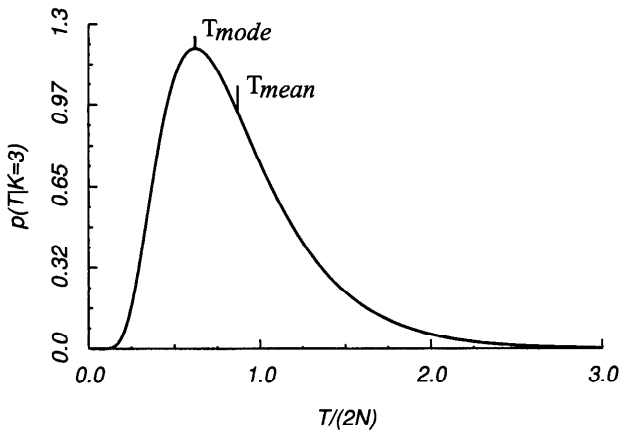

FIG. 2.—The posterior distribution $p(T|K)$ for Hammer's (1995) data.

Table 2
Estimates of T from Knight et al.'s (1996) Sample Based the Wright-Fisher Model with Constant Population Size

N	T_{mode}	T_{mean}	T_{95}	95% Interval Estimate
		$\mu = 0.438 \times 10^{-4}$		
5,000	224	308	586	116 to 674
7,500	294	396	732	156 to 837
10,000	360	466	844	188 to 960
12,500	410	525	930	220 to 1,055
15,000	450	574	1002	246 to 1,134
		$\mu = 0.294 \times 10^{-4}$		
5,000	240	348	680	126 to 784
7,500	336	460	876	174 to 1,005
10,000	408	552	1028	216 to 1,176
12,500	475	630	1155	250 to 1,315
15,000	540	698	1260	282 to 1,434

NOTE.—All estimates are in units of 1,000 years and rounded to the nearest thousand.

was analyzed in detail. They found only two mutations in a sample of 120 *Alu* sequences, each of length 300 bp. They estimated that the mutation rate per site per year is 0.73×10^{-8} or 0.49×10^{-8} if the divergent time between Human and Bonobo is 4.7 Myr or 7 Myr. With 20 years for one human generation, the two estimates correspond, respectively, to $\mu = 0.73 \times 10^{-8} \times 300 \times 20 = 0.438 \times 10^{-4}$ and $\mu = 0.294 \times 10^{-4}$ per sequence per generation. Knight et al. (1996) estimated that the average times of sequence divergence are 37,000 and 55,000 years, respectively, with the two mutation rate estimates, so that their data indicate a recent replacement of the human autosomal genetic complement.

Knight et al., however, failed to consider the population dynamics of genes, especially random drift. In fact, their estimates are independent of N, but the coalescent times of sequences in a population should depend on N. We can estimate the age of the MRCA of these *Alu* sequences with our Monte Carlo method. Note that when there are only two mutations in a sample, both k_{max} and K are equal to 2, so that estimates of T based on either K or k_{max} are the same. Table 2 gives the estimates T_{mode}, T_{mean} and the 95% interval estimates for the two mutations rates.

Assuming N = 10,000, we can see from table 2 that T_{mode} = 360,000, T_{mean} = 466,000 years, and the 95% confidence interval estimate of T is from 188,000 to 960,000 years if $\mu = 0.438 \times 10^{-4}$ or T_{mode} = 408,000, T_{mean} = 552,000 years and the 95% interval estimate of T is from 216,000 to 1,176,000 years if $\mu = 0.294 \times 10^{-4}$. These estimates are much older than those of Knight et al. (1996) and do not necessarily support the hypothesis of the recent origin of modern humans, particularly if the uncertainty in the value of N is taken into account.

LITERATURE CITED

DONNELLY, P., S. TAVARÉ, D. J. BALDING, and R. C. GRIFFITHS. 1996. Estimating the age of the common ancestor of men from the ZFY intron. Science **272**:1357–1359.

DORIT, R. L., H. AKASHI, and W. GILBERT. 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. Science **268**:1183–1185.

FU, Y. X. 1996. Estimating the age of the common ancestor of a DNA sample using the number of segregating sites. Genetics **144**:829–838.

FU, Y. X., and W. H. LI. 1996. Estimating the age of the common ancestor of men from the ZFY intron. Science **272**: 1356–1357.

GRIFFITHS, R. C., and S. TAVARÉ. 1994. Ancestral inference in population genetics. Stat. Sci. **9**:307–319.

HAMMER, M. F. 1995. A recent common ancestry for human Y chromosomes. Nature **378**:376–378.

KNIGHT, A., M. A. BATZER, M. STONEKING, H. K. TIWARI, W. D. SHEER, R. J. HERRERA, and P. L. DEININGER. 1996. DNA sequences of *alu* elements indicate a recent replacement of the human autosomal genetic complement. Proc. Natl. Acad. Sci. USA **93**:4360–4364.

RUVOLO, M. 1996. A new approach to studying modern human origins: hypothesis testing with coalescent time distributions. Mol. Phylogenet. Evol. **5**:202–219.

RUVOLO, M., S. ZEHR, M. VON DORMUM, D. PAN, B. CHANG, and J. LIN. 1993. Mitochondrial CoII sequences and modern human origins. Mol. Biol. Evol. **10**:1115–1135.

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

TAKAHATA, N. 1993. Allelic genealogy and human evolution. Mol. Biol. Evol. **10**:2–22.

TEMPLETON, A. R. 1993. The Eve hypotheses: a genetic critique and reanalysis. Am. Anthropol. **95**:51–72.

WEISS, G., and A. VON HAESELER. 1996. Estimating the age of the common ancestor of men from the ZFY intron. Science **272**:1359–1360.

DAN GRAUR, reviewing editor

Accepted October 29, 1996