Estimating the Age of the Common Ancestor of a DNA Sample Using the Number of Segregating Sites

Yun-Xin Fu

Human Genetics Center, University of Texas, Houston, Texas 77030

Manuscript received March 8, 1996
Accepted for publication July 12, 1996

ABSTRACT

The number of segregating sites in a sample of DNA sequences and the age of the most recent common ancestor (MRCA) of the sequences in the sample are positively correlated. The value of the former can be used to estimate the value of the latter. Using the coalescent approach, we derive in this paper the joint probability distribution of the number of segregating sites and the age of the MRCA of a sample under the neutral Wright-Fisher model. From this distribution, we are able to compute the likelihood function of the number of segregating sites and the posterior probability of the age of the MRCA of a sample. Three point estimators and one interval estimator of the age of the MRCA are developed; their relationships and properties are investigated. The estimation of the age of the MRCA of human *Y* chromosomes from a sample of no variation is discussed.

THERE are considerable interests in the age of the most recent common ancestor (MRCA) of a DNA sample when studying the evolutionary history of a population from which the sample is taken. The current controversy on the age of the MRCA of modern humans attests the need of proper statistical methods for the inferences on common ancestry. Because an inference on the age of the MRCA has to be based on population samples, appropriate population genetics theory should be taken into account. The coalescent theory (KINGMAN 1982a,b; HUDSON 1983; TAJIMA 1983) is a natural choice because it deals with how the sequences in a sample coalesce to their common ancestors. In this paper, we shall present a coalescent theory that is necessary for the estimation of the age of the MRCA of a sample using the number of segregating sites and investigate the properties of newly developed estimators from this theory.

The number of segregating sites in a sample of DNA sequences from a population is the simplest quantity observable. Since WATTERSON's (1975) work, the number of segregating sites has been widely used for estimating the essential population parameter $\theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per sequence per generation, and recently has been used for testing evolutionary hypotheses (*e.g.*, TAJIMA 1989; FU and LI 1993b; FU 1996). Although samples of DNA sequences have been used by several authors to estimate the age of human mitochondria, the sample by DORIT *et al.* (1995), which consists of 38 sequences from the intron of *ZFY* gene in the human *Y* chromosome, presented a special challenge because there is no variation in this sample. Any estimator of the age of the MRCA that is proportional to the number of segregating sites or the mean number of nucleotide differences between two sequences will yield zero as the estimate, which is apparently unacceptable. DORIT *et al.* (1995) attempted to estimate the age of the MRCA of the human males from this sample, but their analysis was not rigorous (FU and LI 1996; DONNELLY *et al.* 1996; WEISS and VON HAESELER 1996). FU and LI (1996) developed a method from the coalescent theory to deal with samples with no variation and reanalyzed DORIT *et al.*'s sample. Their theory is extended in this paper to cope with samples with any number of segregating sites.

THE THEORY

We assume that the population under study evolves according to the Wright-Fisher model, that mutations in the locus from which DNA sequences are obtained are selectively neutral, that the effective population size is constant over time and that there is no recombination within the locus. We shall present our results for a sample of DNA sequences from an autosomal locus so that the parameter θ is defined as $4N\mu$, where N is the effective population size and μ is the mutation rate per sequence per generation. Our results also apply to a DNA sample from a mitochondrial locus by defining θ as $2N_f\mu$, where N_f is the effective size of the female population, and to a DNA sample from a locus in Y chromosome by defining θ as $2N_m\mu$, where N_m is the effective size of the male population.

The genealogy of a sample of n DNA sequences can be divided into $n - 1$ states numbered from 2 to n. State k is the period in which the genealogy has exactly k ancestral sequences (Figure 1). The time length t_k of state k (in number of generations) is the so-called kth

Address for correspondence: Human Genetics Center, University of Texas at Houston, 6901 Bertner Ave., S222, Houston, TX 77030. E-mail: fu@hgc.sph.uth.tmc.edu

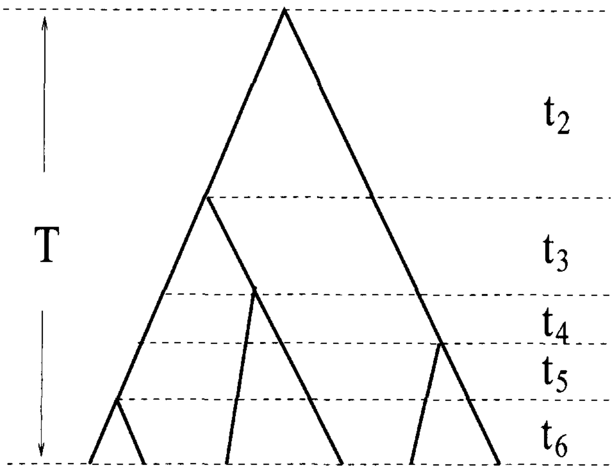

FIGURE 1.—An example of the genealogy of a sample of six sequences. $T = t_2 + \cdots + t_6$ and the total time length of the genealogy is $L = 2t_2 + \cdots + 6t_6$. Dashed lines divide T into five periods (states).

coalescent time. When sample is random, t_k follows approximately an exponential distribution with parameter $k(k-1)/(4N)$ (KINGMAN 1982b). The age (T) of the MRCA of the sample is equal to

$$T = t_2 + \cdots + t_n$$

and the time length in the entire genealogy is $L = 2t_2 + \cdots + nt_n$. The sample genealogy consists of $2(n-1)$ branches. Assume that the number of mutations in branch i conditional on the length l_i of the branch follows Poisson distribution with parameter $l_i\mu$. Then the number K of mutations in the entire genealogy conditional on the coalescent times t_k ($k = 2, \ldots, n$) is the sum of $2(n-1)$ Poisson variables and thus follows the Poisson distribution:

$$P(K|t_2, \ldots, t_n) = \frac{e^{-\mu L}}{K!}(\mu L)^K. \tag{1}$$

When the infinite-sites model is assumed, K is the number of segregating sites in the sample. Since different coalescent times are independent, the joint probability density of $t_2, \ldots, t_n$ is thus

$$\prod_{k=2}^{n} \frac{k(k-1)}{4N} \exp\left[-\frac{k(k-1)}{4N} t_k\right]. \tag{2}$$

The joint probability that there are K segregating sites in the sample and that the kth coalescent time ($k = 2, \ldots, n$) is equal to t_k is the product of (1) and (2), namely

$$\frac{e^{-\mu L}}{K!}(\mu L)^K \prod \frac{k(k-1)}{4N} \exp\left[-\frac{k(k-1)}{4N} t_k\right].$$

If coalescent times are rescaled such that one unit corresponds to $4N$ generations, the above equation becomes

$$\frac{\theta^K n!(n-1)!}{K!} L^K \prod_k e^{-k(\theta+k-1)t_k} \tag{3}$$

Throughout this paper, all times are so scaled when their units are not specified. Note that one unit of the scaled time will correspond to $2N_f$ generations if the locus is in mitochondria and $2N_m$ generations if the locus is in the nonrecombining region of Y chromosome.

It follows from (3) that the probability of the event that there are K segregating sites and that the age of the MRCA of the sample is T is

$$p_n(K, T) = \frac{\theta^K n!(n-1)!}{K!} \int \cdots \int_{t_2+\cdots+t_n=T} L^K \times \prod_k e^{-k(\theta+k-1)t_k} dt_n \cdots dt_2 \tag{4}$$

This joint probability is the foundation for the inferences on T from K. We can show (APPENDIX) that

$$p_n(K, T) = \frac{\theta^K n!(n-1)!}{K!} \sum_{k=2}^{n} \sum_{l=0}^{K} \alpha_{kl} k^l T^l e^{-k(\theta+k-1)T} \tag{5}$$

where

$$\alpha_{kl} = \frac{K!}{l!} \beta_k(\theta) \gamma_{K-l,k} \tag{6}$$

$$\beta_k(\theta) = \frac{(-1)^k(\theta + 2k - 1)}{(k-2)!(n-k)!\prod_{i=1}^{n-1}(\theta + k + i)} \tag{7}$$

and

$$\gamma_{K-l,k} = \sum_{j_2+\cdots+j_n=K-l;j_k=0} \prod_m \frac{1}{(\theta + k + m - 1)^{j_m}}. \tag{8}$$

It is clear that there is only one term in the summation of $\gamma_{K-1,k}$ for $l = K$; while the number of terms for $l < K$ can be shown to be $\sum_{i=1}^{\min(K-l,n-2)} \binom{K-l-1}{i-1}\binom{n-2}{i}$, which is in the order of n^{K-l}. Therefore, it is not convenient to compute a_{kl} directly from (6) when $K - l$ and n are not small. Letting $\alpha_{kl} = \alpha_{kl}(2)$, it can be shown (APPENDIX) that $\alpha_{kl}(2)$ can be computed from the following iteration procedure:

$$\alpha_{kl}(i) = \sum_{j=l}^{K} \frac{\alpha_{kj}(i+1)}{(i-k)(\theta + i + k - 1)^{j-l}} \frac{j!}{l!}, \quad k = i + 1, \ldots, n$$

$$\alpha_{il}(i) = -\sum_{k=i+1}^{n} \alpha_{kl}(i) \tag{9}$$

for $l = 0, \ldots, K$. The initial values for the iteration are

$$\alpha_{nK}(n) = 1, \alpha_{n0}(n) = \cdots = \alpha_{nK-1}(n) = 0. \tag{10}$$

Two marginal distributions $p_n(K)$ and $\phi_n(T)$ can be obtained from $p_n(K, T)$; the former is the distribution of K and the latter the distribution of T. It is simple to show that

$$p_n(K) = \int_0^\infty p_n(K, t)\,dt$$

$$= \frac{\theta^K n!(n-1)!}{K!}\,c_K \qquad (11)$$

where

$$c_K = \sum_{k=2}^{n} \sum_{l=0}^{K} \alpha_{kl} \frac{l!}{k(\theta + k - 1)^{l+1}}\,. \qquad (12)$$

Equation (11) provides an alternative way to compute the probability of K than the formula derived by TAVARÉ (1984). $\phi_n(T)$ can be obtained by summing $p_n(K, T)$ over all possible values of K. Because $\phi_n(T)$ is independent of the mutation rate μ, by setting $\mu = 0$ (thus $\theta = 0$) we have from (5) that

$$\phi_n(T) = n!(n-1)!$$

$$\times \sum_{k=2}^{n} \frac{(-1)^k(2k-1)k(k-1)}{(n-k)!(n+k-1)!}\,e^{-k(k-1)T}$$

$$= \sum_{k=2}^{n} (-1)^k(2k-1)$$

$$\times \left(\prod_{i=1}^{k-1} \frac{n-i}{n+i}\right) k(k-1)\,e^{-k(k-1)T}. \qquad (13)$$

This equation is equivalent to TAJIMA's (1989) Equation 3, except that different time scales are used and that TAJIMA (1990) considered only the case $n = 2N$. Incidentally, since the exponential distribution of a coalescent time is derived under the assumption that $n \ll 2N$, Equation 13 should be applicable only to samples of sizes that are much smaller than $2N$. Nevertheless, TAJIMA (1990) showed that $\phi_{2N}(T)$ is close to KIMURA's (1970) distribution of fixation time of a new neutral mutant.

From the joint probability density $p_n(K, T)$ and the two marginal probabilities $p_n(K)$ and $\phi_n(T)$, two quantities that are essential for the inferences on T can be computed. One is the likelihood function $p_n(K|T)$ of T and the other is the posterior probability $p_n(T|K)$ of T, defined respectively as

$$p_n(K|T) = \frac{p_n(K, T)}{\phi_n(T)}\,, \qquad (14)$$

$$p_n(T|K) = \frac{p_n(K, T)}{p_n(K)}\,. \qquad (15)$$

The posterior probability is equal to

$$p_n(T|K) = c_K^{-1} \sum_{k=2}^{n} \sum_{l=0}^{K} \alpha_{kl} k^l T^l e^{-k(\theta+k-1)T} \qquad (16)$$

from which one can derive the conditional expectation and variance of T. It is a simple matter to show that

$$E(T|K) = c_K^{-1} \sum_{k=2}^{n} \sum_{l=0}^{K} \alpha_{kl} \frac{(l+1)!}{k^2(\theta + k - 1)^{l+2}} \qquad (17)$$

$$\mathrm{Var}\,(T|K) = c_K^{-1} \sum_{k=2}^{n} \sum_{l=0}^{K} \alpha_{kl} \frac{(l+2)!}{k^3(\theta + k - 1)^{l+3}}$$

$$- E^2(T|K). \qquad (18)$$

We now consider several situations in which Equation 5 is convenient to use directly. The first case is when $K = 0$. It is easy to see from (8) that $\gamma_{0,k} = 1$. Therefore $\alpha_{kl} = \beta_k(\theta)$, which implies that

$$p_n(0, T) = n!(n-1)! \sum_{k=2}^{n} \beta_k(\theta)\,e^{-k(\theta+k-1)T}. \qquad (19)$$

Since WATTERSON (1975) showed that

$$p_n(K = 0) = \prod_{k=1}^{n-1} \frac{k}{\theta + k} \qquad (20)$$

the posterior probability $p_n(T|0)$ becomes

$$p_n(T|0)$$

$$= n!\left[\prod_{k=1}^{n-1} (\theta + k)\right] \sum_{k=2}^{n} \beta_k(\theta)\,e^{-k(\theta+k-1)T} \qquad (21)$$

which was derived first by FU and LI (1996). Substituting β_k for α_{kl} in (17) and (18) we have

$$E(T|0) = n!\left[\prod_{k=1}^{n-1} (\theta + k)\right] \sum_{k=2}^{n} \frac{\beta_k(\theta)}{k^2(\theta + k - 1)^2} \qquad (22)$$

$$\mathrm{Var}\,(T|0) = n!\left[\prod_{k=1}^{n-1} (\theta + k)\right] \sum_{k=2}^{n} \frac{2\beta_k(\theta)}{k^3(\theta + k - 1)^3}$$

$$- \mathrm{E}^2(T|0). \qquad (23)$$

The likelihood function of T is given by

$$p_n(0|T) = \frac{\sum_{k=2}^{n} \beta_k(\theta)\,e^{-k(\theta+k-1)T}}{\sum_{k=2}^{n} \beta_k(0)\,e^{-k(\theta+k-1)T}}\,. \qquad (24)$$

The second situation is when $K = 1$. We have from (8) that

$$\gamma_{1,k} = \sum_{i=2}^{n} \frac{1}{\theta + k + i - 1} - \frac{1}{\theta + 2k - 1}\,.$$

It follows that

$$p_n(1, T) = \theta n!(n-1)! \sum_{k=2}^{n} \beta_k(\theta)$$

$$\times \left[\sum_{i=2}^{n} \frac{1}{\theta + k + i - 1}\right.$$

$$\left. - \frac{1}{\theta + 2k - 1} + kT\right] e^{-k(\theta+k-1)T}. \qquad (25)$$

Since it is known from WATTERSON (1975) that

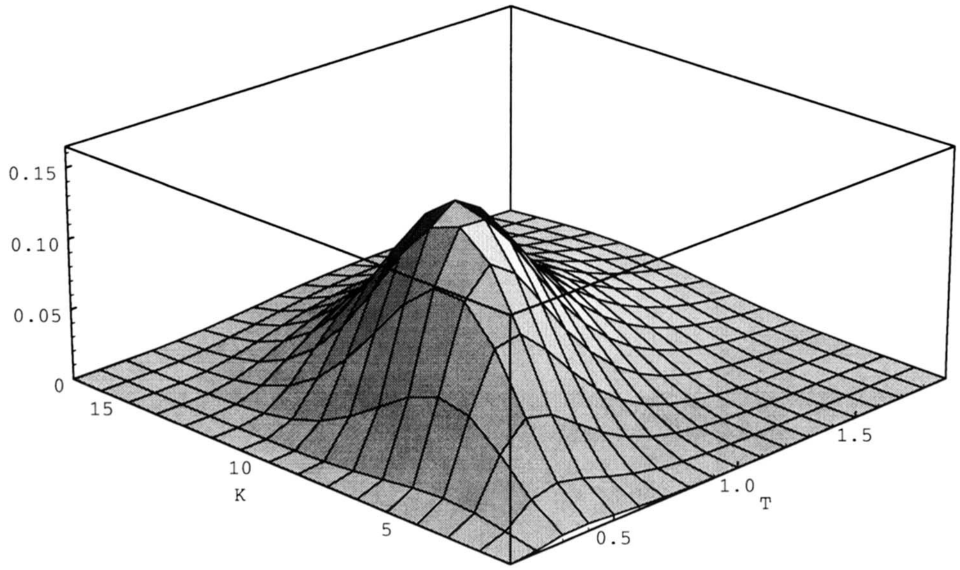

FIGURE 2.—Surface of $p_n(K, T)$ when $n = 30$ and $\theta = 2.0$.

$$p_n(K = 1) = \left(\prod_{k=1}^{n-1} \frac{k}{\theta + k}\right) \sum_{k=1}^{n-1} \frac{\theta}{\theta + k},$$

one can thus compute the values of the likelihood function and the posterior probability without using the iteration procedure specified by (9) and (10).

Finally since

$$\gamma_{2,k} = \sum_{i=2}^{n} \frac{1}{(\theta + k + i - 1)^2} - \frac{1}{(\theta + 2k - 1)^2} + \sum_{2 \le i < j \le n;\, i, j \ne k} \frac{1}{(\theta + k + i - 1)(\theta + k + j - 1)},$$

which is also easy to compute, we have

$$p_n(2, T) = \theta^2 n!(n - 1)! \sum_{k=2}^{n} \beta_k(\theta) \times [\tfrac{1}{2}(kT)^2 + \gamma_{1,k}(kT) + \gamma_{2,k}]\, e^{-k(\theta + k - 1)T}. \quad (26)$$

Before we consider how to estimate T from the value of K, it is helpful to gain some ideas on the shape of the joint probability density $p_n(K, T)$, the likelihood function $p_n(K|T)$ and the posterior probability $p_n(T|K)$. Figure 2 shows the surface of $p_n(K, T)$ for a sample of 30 sequences and $\theta = 2.0$.

It can be seen from Figure 2 that the peak of T shifts with K and vice versa. Figure 3, a and b, shows the likelihood function and the posterior probability of T respectively, for a number of values of K. It is clear by comparing the two panels (a and b) that the value of T corresponding to the peak of a likelihood function is smaller than that of a posterior probability when K is close to zero and becomes larger when K is large. This is a feature that determines the relationship between the maximum likelihood estimator and the other two estimators derived from the posterior probability distribution.

ESTIMATION OF T

Since both the joint probability of K and T and the marginal probability of K depend on θ, therefore, to estimate T from the value of K based on either the likelihood function or the posterior probability, one must know the value of θ or have an estimate of θ prior to the estimation of T. As an initial step, we shall assume in this paper that the value of θ is known.

Before we set forward to develop estimators of T, it is natural to ask whether K is informative about T. One way to answer this question is to examine the correlation coefficient, $\rho_n(\theta)$, between K and T given by

$$\rho_n(\theta) = \frac{E(KT) - E(K)E(T)}{\sqrt{\text{Var }(K) \text{ Var }(T)}}.$$

Since K is positively correlated with the total time length L of the genealogy of the sample and the latter is positively correlated with T, $\rho_n(\theta)$ is thus positive. However, if $\rho_n(\theta)$ is close to zero, it is likely that knowing the value of K is of little help for determining the value of T; on the other hand, if $\rho_n(\theta)$ is close to 1, knowing the value of K will be almost equivalent to knowing the value of T.

Consider the case of two sequences. The joint distribution of K and T is obviously equal to

$$p_2(K, T) = \frac{e^{-2\theta t}(2\theta t)^K}{K!} 2e^{-2t},$$

which can also be obtained from (5). Therefore

$$E(KT) = \int_0^\infty \left[\sum_k kt \frac{e^{-2\theta t}(2\theta t)^k}{k!} 2e^{2t}\right] dt$$
$$= 2\theta \int_0^t t^2 2e^{-2t} dt$$
$$= \theta.$$

and the correlation coefficient between K and T is

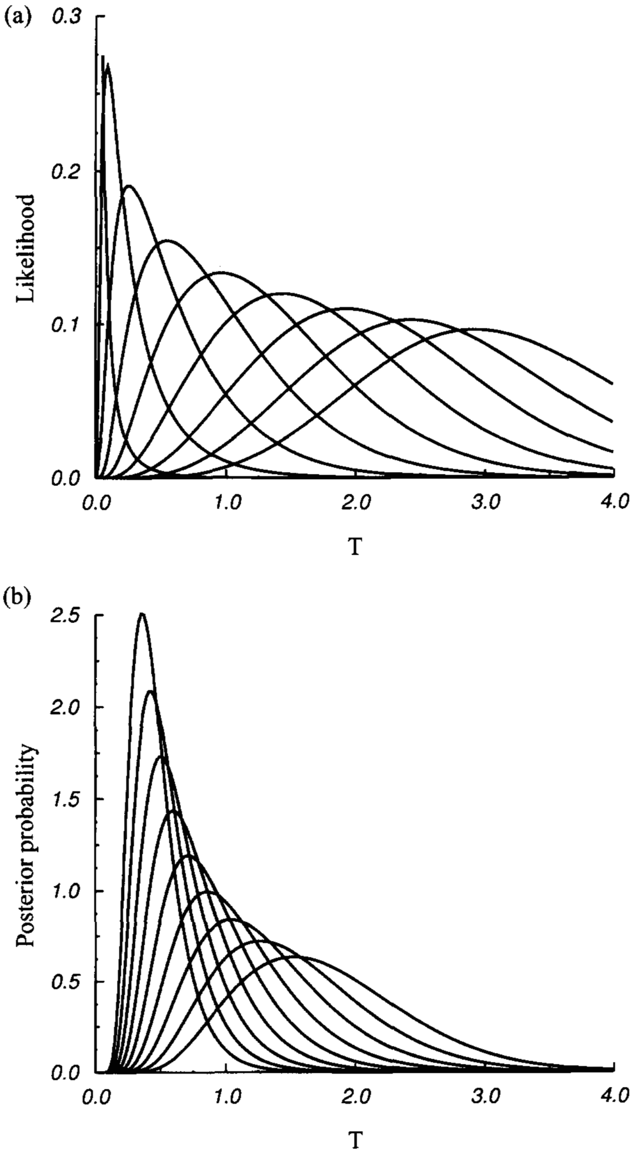

FIGURE 3.—Likelihoods (a) and posterior probabilities (b) for $n = 30$ and $\theta = 2$. In a and b, the curves with descending peaks correspond to $K = 0, 2, \ldots, 16$, respectively.

$$\rho_2(\theta) = \frac{\theta - \theta/2}{\sqrt{(\theta + \theta^2)\,{}^1\!/_4}} = \sqrt{\frac{\theta}{1 + \theta}}.$$

It is thus clear that $\rho_2(\theta)$ increases to 1 when θ approaches infinity. In other words, the value of K is a good indicator of the value of T when the value of K is likely to be large, and is a poor indicator of T when its value is likely to be small. Although we are unable to find simple analytical solution for $\rho_n(\theta)$ when $n > 2$, $\rho_n(\theta)$ can be computed numerically. Table 1 gives the values of $\rho_n(\theta)$ for a number of combinations of n and θ. It is clear from the table that $\rho_n(\theta)$ decreases with n for a given value of θ. This is because for a larger sample, there are more ways that the K segregating sites can be partitioned into states of the sample genealogy and therefore its value has less predictive power on the value of T. It is also true that $\rho_n(\theta)$ increases with θ for a given sample size n. Based on the information in Table 1, it seems reasonable to assume that $\rho_n(\theta)$ will approach 1 when θ approaches infinity for any sample size. To summarize, the informativeness of K on T depends on the value of θ. For the purpose of getting reliable estimate of T, one should examine loci with large mutation rate per site and obtain as longer sequences as possible.

TABLE 1

The correlation coefficient $\rho_n(\theta)$ between K and T

n	$\theta = 0.1$	0.5	1	2	5	10
2	0.30	0.58	0.71	0.82	0.91	0.95
5	0.25	0.49	0.62	0.74	0.86	0.91
10	0.22	0.44	0.57	0.69	0.82	0.88
20	0.20	0.41	0.53	0.65	0.79	0.86
50	0.18	0.37	0.47	0.62	0.76	0.83

We now consider the estimation of T from the value of K. Two types of estimator of T can be devised from the theory developed in the previous section. One is the maximum likelihood estimate and another is the Bayesian estimates. We consider them in turns.

Point estimators of *T*: The first point estimator we consider is the maximum likelihood estimate of T denoted $t_{\max}$, which is the value of T that maximizes the likelihood function of T given by (14). In other words, $t_{\max}$ is the solution for the following equation:

$$\frac{\partial \ln p_n(K|T)}{\partial T} = p_n^{-1}(K, T)\frac{dp_n(K, T)}{dT} - \phi_n^{-1}(T)\frac{d\phi_n(T)}{dT} = 0$$

where

$$\frac{dp_n(K, T)}{dT} = \sum_{k=2}^{n}\left\{\sum_{l=0}^{K} \alpha_{kl}k^l T^l \times [l/T - k(\theta + k - 1)]\right\}e^{-k(\theta+k-1)T} \quad (27)$$

and the value of $d\phi_n(T)/dT$ can be obtained by setting both $\theta = 0$ and $K = 0$ in (27).

Next we consider estimators derived from the posterior probability $p_n(T|K)$. Estimators of this type are commonly called Bayesian estimators. We consider two Bayesian estimators, one denoted t_{mode} is the value of T that maximizes the posterior probability, and another denoted t_{mean} is the conditional expectation of T, *i.e.*, $t_{\text{mean}} = E(T|K)$. Since $p_n(K)$ does not depend on T, t_{mode} is the value of T that maximizes $p_n(K, T)$. Therefore, t_{mode} is the solution for the following equation:

$$\sum_{k=2}^{n}\left\{\sum_{l=0}^{K} \alpha_{kl}k^l T^l\left[\frac{l}{T} - k(\theta + k - 1)\right]\right\}e^{-k(\theta+k-1)T} = 0.$$

To understand the relationship between these three estimators, consider first the case of two sequences. Since the likelihood function of T for a sample of two sequences is

$$p_2(K|T) = \frac{p_2(K, T)}{\phi_2(T)} = \frac{e^{-2\theta t}(2\theta t)^K}{K!}$$

and the posterior probability is

$$p_2(T|K) = (\theta + 1)\left(\frac{\theta + 1}{\theta}\right)^K \frac{e^{-2\theta t}(2\theta t)^K}{K!} 2e^{-2t},$$

it is easy to show that

$$t_{\max} = \frac{K}{2\theta}$$

$$t_{\text{mode}} = \frac{K}{2(\theta + 1)}$$

and

$$t_{\text{mean}} = \frac{K + 1}{2(\theta + 1)}.$$

We thus have the relationship $t_{\text{mode}} < t_{\text{mean}}$ for any given value of θ. Furthermore $t_{\text{mode}} \leq t_{\max} \leq t_{\text{mean}}$ when $K \leq \theta$ and $t_{\text{mode}} < t_{\text{mean}} < t_{\max}$ when $K > \theta$. Note that $E(K) = \theta$ when $n = 2$.

Examining these three estimators for $n = 3$, we found that none of them can be expressed as a linear function of K. When $n > 3$, these estimators become too complicated to be derived analytically. Therefore, we compared the numerical values of these estimators for a number of combinations of n, K and θ. Figure 4 gives two examples of the values these estimators. Figure 4a corresponds to $\theta = 2$ and sample size 10, and Figure 4b corresponds to $\theta = 5$ and sample size 30. The pattern of the values of the three estimators in a and b, as well as those in many other parameter settings not shown here, enable us to conclude that

1. The value of each of the three estimators increases with K.
2. For any values of θ and sample size n, t_{mode} is smaller than t_{mean}. This is because the posterior probability of T is skewed to the left.
3. The maximum likelihood estimate $t_{\max}$ is equal to zero when $K = 0$ and is the smallest among the three estimators when K is small.
4. The value of the maximum likelihood estimator $t_{\max}$ increases with K most rapidly and eventually becomes the largest among the three estimators after K is larger than a value that is larger than $E(K)$.

Interval estimate of *T*: Besides the two Bayesian point estimators t_{mode} and t_{mean}, one can construct interval estimates of T from the posterior probability $p_n(T|K)$. For example, the 95% interval estimate of T can be defined as $(T_{2.5}, T_{97.5})$, where T_x is the value of S such that

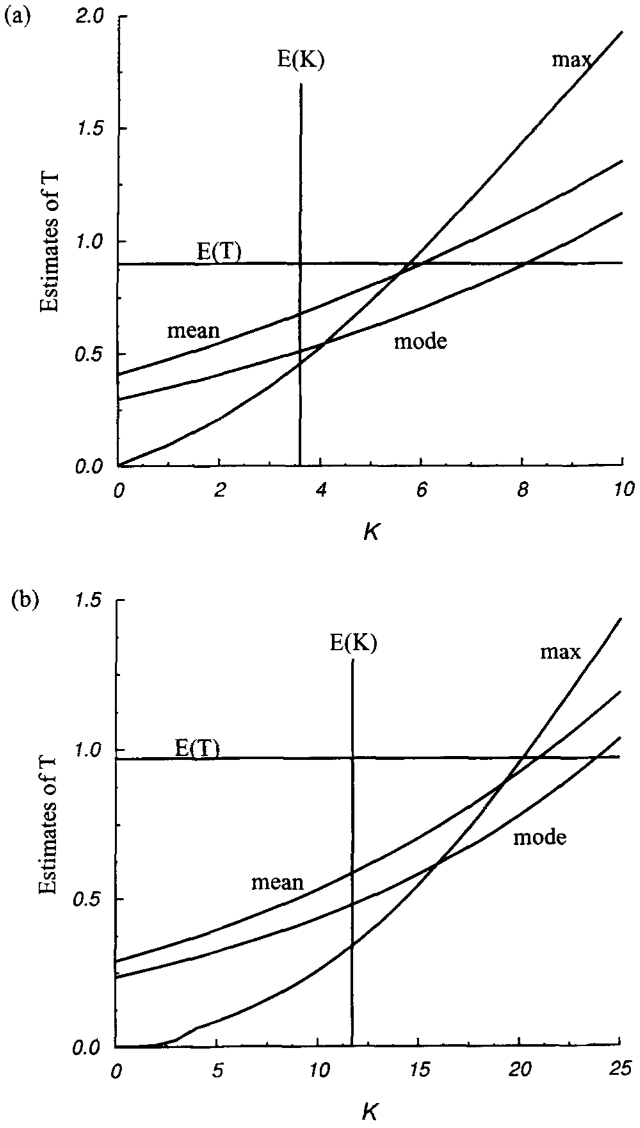

FIGURE 4.—Estimates of T for given values of K. (a) $n = 10$ and $\theta = 2$; (b) $n = 30$ and $\theta = 5$.

$$x\% = \int_0^S p_n(t|K)\,dt = \frac{1}{p_n(K)}\int_0^S p_n(K, t)\,dt$$

where $\int_0^S p_n(K, t)\,dt$ can be shown to be

$$p_n(K) - \frac{\theta^K n!(n-1)!}{K!}\sum_{k=2}^{n}\sum_{l=0}^{K}\alpha_{kl} \times \sum_{i=0}^{l}\frac{l!k^{l-i-1}}{(l-i)!(\theta + k - 1)^{i+1}} S^{l-i}e^{-k(\theta+k-1)S}.$$

Obviously $T_{2.5}$ should be smaller than $T_{97.5}$.

Figure 5 gives examples of the 95% interval estimate of T for several values of θ in a sample of 50 sequences. It is clear that the length of 95% interval of T becomes shorter with increasing θ. Because a shorter interval of

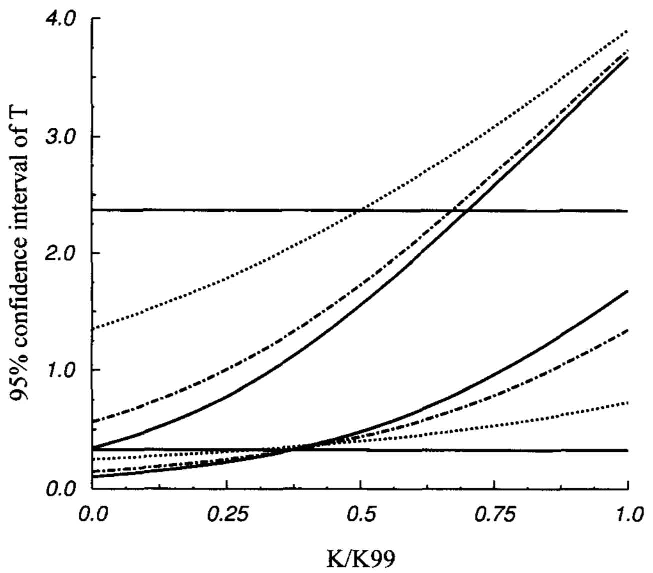

FIGURE 5.—The 95% interval estimate of T for a sample of 50 sequences. The two dotted lines, dash-dotted lines and solid lines correspond to the upper and lower limits of the interval estimate for $\theta = 1$, 5 and 10, respectively; the two horizontal lines correspond to the interval estimate based on the prior distribution, $\phi_n(T)$, of T. $K99$ is the number of segregating sites such that $p(K \leq K99 | \theta) \approx 0.99$. The values of $K99$ for $\theta = 1$, 5 and 10 are 12, 46 and 90, respectively.

T implies better estimate of T, Figure 5 concurs with our earlier analysis of the correlation coefficient between K and T. Figure 5 also shows that a large θ improves mainly the estimate of the upper bound of T when K is small and the lower bound of T when K is large.

AN EXAMPLE: THE HUMAN Y CHROMOSOME

We shall consider the sample of DNA sequences by DORIT *et al.* (1995) from an intron of *ZFY* gene in the human Y chromosome. The sample consists of 38 sequences of 738 base pairs and has no sequence variation ($K = 0$). Since FU and LI (1996) (also see DONNELLY *et al.* 1996, WEISS and VON HAESELER 1996) have already analyzed this sample, we shall give a supplementary analysis below.

To estimate the age of the MRCA of this sample, one has to obtain an estimate of the value of $\theta = 2N_m\mu$. Because homologous DNA sequences from several primates were also available, DORIT *et al.* (1995) estimated the mutation rate per sequence per years as 0.98×10^{-6}. Assume 20 years as one human generation, the mutation rate (μ) per sequence per generation is thus 1.96×10^{-6}. In additions to the value of μ, we need to know the value of N_m. Figure 6 shows the curves of the posterior probability for several values of N_m. One can see that a larger value of N_m results in a more concentrated distribution of T. If one fixes the value of N_m and varies the value of μ, the effect on the posterior probability would be the similar to that shown in Figure 5. In other words, with increasing mutations rate, the posterior probability distribution will be more concentrated, therefore the inference on T will be more accurate.

Assuming equal sex ratio, FU and LI (1996) took $N_m = 5000$ according to TAKAHATA (1993). This results in $\theta = 0.196$. FU and LI (1996) obtained $t_{\text{mode}} = 114{,}000$ yr, $t_{\text{mean}} = 174{,}000$ yr and the 95% interval estimate of T is from 60,000 to 408,000 yr. The maximum likelihood estimate $t_{\max}$ of T is equal to zero as pointed out earlier.

One can also compute the Bayesian estimates t_{mode} and t_{mean}, and the 95% interval estimate of T directly from the prior distribution $\phi_n(T)$. This yields that $t_{\text{mode}} = 124{,}000$, $t_{\text{mean}} = 195{,}000$ yr and the 95% interval estimate of T from 65,000 to 473,000 yr. Comparing these point estimates to those based the posterior probability distribution, we can see that the former are smaller. The interval estimate of T based on the posterior probability, which is a better indicator of the quality of the information in the sample, is 60,000 yr narrower than that based on the prior distribution of T. The improvement is apparently significant though not dramatic, which is not surprising for two reasons. First, when $\theta = 0.196$ the correlation coefficient between K and T is 0.25; therefore, the value of K provides only a modest amount of informative about T. Second, one can compute the probability of no variation from (20), and with $\theta = 0.196$ this probability is 0.42, which is not small at all. Therefore, the posterior distribution of T is not too different from the prior distribution of T, which is equivalent to the posterior probability of T with $\theta = 0$.

Since our analytical results are derived under the WRIGHT-FISHER model with a constant effective population size and since the human population is apparently subdivided and is growing, the above analysis should be viewed as preliminary. However, NEI and TAKAHATA (1993) showed that, when population subdivision is not substantial (*i.e.*, $4Mm$ is not too small where m is the migration rate), the formula, $4N(1 - 1/n)$, of the mean age of the MRCA of a sample from a random mating population is also a good approximation to that of a sample from a subdivided population with N replaced by the effective population size of the subdivided population. Therefore, the theory and estimators developed in this paper should be an useful starting point for the inferences on T.

DISCUSSION

We have focused on the age of the MRCA of a sample from a population. It is often more interesting to be able to estimate the age of the MRCA of a population, such as the cases of the human mitochondria and Y chromosomes. The age of the MRCA of a sample can be different from that of a population and thus younger. SAUNDERS *et al.* (1984) showed that the probability the two are the same is

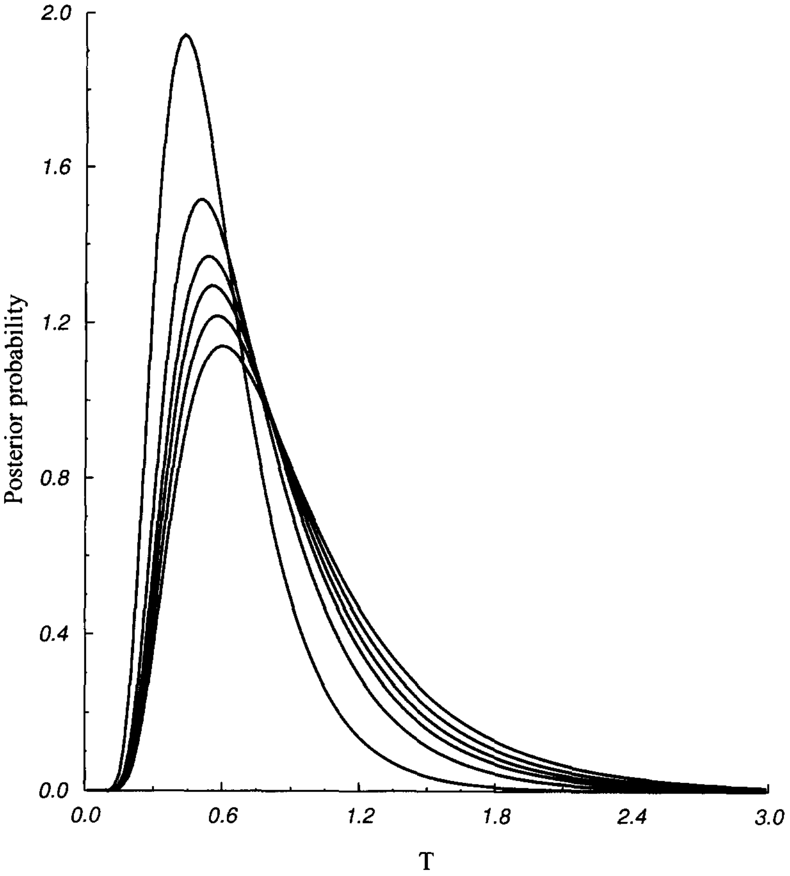

FIGURE 6.—Posterior probability $p_n(T|0)$ with different effective population sizes for a sample of 38 sequences, given that $\mu = 0.98 \times 10^{-6} \times 20$. The curves with descending peaks correspond to N_m = 30,000, 15,000, 10,000, 7500, 5000 and 2500, respectively.

$$\frac{(n-1)(N+1)}{(n+1)(N-1)}.$$

Because sample size is usually much smaller than the effective population size N, the above probability is approximately equal to $(n-1)/(n+1)$. It follows that when n is large, it is reasonable to treat the MRCA of a sample as that of a population. For example, the probability that the MRCA of a random sample of 38 sequences is the same as the MRCA of a population is 0.95. Therefore, it is reasonable to treat the estimate of the age of the MRCA of the sample by DORIT *et al.* (1995) as that of the male human population, although one would feel safer if the sample size had been 100, which gives 0.98 probability that the two MRCAs are the same.

We presented in this paper three point estimators of T and showed that their values for a given sample are usually different. In particular, the maximum likelihood estimate $t_{\max}$ can be substantially different from the two Bayesian estimates t_{mode} and t_{mean}. This raises the question on which of the three estimators should be preferred. As we have seen that when there is no variation in a given sample, the maximum likelihood estimate $t_{\max}$ of T is 0, which is by all means a bad estimate. The maximum likelihood estimator ignores the fact that T has a bell-shaped distribution so that it is unlikely to be either too small or too large and thus yields estimates that seems to be too small when K is close to zero and too large when K is large. Therefore, Bayesian estimates should be preferred over the maximum likelihood estimate of T from the value of K. Between the two Bayesian estimators, t_{mode} should be preferred over t_{mean}, because the former is the most likely value of T for the given value of K while the latter is the average value of T. When one has to draw conclusions about T from a single sample, the average value of T appears to be less relevant. However, this judgment is necessarily subjective to some extent and I recommend to report the values of all the three estimators when analyzing real samples.

We also presented an interval estimate of T derived from the posterior probability distribution of T. It should be emphasized that the resulting 95% interval of T is not the 95% confidence interval of any of the three point estimators discussed in this paper. This fact can be overlooked easily and when the phrase "interval of T" is used loosely, it is tempting to interpret it as the confidence interval of a point estimator, although the two intervals should be correlated. Because the interval estimate of T allows one to make a very informative probabilistic statement, such as, with 0.95 probability T is in a certain interval, I strongly recommend the use of interval estimate of T.

We showed that the usefulness of the value of K as a predictor of the value of T depends on the value of θ. The larger the θ is, the more informative the value of K becomes. This observation is in line with the finding that the accuracy in the estimation of θ from K increases with the value of θ (FELSENSTEIN 1992; FU and LI 1993a). Because we assume that θ is known in this paper, while in reality the same sample will probably be used to estimate both θ and T, a sample of DNA sequences from a locus with large value of θ will improve the estimations of both θ and T.

Finally, it has been demonstrated that phylogenetic information in a sample can improve the accuracy in the estimation of θ (*e.g.*, FU 1994); it is thus of interest to explore the possibility of incorporating phylogenetic information in a sample into the estimation of the age of the MRCA of the sample. One such approach has been developed by GRIFFITHS and TAVARÉ (1994). The extent of the improvement of inference by such approaches remains to be seen, but the estimation of the age of the MRCA based on the number of segregating sites should be efficient at least for DNA samples with few segregating sites.

I thank Drs. J. FELSENSTEIN, W. H. LI and N. TAKAHATA, and a reviewer for their comments and suggestions. This research was supported by National Institutes of Health grant R29 GM-50428.

LITERATURE CITED

DONNELLY, P., S. TAVARÉ, D. J. BALDING and R. C. GRIFFITHS, 1996 Estimating the age of the common ancestor of mem 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.

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

FU, Y. X., 1994 A phylogenetic estimator of effective population size or mutation rate. Genetics **136:** 685–692.

FU, Y. X., 1996 New statistical tests of neutrality for DNA samples from a population. Genetics **143:** 557–570.

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.

FU, Y. X., and W. H. LI, 1996 Estimating the age of the common ancestor of mem 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.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Pop. Biol. **23:** 183–201.

KIMURA, M., 1970 The length of time required for a selectively neutral mutant to reach fixation through random frequency drift in a finite population. Genet. Res. **15:** 131–133.

KINGMAN, J. F. C., 1982a The coalescent. Stochastic processes and their applications. **13:** 235–248.

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

NEI, M., and N. TAKAHATA, 1993 Effective population size, genetic diversity, and coalescent time in subdivided populations. J. Mol. Evol. **37:** 240–244.

SAUNDERS, L. W., S. TAVARÉ and G. A. WATTERSON, 1984 On the genealogy of nested subsamples from a haploid population. Adv. Appl. Prob. **16:** 471–491.

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.

TAJIMA, F., 1990 Relationship between DNA polymorphism and fixation time. Genetics **125:** 447–454.

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

TAVARÉ, S., 1984 Line of descent and genealogical process and their applications in population genetics models. Theor. Popul. Biol. **26:** 119–164.

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

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

Communicating editor: N. TAKAHATA

APPENDIX: DERIVATION OF $p_n(K, T)$

Let

$$g_k = k(\theta + k - 1)$$

$$T_k = T - t_2 - \cdots - t_k$$

$$L(k, i) = kT + \sum_{j=2}^{i-1} (j - k) t_j,$$

and

$$G(k, i) = g_k T + \sum_{j=2}^{i-1} (g_j - g_k) t_j.$$

Because of the constraint $t_2 + \cdots + t_n = T$, t_n is equal to T_{n-1}. It follows that Equation 4 can be written as

$$p_n(K, T) = \frac{\theta^K n!(n-1)!}{K!} \int_0^T \int_0^{T_2} \cdots \int_0^{T_{n-2}} L^K(n, n) e^{-G(n,n)} dt_{n-1} \cdots dt_2$$

which can be computed by integrating with respect to $t_{n-1}, \ldots, t_2$ in turns. Note that it is equivalent to write $L^K(n, n) e^{-G(n,n)}$ as

$$f_{n-1} = \sum_{k=n}^{n} \sum_{l=0}^{K} \alpha_{kl}(n) L^l(k, n) e^{-G(k,n)},$$

where

$$\alpha_{nK}(n) = 1, \alpha_{n0}(n) = \cdots = \alpha_{nK-1}(n) = 0. \quad (28)$$

Suppose that the function to be integrated with respect to t_i is

$$f_i = \sum_{k=i+1}^{n} \sum_{l=0}^{K} \alpha_{kl}(i+1) L^l(k, i+1) e^{-G(k,i+1)}.$$

Then because

$$\frac{d^j L^l(k, i+1)}{dt_i^j} = \frac{l!(i-k)^j}{(l-j)!} L^{l-j}(k, i+1),$$

$$\int_0^x L^l(k, i+1) e^{-G(k,i+1)} dt_i = \sum_{j=0}^{l} \frac{l!(i-k)^j}{(l-j)!(g_i - g_k)^{j+1}} L^{l-j}(k, i+1) \Bigg|_x^0.$$

The integration with respect to t_i results in

$$
\begin{aligned}
\int_0^{T_{i-1}} f_i\,dt_i &= \sum_{k=i+1}^{n}\sum_{l=0}^{K}\alpha_{kl}(i+1)\sum_{j=0}^{l}\frac{l!\,(i-k)^j}{(l-j)!\,(g_i-g_k)^{j+1}}L^{l-j}(k,i+1)\,e^{-G(k,i+1)}\Bigg|_{T_{i-1}}^{0}\\
&= \sum_{k=i}^{n}\sum_{l=0}^{K}\alpha_{kl}(i+1)\sum_{j=0}^{l}\\
&\qquad\left\{\frac{l!\,(i-k)^j}{(l-i)!\,(g_i-g_k)^{j+1}}\left[L^{l-j}(k,i)\,e^{-G(k,i)}-L^{l-j}(i,i)\,e^{-G(i,i)}\right]\right\}\\
&= \sum_{k=i}^{n}\sum_{l=0}^{K}\alpha_{kl}(i)\,L^{l}(k,i)\,e^{-G(k,i)},
\end{aligned}
$$

where

$$
\alpha_{kl}(i) = \sum_{j=l}^{K}\frac{\alpha_{kj}(i+1)}{g_i-g_k}\frac{j!}{l!}\left(\frac{i-k}{g_i-g_k}\right)^{j-l},
$$

$$
k = i+1,\ldots,n
$$

$$
\alpha_{il}(i) = -\sum_{k=i+1}^{n}\alpha_{kl}(i) \tag{29}
$$

for $l = 0,\ldots,K$.

The last integration with respect to t_2 yields

$$
\begin{aligned}
p_n(K,T) &= \frac{\theta^K n!\,(n-1)!}{K!}\sum_{k=2}^{n}\sum_{l=0}^{K}\alpha_{kl}(2)\\
&\qquad\times L^{l}(k,2)\,e^{-G(k,2)}\\
&= \frac{\theta^K n!\,(n-1)!}{K!}\sum_{k=2}^{n}\sum_{l=0}^{K}\alpha_{kl}(2)\\
&\qquad\times (kT)^{l}e^{-k(\theta+k-1)T}
\end{aligned}
\tag{30}
$$

Therefore, $p_n(K,T)$ can be calculated from (30) once we know the values of $a_{kl}(2)$, which can be obtained sequentially from the iteration (29) with initial conditions given by (28). Substituting $k(\theta+k-1)$ for g_k in (29) results in the iteration procedure defined by (9) and (10).

We now show that α_{kl} (2) is also given by (6). It is easy to see from the iteration procedure described above that

$$
\alpha_{n\,l}(n-1) = \frac{K!}{l!\,(g_{n-1}-g_n)}\left(\frac{(n-1)-n}{g_{n-1}-g_n}\right)^{K-l}
$$

$$
\alpha_{n-1\,l}(n-1) = \frac{K!}{l!\,(g_n-g_{n-1})}\left(\frac{n-(n-1)}{g_n-g_{n-1}}\right)^{K-l}.
$$

Suppose that

$$
\alpha_{kl}(i) = \frac{K!}{l!}\left(\prod_{m\geq i,m\neq k}\frac{1}{g_m-g_k}\right)
\times\sum_{j_i+\cdots+j_n=K-l,\,j_k=0}\prod_{m}\left(\frac{m-k}{g_m-g_k}\right)^{j_m},
$$

which is obviously true for $i = n-1$. Then we have from (29) that for $k \geq i$

$$
\begin{aligned}
\alpha_{kl}(i-1) &= \sum_{j\geq l}^{K}\frac{a_{kj}(i)\,j!}{(g_{i-1}-g_k)\,l!}\left(\frac{i-1-m}{g_{i-1}-g_k}\right)^{j-l}\\
&= \frac{K!}{l!}\left(\prod_{m\geq i-1,m\neq k}\frac{1}{g_m-g_k}\right)\\
&\qquad\times\sum_{j_{i-1}+\cdots+j_n=K-l,\,j_k=0}\prod_{m}\left(\frac{m-k}{g_m-g_k}\right)^{j_m}.
\end{aligned}
$$

Although it is not easy to show analytically that this equation also holds for $k = i-1$, comparing the numerical values of $\alpha_{kl}(i-1)$ computed by the above equation and by the iteration procedure indicates that it indeed holds for all values of $k = i-1,\ldots,n$. It thus follows that

$$
\alpha_{kl} = \alpha_{kl}(2) = \frac{K!}{l!}\left(\prod_{i\neq k}\frac{1}{g_i-g_k}\right)
\times\sum_{j_2+\cdots+j_n=K-l,\,j_k=0}\prod_{m}\frac{1}{(\theta+m+k-1)^{j_m}},
$$

and furthermore

$$
\begin{aligned}
\prod_{i\neq k}\frac{1}{g_i-g_k} &= \prod_{i\neq k}\frac{1}{(i-k)(\theta+i+k-1)}\\
&= \frac{\theta+2k-1}{(2-k)\cdots(-1)\cdot1\cdots(n-k)(\theta+k+1)\cdots(\theta+n+k-1)}\\
&= \frac{(-1)^k(\theta+2k-1)}{(k-2)!\,(n-k)!\prod_{i=1}^{n-1}(\theta+k+i)}.
\end{aligned}
$$

We thus have Equations 6–8.