A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences

ANALABHA BASU and PARTHA P. MAJUMDER*

Anthropology & Human Genetics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700 108, India

Abstract

We have compared two statistical methods of estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences, which have been proposed by Templeton (1993) and Bandelt *et al*. (1995). Monte-Carlo simulations were used for generating DNA sequence data. Different evolutionary scenarios were simulated and the estimation procedures were evaluated. We have found that for both methods (i) the estimates are insensitive to demographic parameters and (ii) the standard deviations of the estimates are too high for these methods to be reliably used in practice.

[Basu A. and Majumder P. P. 2003 A comparison of two popular statistical methods for estimating the time to most recent common ancestor (TMRCA) from a sample of DNA sequences. *J. Genet*., **82**, 7–12]

Introduction

The assumption that underlies the statistical reconstruction of the evolutionary history of a set of contemporary populations is that new populations evolve over time by binary fission from ancestral populations. Looking backwards in time, therefore, a set of contemporary populations will coalesce pairwise at different points of time, until finally there is a coalescent event to the most recent common ancestor (MRCA) of all the populations. Such reconstruction can be done by using DNA sequence data generated from samples of individuals drawn from each of the contemporary populations under consideration. The two major features and parameters to be estimated from such data are (i) the topology of the coalescence events and (ii) the times of coalescence to common ancestors of the populations, including the time to MRCA (TMRCA). Both these features and parameters are known to be affected by demographic scenarios that prevailed during the process of evolution (Hudson 1991; Nordborg 2001). Coalescent theory (Kingman 1982a,b) provides a probabilistic framework and a method for reconstruction of evolution from DNA sequence data. The framework is simpler when one is dealing with a haploid, nonrecombining DNA molecule, such as mitochondrial DNA. Even with haploid DNA sequence data, estimating TMRCA based on a sample remains a major challenge. Saunders *et al*. (1984) have shown that, although the TMRCA estimated from a sample can be different from the true TMRCA, the probability that the estimate will coincide with the true value is

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

where n is the sample size and N ($\gg n$) the population size (assumed to have been large and constant over evolutionary time). Thus, provided that we are dealing with numerically large and temporally constant-size populations, even with a sample of 38 haploid DNA sequences (n), the probability of correctly estimating the true TMRCA is 0.95. Thus the TMRCA of a sample is a reasonably good estimate of the TMRCA of the population (Saunders *et al*. 1984). Statistical methods have been developed to estimate TMRCA from a sample. However, the temporal constancy of population size is a crucial assumption underlying these methods. In practice, a population is expected to encounter demographic pressures (such as bottlenecks

*For correspondence. E-mail: ppm@isical.ac.in.

Keywords. coalescence; Monte-Carlo simulation; evolution.

and expansions), resulting in violation of this assumption. The purpose of this study is to evaluate the impact of evolutionarily variable demographic scenarios on the estimates of TMRCA obtained by using two popular statistical methods (Templeton 1993; Bandelt *et al*. 1995; Saillard *et al*. 2000).

Methodology

***The coalescent*:** For completeness, we provide some key results of coalescent theory and briefly describe the two popular statistical methods. The Kingman coalescent (Kingman 1982a,b) is a probability model for the genealogical tree of a random sample of n genes drawn from a large population. A genealogical tree for a sample of size $n = 5$ is depicted in figure 1. Time is measured continuously in the coalescent. The time t_j during which the sample has j distinct lineages ($2 \le j \le n$) follows an exponential distribution with parameter $j(j-1)/2$ (Tajima 1983; Hudson 1991; Nordborg 2001). The random variables denoting the times for different js are independent. This description provides a close approximation to a range of population genetics models in which time is expressed in generations. An even larger class of models is approximated if a unit of coalescence time is interpreted as N/s^2 generations, where s^2 is the variance in the number of offspring produced by an individual (Kingman 1982a). We shall assume $s = 1$. We are primarily interested in the height of the tree T_n, i.e. the TMRCA.

It may be noted that

$$T_n = t_n + t_{n-1} + \ldots + t_2 = \sum t_i, \quad \text{and}$$
$$E(t_j) = 2/\{j(j-1)\},$$

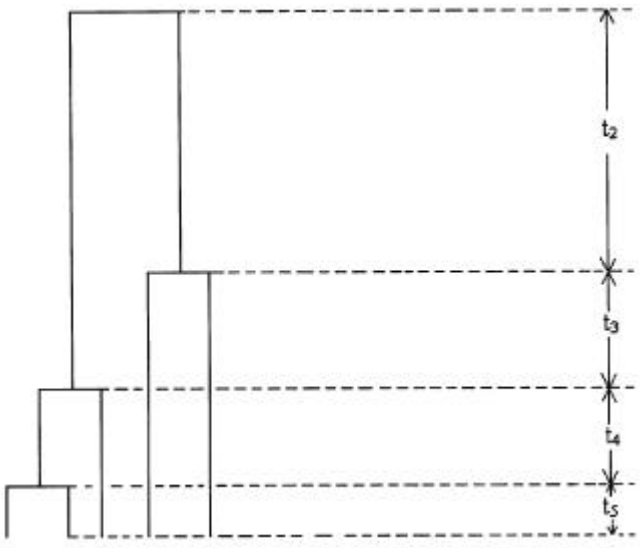

Figure 1. Genealogical tree of a sample of five genes.

$$E(T_n) = 2(1 - 1/n),$$
$$\mathrm{Var}(T_n) = 8\sum 1/j^2 - 4(1 - 1/n)^2 \to 4p^2/3 - 12 \approx 1.16.$$

In all the above expressions time is measured in units of N generations.

It is clear that as n increases $E(T_n)$ very rapidly converges to 2. Also, T_n, the TMRCA, has a large variance relative to the mean and this ratio does not decrease much with increase in sample size.

The times at which mutations occur are modelled in the coalescent by assuming that these times form a Poisson process of constant rate m, where m is the mutation rate per sequence per generation. This means that the number of mutations that may have accumulated on a branch of time length l is a realization from the Poisson distribution with mean ml. For DNA sequence data, if we assume that mutation rate has remained constant across sites and over time, then m is equal to the sequence length times the mutation rate per site per generation.

Although there are several methods available for estimating TMRCA from a sample of DNA sequences (Fu and Li 1997; Tavare *et al*. 1997), two methods are widely used (e.g. Mountain *et al*. 1995; Saillard *et al*. 2000) primarily because of conceptual simplicity and ease of interpretation.

Under the infinite sites model (Ewens 1979), all information in two DNA sequences is captured by the total number of segregating sites (S_2). Since $E(S_2|T_2) = qT_2$, one estimator of T_2, which for a sample of two sequences is the TMRCA, is S_2/q. This and similar approaches (Hammer 1995) are not capable of utilizing prior historical demographic information.

Using Bayes's theorem, Tajima (1983) noted that if $S_2 = k$ then the distribution of T_2 is gamma with parameters $1 + k$ and $1 + q$. In particular,

$$E(T_2|S_2 = k) = (1 + k)/(1 + q),$$
$$\mathrm{Var}(T_2|S_2 = k) = (1 + k)/(1 + q)^2.$$

Templeton (1993) considered the problem of estimating the TMRCA of n (> 2) sequences by extending the analytical results that hold for $n = 2$ and calculated the number of differences for each pair of sequences whose common ancestor is the root of the tree and then averaged these pairwise differences. He also observed that this value, $\hat{k}$, of k varied little over plausible reconstructed trees. He then substituted k by $\hat{k}$ in the previous equations for $E(T_2|S_2 = k)$ and $\mathrm{Var}(T_2|S_2 = k)$. In a different study Hammer (1995) estimated the TMRCA for multiple sequences by substituting the largest value of k among all pairs in the previous equations. This is not a proper approach, because Donnelly and Kurtz (1997) have shown that the maximum number of differences between a pair of sequences chosen from this set of n sampled sequences goes to infinity as n goes to infinity. This is true even when T_n is bounded.

A popular alternative to the above procedures of estimating TMRCA is to use median-joining network (MJN) analysis (Bandelt *et al*. 1995). In this analysis, a genealogy of n individuals is considered as an ultrametric tree, in which the lengths of links are scaled to time and each interior node corresponds to a coalescent event. If there are k ($\leq 2n-2$) links of lengths $t_1, t_2, \ldots, t_k$ time units and if the clade defined by the ith link carries n_i individuals ($i = 1, 2, \ldots, k$) then the coalescent time t can be expressed as

$$t = (n_1t_1 + n_2t_2 + \ldots + n_kt_k)/n.$$

If m denotes the mutation rate, expressed as the expected number of (scored) mutations in a sequence segment per time unit, one may associate to the ith link a Poisson-distributed random variable X_i with parameter $m_i = t_im$. The random variable $X = (n_1X_1 + n_2X_2 + \ldots + n_kX_k)/n$ has the expected value

$$E(X) = \{(n_1t_1 + n_2t_2 + \ldots + n_kt_k)/n\}m = tm,$$

and variance

$$V(X) = \{(n_1^2t_1 + n_2^2t_2 + \ldots + n_k^2t_k)/n^2\}m,$$

assuming independence of $X_1, X_2, \ldots, X_k$.

***Simulation method*:** We evaluated the performance of these two methods (Templeton 1993; Bandelt *et al*. 1995) for estimating the coalescent times from DNA sequence data. The data set consisted of nucleotide sequences from homologous segments of DNA sampled from different individuals. The data generated are similar to haploid nucleotide sequences, such as of the mitochondrial DNA HVS1 (http://www.hvrbase.org).

We have used a forward propagating algorithm to generate simulated DNA sequence data. In this algorithm a nucleotide sequence of specified length and base composition is created by a multinomial random number generator with cell probabilities equal to the probabilities of the four bases. A completely homogeneous founding population of a given size is then formed by making the appropriate number of copies of the randomly generated nucleotide sequence. The founding population then evolves in accordance with the Wright–Fisher model (Ewens 1979), i.e. a new generation is formed by sampling from the previous generation with replacement. The numerical size of the succeeding generations is controlled after the founding population is created. In this study we have considered two demographic scenarios: (i) constancy of population size over generations and (ii) exponential growth in size, allowing for variability in the growth parameter over generations; that is, when the size of a new generation is determined, we randomly selected the appropriate number of sequences from the gene pool of the previous generation with replacement. Then, using the assumed value of the mutation rate, we calculated the expected number of mutations per generation, and determined the number of new mutations to be introduced in each generation. If the expected number of mutations per generation is denoted by y, then we randomly chose and mutated $[y]$ or $[y] + 1$ sites, where $[y]$ denotes the largest integer $\leq y$. Choice between $[y]$ or $[y] + 1$ was made randomly by generating a random number u from the uniform [0, 1] distribution, where $[y]$ was chosen if u was less than $y - [y]$. Suppose there are N_t individuals in generation t, each with data on a sequence of L nucleotide sites. To introduce a new mutation in generation t, a site was chosen with probability $1/(N_t \times L)$ and mutated. If x_1 is one such observation, then the mutation is introduced at the nucleotide position $((x_1/L) - [x_1/L]) \times L$ of the $[x_1/L]$th individual. While introducing the mutation, we did not consider any prior information on mutational histories of the site or the individual, thus allowing for parallel, recurrent and back mutations to occur. This process was repeated for a stipulated number of generations. The population thus generated was treated as the present population and a random sample of size n was drawn without replacement. This sample of n sequences was then used to estimate the TMRCA of the population. The estimated TMRCA was compared to the actual number of generations used in the simulation.

***Simulation parameters*:** Since estimates of TMRCA can be affected by various parameters, we have investigated the effects of variation in four crucial parameters. These are:

(i) The number of bases (L) of the nucleotide sequence; we have used two different values of L, namely 200 and 400.

(ii) Variability in population size over generation, which was introduced through a parameter a. We have used an exponential growth model. In this model, if N_t denotes the population size in generation t, then $N_{t+1} = N_te^a$. In order that N_{t+1} is an integer, we have chosen either $[N_te^a]$ or $([N_te^a] + 1)$. Choice between $[N_te^a]$ or $[N_te^a] + 1$ was made randomly by generating a random number u from the uniform [0, 1] distribution; $N_{t+1} = [N_te^a]$ was chosen if u was $< (N_te^a) - [N_te^a]$; otherwise $N_{t+1} = [N_te^a] + 1$ was chosen. We have used three different values of a, namely 0, 0.001, 0.005.

(iii) The number of generations (g); three different values of g, namely 250, 500 and 1000, were used.

(iv) Mutation rate (m); two values were used, $m = 10^{-5}$ per site per generation and $m = 5 \times 10^{-5}$ per site per generation. These values roughly correspond to the observed rates in human autosomal and mitochondrial hypervariable segment-1, respectively. We note that, although the relevant theoretical equations are functions of Nm, we have varied N and m independently to study the effect of parallel and back mutations, which are possibly introduced when m is large.

Table 1. Mean (± s.d.) values of TMRCA estimated by Templeton's (T_1) and median-joining-network (T_2) methods for various sets of parameter values.

Sequence ength (L)	No. of generations (g)	Growth rate (a)	$m = 10^{-5}$			$m = 5 \times 10^{-5}$		
			$\hat{T}_1 \pm$ s.d.	$\hat{T}_2 \pm$ s.d.	Correlation	$\hat{T}_1 \pm$ s.d.	$\hat{T}_2 \pm$ s.d.	Correlation
200	250	0	361 ± 316.1	398 ± 313.9	0.95	216 ± 132.4	178 ± 123.7	0.98
		0.001	363 ± 338.5	355 ± 329.5	0.89	198 ± 135.8	108 ± 114.4	0.88
		0.005	346 ± 278.2	276 ± 234.1	0.86	187 ± 106.1	95 ± 75.6	0.88
	500	0	531 ± 503.1	583 ± 620.0	0.93	374 ± 179.8	387 ± 262.2	0.80
		0.001	507 ± 412.7	446 ± 414.2	0.87	326 ± 178.4	298 ± 191.7	0.94
		0.005	428 ± 346.2	251 ± 306.4	0.85	394 ± 139.6	303 ± 112.0	1.00
	1000	0	908 ± 808.0	1167 ± 1080.0	0.93	629 ± 318.6	639 ± 376.2	0.94
		0.001	825 ± 694.5	890 ± 805.4	0.96	498 ± 322.8	505 ± 376.7	0.94
		0.005	858 ± 394.2	644 ± 312.5	1.00	910 ± 178.5	696 ± 153.0	0.99
400	250	0	218 ± 212.3	374 ± 422.5	0.93	205 ± 91.0	342 ± 167.5	0.96
		0.001	231 ± 177.4	305 ± 294.0	0.82	199 ± 89.0	217 ± 139.0	0.88
		0.005	225 ± 167.5	288 ± 345.2	0.81	206 ± 74.7	215 ± 103.5	0.89
	500	0	429 ± 370.7	833 ± 905.4	0.92	387 ± 162.3	701 ± 331.7	0.98
		0.001	439 ± 306.8	962 ± 774.3	0.89	351 ± 150.0	620 ± 300.6	0.98
		0.005	375 ± 201.4	590 ± 317.6	1.00	437 ± 124.8	681 ± 215.8	0.99
	1000	0	637 ± 464.3	1331 ± 1031.4	0.95	611 ± 219.0	1190 ± 518.0	0.94
		0.001	751 ± 527.3	1520 ± 1190.5	0.91	606 ± 204.8	1129 ± 444.3	0.96
		0.005	834 ± 287.8	1262 ± 469.2	1.00	912 ± 81.6	1375 ± 145.6	0.99

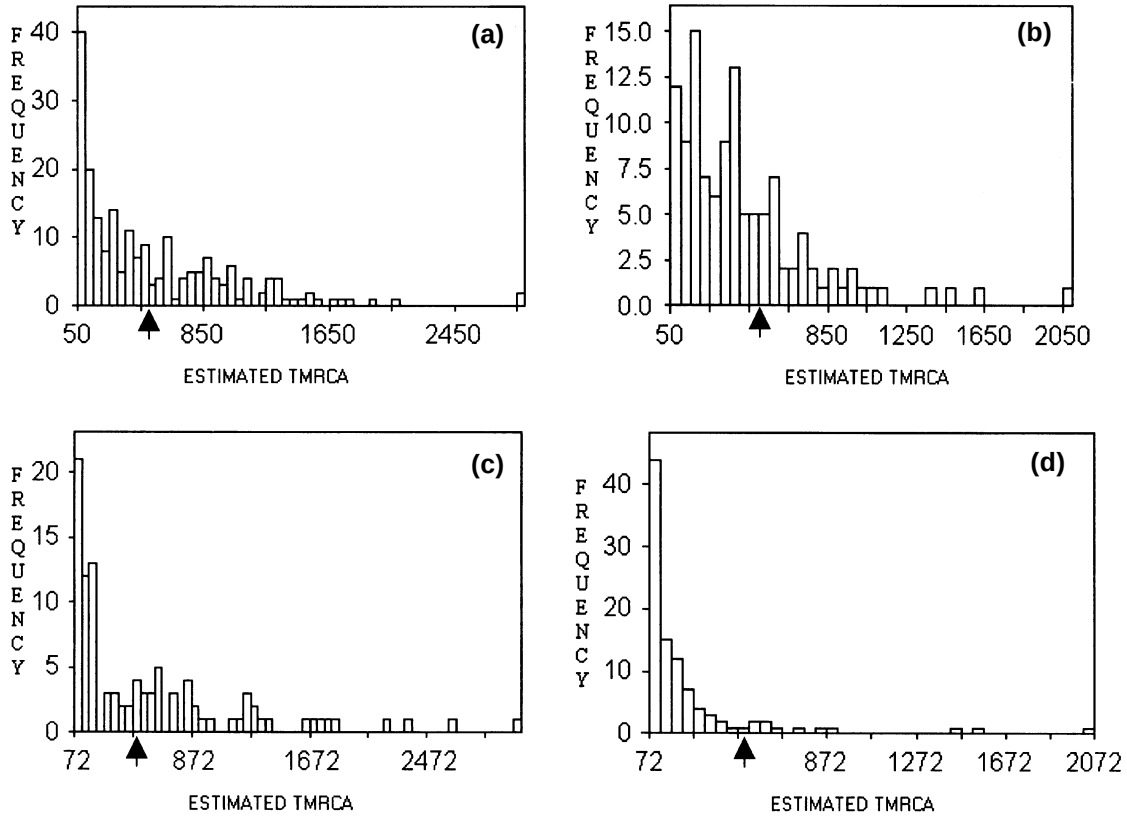

Figure 2. Frequency distributions of TMRCA estimated by two methods on simulated DNA sequence data, with simulation parameters $L = 200$ nucleotides, $m = 10^{-5}$ per site per generation and $g = 500$ generations (marked with an arrow on the X-axis), comprising 100 replications of each data set. (a) $a = 0$, Templeton's estimation method; (b) $a = 0.005$, Templeton's estimation method; (c) $a = 0$, MJN estimation method; (d) $a = 0.005$, MJN estimation method.

Results and discussion

Simulated data were generated using different combinations of the parameter values stated above. For each simulated data set, estimation of TMRCA was carried out using two different methods (Templeton 1993; Bandelt *et al*. 1995). TMRCA was estimated from a sample of $n = 100$ sequences. Since both estimation procedures crucially depend on the number of segregating sites, for a data set to be 'informative' the sample of sequences must have at least two segregating sites. We encountered noninformative data sets in our simulation runs, especially when g and m were both small. Our comparisons are all based on 100 informative data sets; that is, 100 data sets each of 100 sequences containing at least two segregating sites.

We first note that a large number of simulation runs was required to generate 100 informative data sets, because often the generated data set did not contain even two segregating sites. This number was particularly large when either g or m was small. For the MJN analysis, a further problem was encountered for an informative data set that had a single segregating site. For such a data set, while it was possible to calculate $\hat{k}$, it was not possible to draw the network (using the MJN software) and, therefore, to estimate TMRCA from the MJN. We had to discard such data sets from the MJN analysis. To keep the results comparable, however, we generated 100 informative data sets on which both methods of estimating TMRCA could be implemented.

Our results are summarized in table 1. It is evident from table 1 that the standard deviations of the TMRCA estimates were very large, irrespective of the parameter values used in the simulation. Generally, both methods underestimated the true TMRCA, except for short sequence lengths ($L = 200, 500$) and a short evolutionary time ($g = 250, 500$) with a low mutation rate ($m = 10^{-5}$). However,

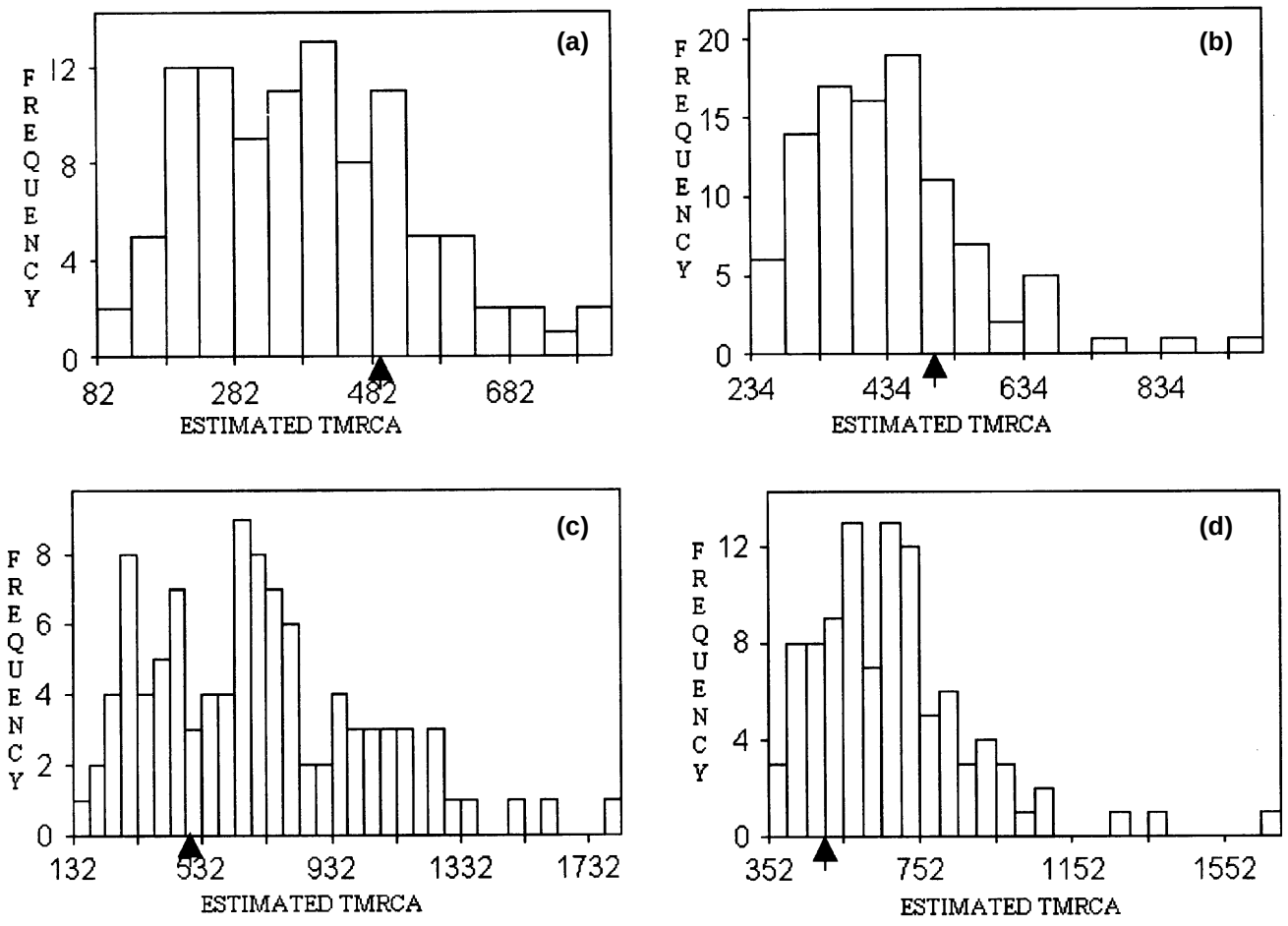

Figure 3. Frequency distributions of TMRCA estimated by two methods on simulated DNA sequence data, with simulation parameters $L = 400$ nucleotides, $m = 5 \times 10^{-5}$ per site per generation and $g = 500$ generations (marked with an arrow on the X-axis), comprising 100 replications of each data set. (a) $a = 0$, Templeton's estimation method; (b) $a = 0.005$, Templeton's estimation method; (c) $a = 0$, MJN estimation method; (d) $a = 0.005$, MJN estimation method.

the means of the estimated TMRCA values were not significantly different from the true values because of the large standard deviations. The correlation coefficient of the TMRCA estimates by the two methods is large for all sets of simulation parameter values. Thus, both methods seem to be unreliable to a similar degree in practice and it is difficult to choose between the two.

The means of the estimated TMRCA values for most combinations of simulation parameter values decreased as the mutation rates increased. This is because with a higher mutation rate there is a higher probability of parallel and back mutations, especially when the lengths of the sampled sequences are short. Both methods were rather insensitive to the population growth parameter (a), and there was no consistent trend with respect to a of either the mean values of the TMRCA estimates or the standard deviations, although in many cases the standard deviations decreased with increase in a. The frequency distributions of the TMRCA estimates (figures 2 and 3) were all highly positively skewed with a very long upper tail for both methods. Our results indicate that in practice considerable caution needs to be exercised in interpreting coalescence times estimated by either of these two popular methods.

References

Bandelt H.-J., Forster P., Sykes B. C. and Richards M. B. 1995 Mitochondrial portraits of human populations. *Genetics* 141, 743–753.

Donnelly P. and Kurtz T. G. 1997 The asymptotic behaviour of an urn model arising in population genetics. *Stoch. Proc. Appl.* **64**, 1–16.

Ewens W. J. 1979 *Mathematical population genetics*. Springer, New York.

Fu Y.-X. and Li W. H. 1997 Estimating the age of the common ancestor of a sample of DNA sequences. *Mol. Biol. Evol.* **14**, 195–199.

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

Hudson R. R. 1991 Gene genealogies and the coalescent process. In *Oxford surveys of evolutionary biology* (ed. D. Futuyma and J. Antonovics), vol. 7, pp. 1–44. Oxford University Press, Oxford.

Kingman J. F. C. 1982a On the genealogy of large populations. *J. Appl. Prob.* **19A**, 27–43.

Kingman J. F. C. 1982b The coalescent. *Stoch. Proc. Appl.* **13**, 235–248.

Mountain J. L., Hebert J. M., Bhattacharyya S., Underhill P. A., Ottolenghi C., Gadgil M. *et al.* 1995 Demographic history of India and mtDNA sequence diversity. *Am. J. Hum. Genet.* **56**, 979–992.

Nordborg M. 2001 Coalescent theory. In *Handbook of statistical genetics* (ed. D. Balding, M. Bishop and C. Cannings), pp. 179–212. Wiley, Chichester.

Saillard J., Forster P., Lynnerup N., Bandelt H.-J. and Nørby S. 2000 mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am. J. Hum. Genet.* **67**, 718–726.

Saunders I. W., Tavare S. and Watterson G. A. 1984 On the genealogy of nested subsamples from a haploid populaton. *Adv. Appl. Prob.* **16**, 471–491.

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

Tavare S., Balding D. J., Griffiths R. C. and Donnelly P. 1997 Inferring coalescence times from DNA sequence data. *Genetics* **145**, 505–518.

Templeton A. R. 1993 The “eve” hypothesis: a genetic critique and reanalysis. *Am. Anthropol.* **95**, 51–72.

Received 29 April 2003