Pairwise Comparisons of Mitochondrial DNA Sequences in Stable and Exponentially Growing Populations

Montgomery Slatkin* and Richard R. Hudson†

**Department of Integrative Biology, University of California, Berkeley, California 94720, and †Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92717*

Manuscript received February 1, 1991
Accepted for publication June 6, 1991

ABSTRACT

We consider the distribution of pairwise sequence differences of mitochondrial DNA or of other nonrecombining portions of the genome in a population that has been of constant size and in a population that has been growing in size exponentially for a long time. We show that, in a population of constant size, the sample distribution of pairwise differences will typically deviate substantially from the geometric distribution expected, because the history of coalescent events in a single sample of genes imposes a substantial correlation on pairwise differences. Consequently, a goodness-of-fit test of observed pairwise differences to the geometric distribution, which assumes that each pairwise comparison is independent, is not a valid test of the hypothesis that the genes were sampled from a panmictic population of constant size. In an exponentially growing population in which the product of the current population size and the growth rate is substantially larger than one, our analytical and simulation results show that most coalescent events occur relatively early and in a restricted range of times. Hence, the "gene tree" will be nearly a "star phylogeny" and the distribution of pairwise differences will be nearly a Poisson distribution. In that case, it is possible to estimate r, the population growth rate, if the mutation rate, μ, and current population size, N_0, are assumed known. The estimate of r is the solution to $r\bar{i}/\mu = \ln(N_0r) - \gamma$, where $\bar{i}$ is the average pairwise difference and $\gamma \approx 0.577$ is Euler's constant.

THE analysis of within-species variation in DNA sequences has the potential for providing insight into population genetic processes. New statistical methods are needed to analyze within-species sequence data, however, because DNA sequences provide new kinds of information about the genome.

In this paper, we point out some features of a commonly used way to describe within-species variation in DNA sequences, particularly of mitochondrial DNA (mtDNA). We will be concerned with two related questions: first, is it possible to use the sample distribution of pairwise differences in DNA sequence to test the hypothesis that the sequences were drawn from a panmictic population of constant size, and second, can the sample distribution of pairwise differences indicate that the genes sequenced were drawn from a population that has been growing exponentially in size for a long time? To answer these questions, we will review and develop the necessary analytic theory for pairs of genes and then present results obtained from a simulation program that yields the distribution of pairwise differences for samples of genes.

A typical data set consists of the sequences or fine scale restriction maps of mtDNA from several individuals. The numbers of differences in sequence between all pairs of individuals can be used to summarize information in the data (AVISE, BALL and ARNOLD 1988). It is also possible to estimate the times until each pair of mtDNA had a most recent common ancestor by using an estimate of the substitution rate per base pair. For mtDNA in animals, the rate of 0.01 substitutions per base pair per million years is usually used (BROWN, GEORGE and WILSON 1979; AVISE, BALL and ARNOLD 1988).

To illustrate this procedure we generated a sample data set using a simulation program described below. In Figure 1, we plot the frequencies of sample pairs that differ at i sites, $i \geq 0$. The conversion to divergence times would be obtained by multiplying i/L by 10^8 years where L is the number of base pairs in the sequence examined. This way of describing differences among sequences provides a convenient way to summarize some of the information in the data set.

CONSTANT POPULATION SIZE

Whether the graph of pairwise differences in Figure 1 is consistent with the hypothesis that the sample of mtDNAs is drawn from a panmictic population of constant size depends on what the null hypothesis predicts. WATTERSON (1975) and others have shown that under a neutral infinite-sites model with constant population size and no recombination among the sites, the distribution of the number of differences between

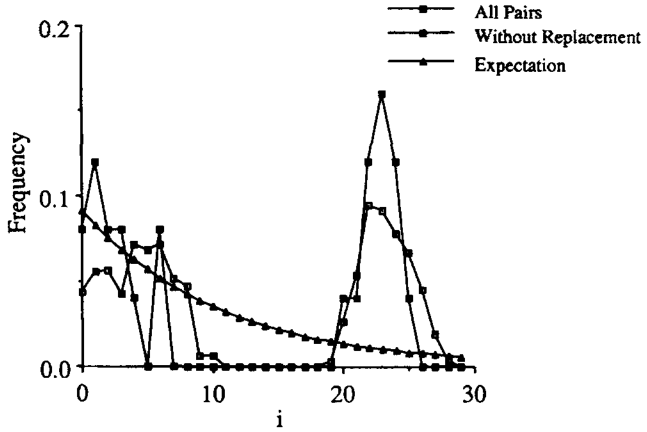

FIGURE 1.—Graphs of pairwise differences in 50 DNA sequences in a simulated data set compared with the geometric distribution under the model of constant population size. The "All pairs" curve shows the frequency distribution of the numbers of differences between all 1225 pairs; the "Without Replacement" curve shows the frequency distribution for the 25 pairs, 1 *vs.* 2, 3 *vs.* 4, etc.; and the "Expectation" curve plots Equation 1 for $\theta = 10$.

a pair of genes follows a geometric distribution:

$$Q(i) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^i, \tag{1}$$

where $\theta = 2N\mu$ with N being the size of the haploid population and μ being the mutation rate per generation. For mtDNA in higher animals, N is the effective size of the female population. The mean of i, $\bar{i}$, is θ, and the variance of i, σ_i^2, is $\theta(1 + \theta)$ (WATTERSON 1975).

This result is equivalent to an exponential distribution of divergence times of pairs of genes

$$R(t) = \frac{1}{N}e^{-t/N} \tag{2}$$

(TAJIMA 1983). The relationship between the distribution of pairwise differences and the distribution of coalescence times is obtained by noting that for a given coalescence time, the number of mutations that have occurred follows a Poisson distribution with mean $2\mu t$. We can find the mean and variance of i from the distribution of coalescence times, a procedure we will use later. Given that two genes have a coalescence time t, the mean number of differences is $\bar{i}(t) = 2\mu t$. The variance is also $2\mu t$ which implies that the mean square value is, $\overline{i^2}(t) = 2\mu t + 4\mu^2t^2$. We can then use (2) to find $\bar{t} = N$ and $\overline{t^2} = 2N^2$, from which we find, by averaging over t, the mean and variance of i to be θ and $\theta(1 + \theta)$ as before. We mention this now because in an exponentially growing population, we will derive the mean and variance of i directly from the analog of Equation 2.

The distribution given by Equation 1 would be useful for testing the null hypothesis if a large number of pairs of sequences were drawn, each pair from a different replicate population. That is not the kind of sample that is available, however, and we will show that (1) is not the distribution expected when pairs of genes in a single sample are compared. Differences between sample pairs from the same population are correlated because of their common history (BALL, NEIGEL and AVISE 1990). That history can be represented by a gene tree with each node indicating a coalescent event. Figure 2 shows the gene tree for a simulated data set, for which we know the extract tree.

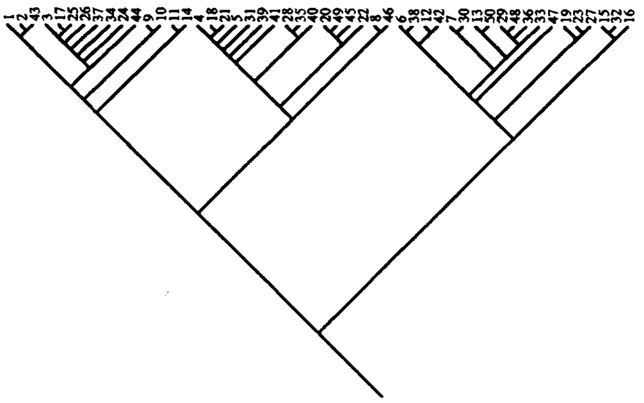

FIGURE 2.—The cladogram of a simulated data set. The correct cladogram is obtained using a parsimony criterion because the data were generated using the assumption that no site changed more than once. In this cladogram, the branch lengths do not represent time.

In Figure 1, we also plot Equation 1 for comparison with the data. The data were generated by simulating the null model with $\theta = 10$. We can see there is poor agreement between the data and the expectation under the null model. Figure 2 shows why. The cladogram is roughly balanced, meaning that there are approximately equal numbers of descendent genes on either side of the root. As a consequence, the number of base pair differences between genes from opposite sides of the root will all reflect the fact that they are separated by the maximum time possible. Because almost half of the pairwise comparisons are from genes on opposite sides of the root there are two modes in the distribution shown in Figure 1. The sample distribution appears to differ substantially from the exponential distribution even though they were generated by simulating the null model.

To determine what would be expected under the null hypothesis, we simulated 20 independent replicates of samples of genes from a single panmictic population and plotted graphs of pairwise differences, as in Figure 1. The simulation method was based on the coalescent process for a neutral infinite-sites model without recombination. Each sample is obtained by first producing the genealogy of the sample under the assumption of a large constant population size and no selection. Once the genealogy is produced, mutations are randomly placed on the genealogy. Assuming an infinite-site model, in which each mutation occurs at a previously unmutated site, the genotypes of the sampled sequences are then determined from the

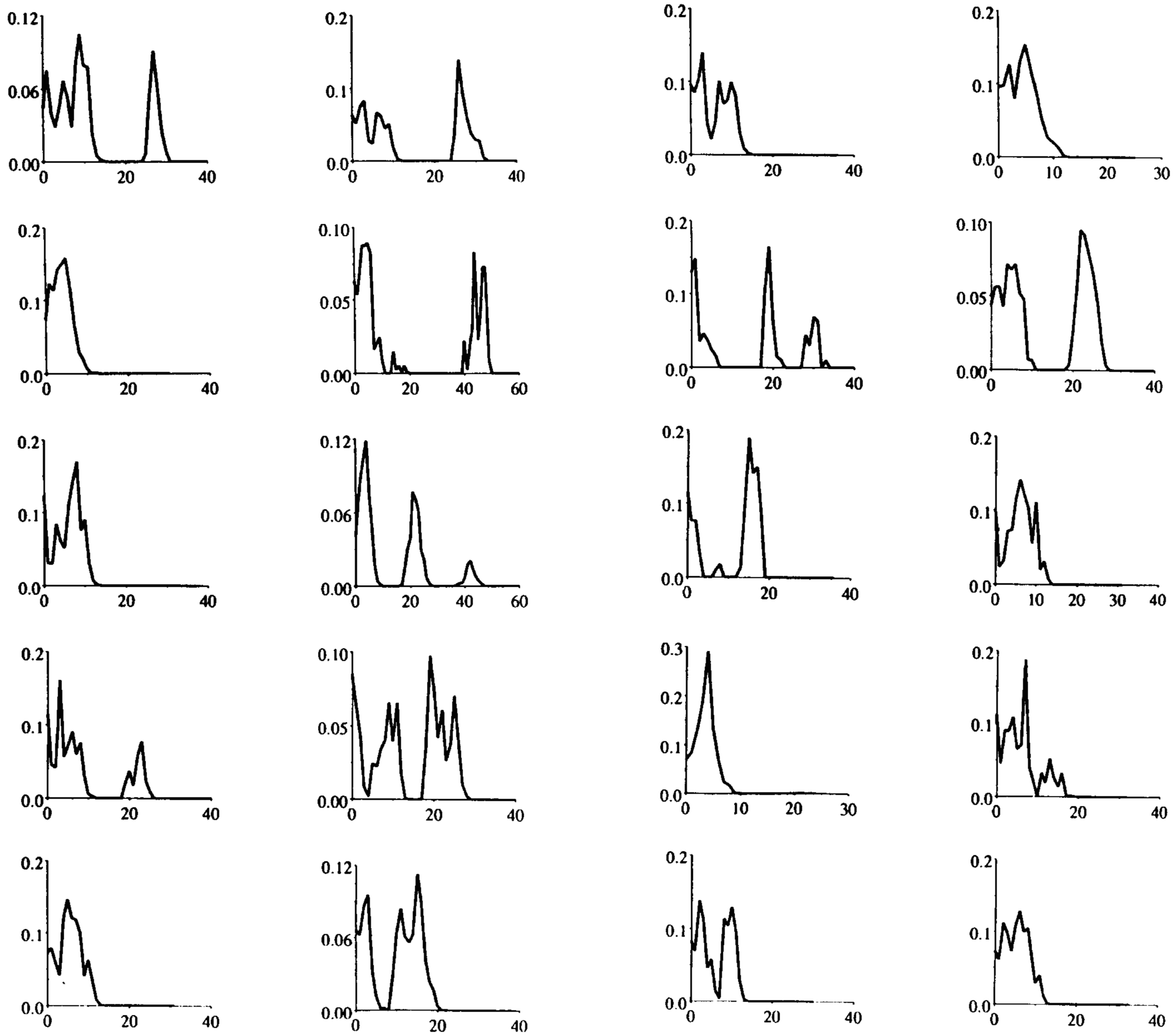

FIGURE 3.—Frequency distributions of pairwise differences for 20 replicate simulations. The distributions of all 1225 pairs in samples of 50 genes from a panmictic population are plotted. The data were generated using a simulation program described in the text. In each graph, the abscissa is the number of sites at which two samples differ and the ordinate is the fraction of pairs that differ.

genealogy with its mutations. The method is described in the appendix to HUDSON (1983), and described in greater detail in HUDSON (1990). We would be happy to distribute copies of the programs (written in *C*) that generated these results.

The simulation results are shown in Figure 3, for the case where $\theta = 2N\mu$ was 10. If we assume $\mu = 2 \times 10^{-4}$ (corresponding to a region 1000 nucleotides long with a per site neutral mutation rate of 0.01 per million years and 20-year generations) then $\theta = 10$ corresponds to $N = 25{,}000$. In Figure 3, we can see that a variety of shapes can be found, including bimodal and even trimodal distributions and distributions with modes at zero or at higher values. None of them resembles the geometric distribution shown in Figure 1. Even with this small sample of simulated results, we found a wide variety of distributions of pairwise differences are consistent with the predictions of the null hypothesis.

Bimodal distributions of sequence differences are reasonably common indicating that roughly balanced trees, such as the one shown in Figure 2, are common. TAJIMA (1983) showed that the number of genes on the right (or left) side of the root in a random gene tree generated by the coalescent process in a panmictic population follows a uniform distribution: that is, if there are n genes sampled, the probability of i genes on the left branch is $1/(n-1)$ for $i = 1, \ldots, n-1$.

EXPONENTIALLY GROWING POPULATION

It is likely that many species have undergone a sustained increase in population size, possibly because of a prior catastrophic decline in size or because a species is expanding its geographic range for the first time. For such species, a model of exponential growth at a constant rate is simple and reasonable. A model of exponential growth of human populations was suggested to us by J. BROOKFIELD (personal communication).

We examined the consequences of exponential population growth by considering first the theory of pairs of genes, for which relatively simple analytic results can be obtained, and a simulation model of the coalescent process for samples of n genes, as we did above. We will show that, in contrast to the model with constant population size, sample distributions of pairwise differences do provide useful information about the history of coalescent events and that under some conditions, the average pairwise difference, $\bar{i}$, leads to an estimate of the population growth rates.

Assume that the (haploid) population of interest is of current effective size N_0 and has been growing exponentially at a rate r. The population size at time t in the past is then $N(t) = N_0e^{-rt}$. Following the usual theory of coalescent processes (KINGMAN 1982), the probability that two genes sampled do not coalesce in generation t given that they did not coalesce before t is approximately $[1 - 1/N(t)]$ and the probability that they do coalesce in generation t is approximately $1/N(t)$. Therefore the probability that the first coalescence is in generation t, $P(t)$, is approximately

$$P(t) = \frac{1}{N(t)} \prod_{t'=0}^{t-1} \left[1 - \frac{1}{N(t')}\right] \qquad (3)$$

or approximately

$$P(t)dt = \frac{1}{N(t)} \exp\left[-\int_0^t \frac{1}{N(t')}\,dt'\right]dt, \qquad (4)$$

where now $P(t)dt$ is the probability of a coalescent event between t and $t + dt$. Assuming exponential growth, Equation 4 reduces to

$$P(t)dt = \frac{e^{rt}}{N_0} \exp\left(-\frac{e^{rt} - 1}{N_0r}\right)dt \qquad (5)$$

as was pointed out to us by J. BROOKFIELD (personal communication). This distribution is related to the Gompertz hazard distribution (JOHNSON and KOTZ 1970). It is convenient to simplify (5) by measuring time in units of $1/r$ ($\tau = rt$) and defining $\alpha = N_0r$ to obtain

$$P(\tau)d\tau = \frac{e^{\tau}}{\alpha} \exp\left(-\frac{e^{\tau} - 1}{\alpha}\right)d\tau. \qquad (6)$$

We can compute the mean coalescence time from (6):

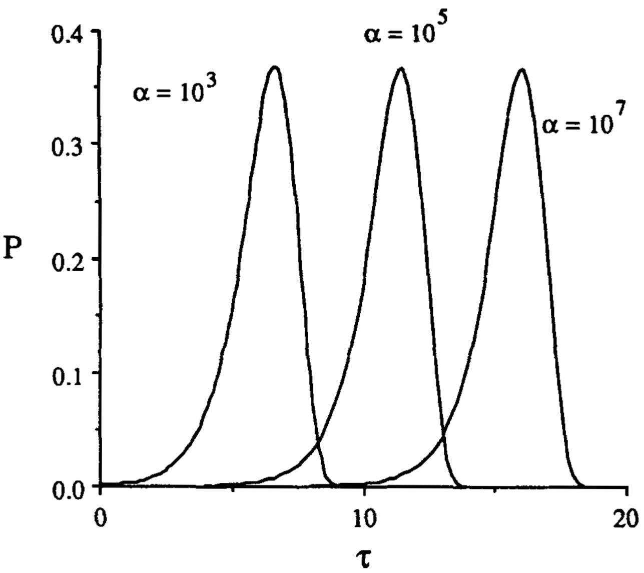

FIGURE 4.—The probability density of τ ($=rt$), the time to coalescence of two genes in an exponentially growing population. This density is given by Equation 6. ($\alpha = 2N_0r$, where r is the growth rate of the population per generation.)

$$\bar{\tau} = \int_0^\infty \tau \frac{e^{\tau}}{\alpha} \exp\left(-\frac{e^{\tau} - 1}{\alpha}\right)d\tau = \frac{e^{1/\alpha}}{\alpha} \int_1^\infty \ln(u)e^{-u/\alpha}du = -e^{1/\alpha}Ei\left(\frac{-1}{\alpha}\right), \qquad (7)$$

where $Ei(\cdot)$ is the exponential integral (GRADSHTEYN and RYZHIK 1965, §4.331.1). If $\alpha \gg 1$, then $-Ei(-1/\alpha) \approx \ln(\alpha) - \gamma$, where γ is Euler's constant (0.577 . . .) (GRADSHTEYN and RYZHIK 1965, §8.214.1), in which case $\bar{\tau} \approx \ln(\alpha) - \gamma = \ln(N_0r) - \gamma$. The distribution $P(t)$ is slightly asymmetric: by differentiating $P(\tau)$ and setting the result to zero, the modal value of τ is found to be $\ln(\alpha)$, which exceeds $\bar{\tau}$ by γ when $\alpha \gg 1$.

To find the variance, we have to evaluate

$$\overline{\tau^2} = \int_0^\infty \tau^2 \frac{e^{\tau}}{\alpha} \exp\left(-\frac{e^{\tau} - 1}{\alpha}\right)d\tau = \frac{e^{1/\alpha}}{\alpha} \int_1^\infty [\ln(u)]^2e^{-u/\alpha}du, \qquad (8)$$

which appears not to be expressable in terms of tabulated functions. It is relatively easy to compute either of the integrals in (8) numerically.

The function $P(\tau)$ is plotted in Figure 4 for different values of α. This figure conveys the important idea that if $\alpha \gg 1$, then coalescence events tend to occur in a restricted range of times, concentrated at $t = \ln(N_0r)/r$. This conclusion was reinforced by our calculations of the coefficients of variation (c.v.) of coalescence times, which were small (c.v. $\approx$ 0.25 for $\alpha = 10^3$) and decreased with increasing α (c.v. $\approx$ 0.09 for $\alpha = 10^8$). At $t = \ln(N_0r)/r$, the population size is approximately $1/r$ and is independent of N_0.

Our results for coalescence times can be expressed in terms of the numbers of pairwise differences between samples if we assume a constant mutation rate. As discussed above, the average number of differences given a coalescence time of t, $\bar{i}(t)$, is $2\mu t$, and the mean square number of differences given t is $\overline{i^2}(t)$ is $2\mu t + 4\mu^2 t^2$, because the distribution of the numbers of mutations is assumed to be Poisson. Averaging these values over t, we find $\bar{i} = 2\mu\bar{t}$ and the variance in i, σ_i^2 is $2\mu\bar{t} + 4\mu^2\overline{t^2} - 4\mu^2\bar{t}^2$. We can express σ_i^2 in terms of the mean and variance of t: $\sigma_i^2 = 2\mu\bar{t} + 4\mu^2\sigma_t^2 = \bar{i} + \overline{i^2}(\sigma_t/\bar{t})^2$. We contrast this result with that for a population of constant size for which $\sigma_i^2 = \bar{i} + \overline{i^2}$. In an exponentially growing population, the value of σ_i^2 differs from the variance under a Poisson distribution, $\bar{i}$, by a term that depends on the square of the coefficient of variation of t, which is small if $\alpha \gg 1$. Therefore, if $\alpha \gg 1$, the distribution of i is nearly Poisson which is the distribution of i if the phylogeny of genes sampled were a "star" phylogeny with all genes coalescing at the same time. Hence we conclude that in an exponentially growing population, the phylogeny of genes is likely to be nearly a star phylogeny, meaning that all coalescent events will occur near the root and few if any will occur later. In that case, we might guess that correlations induced by the phylogeny are relatively unimportant. Our simulations will support that guess.

As an illustration of our result, consider a naive model of human population growth. Assume $N_0 = 10^9$ and assume that 50,000 years ago, the size of the female population was 5000. Age structure will make the effective sizes smaller but we will assume by the same proportion at every time. If exponential growth had been occurring at a constant rate and the generation time is 20 years, then $r \approx 0.00488$ per generation. Under these assumptions, $\alpha = 4.88 \times 10^6$ and $\bar{t} \approx (\ln(\alpha) - \gamma)/r = 3030$ generations or approximately 60,600 years ago. The standard deviation of coalescent times for these parameter values is approximately 333 generations or 6,700 years.

The function $P(t)$ describes the distribution of the coalescence time for a single pair of genes sampled from an exponentially growing population. For reasons we discussed above, that does not provide us with the joint distribution of coalescence times between pairs of genes in a sample of n genes because that requires taking into account the correlation imposed by their common history. To examine the effects of this correlation we carried out the same kind of simulation that we did for the constant population size case. The simulations of the coalescent process with an exponentially growing population are very similar to those with a constant population size except that the distribution of the times between coalescent events are different. The required generalization of Equations 3–6 follows from the fact that the probability that the first coalescence among i lineages occurs in generation t is approximately

$$P(t) = \frac{\binom{i}{2}}{N(t)} \prod_{t'=0}^{t-1} \left[1 - \frac{\binom{i}{2}}{N(t')}\right].$$

Recall that as one traces the genealogy of the sampled sequences back in time, coalescent events occur and the number of linages that are being traced decreases by one for each coalescent event. The time interval, t_i, measured in units of $1/r$, during which there are i lineages can be generated by

$$t_i = \ln\left[1 + \alpha e^{-\tau_i} \frac{-2}{i(i+1)} \ln(U)\right]$$

where

$$\tau_i = \sum_{k=n}^{i+1} t_k$$

is the time of the coalescent event that reduced the number of lineages to i and U is a random variable uniformly distributed on the interval (0,1).

In ten replicate simulations, with $\alpha = 10^4$ and $\theta = 1.1 \times 10^4$, we found that the distribution of pairwise differences is unimodal and approximately Poisson in form. Two of the ten replicates are shown in Figure 5. As our analytic results suggest, a history of exponential growth tends to force coalescent events to occur in a relatively restricted range of times. As a consequence, correlations between coalescence times created by their history are relatively unimportant. Another consequence of having a "star" genealogy is that each mutation that occurs on the genealogy is likely to be inherited by only a single gene in the sample. In other words, the polymorphisms at individual nucleotide sites will consist of one mutant nucleotide and the rest of the sample will have the ancestral nucleotide at the site. A significant excess of this pattern of polymorphism is potentially detectable by the test of TAJIMA (1989) which is based on the total number of segregating sites and the average pairwise difference to the simulated data. In fact, for each of our ten replicates, TAJIMA's test indicated that the data were not consistent with the hypothesis that they were drawn from a randomly mating population of constant size.

To illustrate how similar the distributions of pairwise differences are, in our replicates, to a Poisson distribution, we show results from two replicates in Figure 5. The distribution in part A was chosen because its mean and variance were similar ($\bar{i} = 11.474$, $\sigma_i^2 = 10.254$) and because the distribution looked most like a Poisson. We can use the standard χ^2 statistic as a description of goodness of fit. For part A, $\chi^2 = 32.67$, which, if used in a statistical test, would indicate

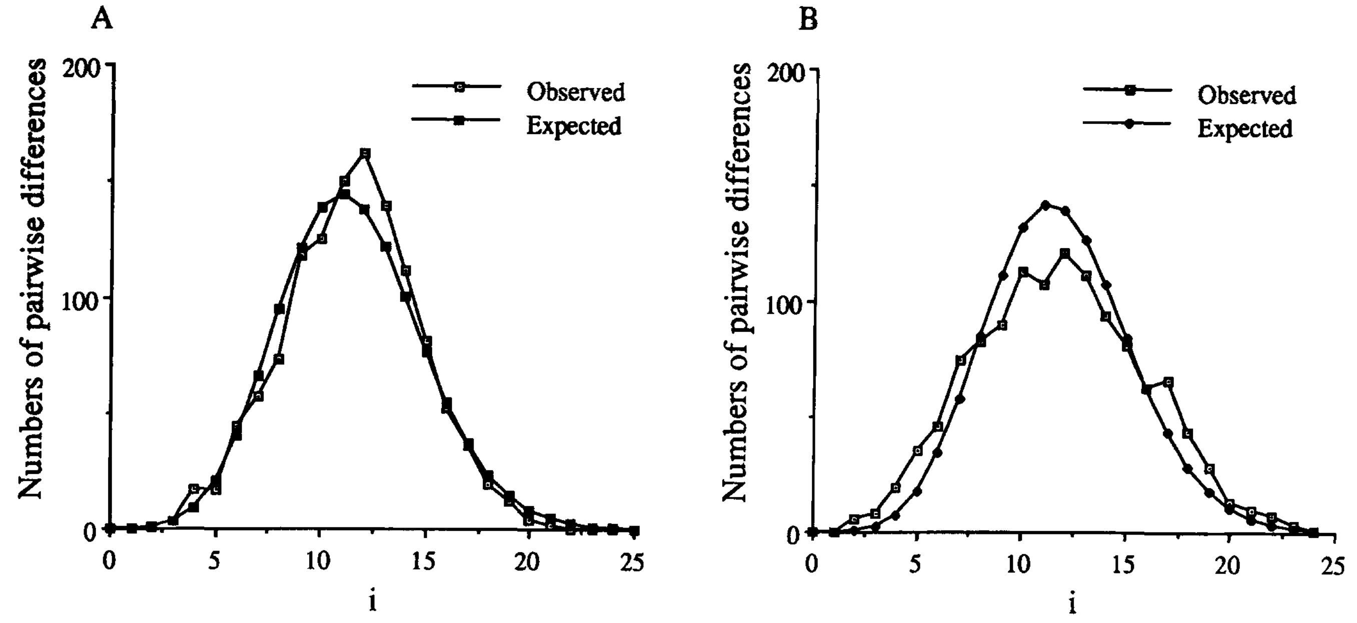

FIGURE 5.—Comparison of the observed numbers of pairwise differences in two simulations with the numbers expected under a Poisson distribution with the same mean. The two data sets are two of ten replicate samples generated as described in the text. In these simulations, $\alpha = 10^4$ and $\theta = 1.1 \times 10^4$. In part A, $\bar{i} = 11.474$ and $\sigma_i^2 = 10.254$ and in part B, $\bar{i} = 11.814$ and $\sigma_i^2 = 16.331$. Both distributions differ significantly from a Poisson, for part A, $\chi^2_{17} = 32.67$ ($P < 0.025$) and for part B, $\chi^2_{19} = 135.1$ ($P < 0.005$).

a marginally significant deviation (17 d.f.; $P < 0.025$). The distribution in part B was chosen because it appeared to differ substantially from a Poisson ($\bar{i} = 11.814$, $\sigma_i^2 = 16.331$) and indeed the differences between the observed and expected distributions are much greater ($\chi^2 = 135.1$; 19 d.f.; $P < 0.005$). These P values are meaningful only as a measure of fit to a Poisson because the pairwise differences are of course not independent. Even for the distribution shown in Figure 5B, however, the observed distribution does not appear to be very different from a Poisson.

Another way to interpret these results is to note that in Figure 5, none of the pairwise differences were zero, which means that there were no short branches in the gene tree. In the other replicates, less than one in one thousand pairwise comparisons had zero differences. In contrast, there were always large fractions of the pairwise comparisons with zero differences for the model of constant population size, as shown in Figure 3.

We conclude then that if the observed distribution of pairwise differences is close to a Poisson, that it is consistent with the hypothesis that the population from which those genes were sampled has been growing exponentially in size.

Estimating population growth rates: Our results for the model of exponential population growth suggest that it is possible to estimate the population growth rate, r, under some conditions. In particular, if the distribution of pairwise differences were similar to a Poisson distribution, that would indicate that the gene tree is nearly a star phylogeny which we have shown is consistent with a model of exponential growth with $\alpha = N_0 r \gg 1$.

To estimate r, we have to assume that N_0 and μ are known. Then we use the approximate estimate of the mean pairwise coalescent time, $\bar{t} \approx (\ln(N_0 r) - \gamma)/r$, to obtain the estimate of the mean pairwise difference,

$$\bar{i} = 2\mu[\ln(N_0 r) - \gamma]/r. \qquad (9)$$

If the value of $\bar{i}$ is estimated from the data, then Equation 9 can be solved numerically for r. To illustrate this result, we again use hypothetical data. Assume that 2000 base pairs have been sequenced and that $\bar{i} = 10.5$. Assume that the population from which the sample was taken has an effective size of the female population of $N_0 = 10^6$. If the mutation rate per site per year is 10^{-8}, the mutation rate per generation is 2×10^{-7}, when the generation time is 20 years, and the value of μ in (9) is $2000 \times 2 \times 10^{-7} = 4 \times 10^{-4}$. Using a program that solves (9) numerically, we find that r is approximately 4.15×10^{-4}. We will distribute a copy of this program upon request.

In making such an estimate of r, it is important to realize that an approximately Poisson distribution of pairwise differences does not imply that there has been exponential growth of the population at a constant rate. There are other possible explanations as well. A very rapid increase in population size followed by a period of large and constant population size would also result in a starlike gene tree because all coalescent events would occur relatively quickly be-

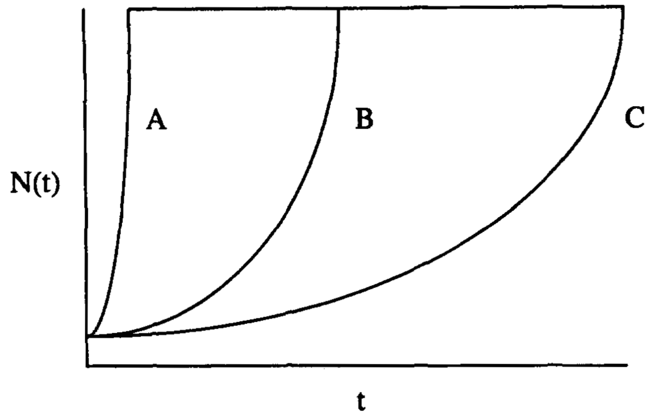

FIGURE 6.—An illustration of different population growth trajectories that could lead to a nearly starlike gene tree and hence an approximately Poisson distribution of pairwise differences in sequence. Consequently it is not possible to use the observation of an approximately Poisson distribution of pairwise differences to conclude that there was exponential growth during the history of the population sampled.

fore the time of rapid increase. Hence a distribution of pairwise differences that is nearly Poisson would result. If that assumption about population growth is accepted, the value of $\hat{\imath}$ can be used to estimate the time of the sudden increase in population size: $t = \hat{\imath}/(2\mu)$. Using the numbers in the preceding paragraph, a value of $\hat{\imath}$ of 10.5 when 2000 base pairs are sequenced is consistent with a time of very rapid increase in population size of $10.5/(8 \times 10^{-4}) = 13{,}125$ generations or 262,500 years ago. Figure 6 illustrates different growth trajectories that would all lead to nearly a Poisson distribution of pairwise differences.

Yet another possibility is natural selection in favor of one mitochondrial genotype over previously existing ones. Such selection would result in a rapid increase in the number of individuals carrying the favored mitochondrial type. As KAPLAN, HUDSON and LANGLEY (1989) have shown, the fixation of an advantageous gene in the recent past can result in the coalescence of most lineages near the time of fixation. For a population genetics perspective, there is no difference between a rapid increase in population size and a rapid increase in the size of the population carrying the only mitochondrial type to leave descendents. We simulated this possibility as well and found that distributions of pairwise differences are very similar to those found for an exponentially growing population. Results of those simulations are available on request.

DISCUSSION

Our analysis has been motivated by repeated observations that observed distributions of pairwise differences in DNA sequences in samples of mitochondrial DNAs differ substantially from the geometric distribution expected in populations that have remained constant in size. AVISE, BALL and ARNOLD (1988) summarized data for hardhead catfish, American eels and redwing blackbirds and found that the distributions of pairwise differences differed substantially from expectations based on rough estimates of effective population sizes obtained from censuses (AVISE, BALL and ARNOLD 1988, Figures 2, 3, and 4). They concluded that effective population sizes were in fact much smaller than current census sizes, suggesting that past bottlenecks in population size had occurred. The distribution of pairwise differences for redwings does have a unimodal distribution of a form similar to those in Figure 5.

Distributions of pairwise differences that are similar to a Poisson distribution are also found for human data. CANN, STONEKING and WILSON (1987, Figure 1) show a unimodal distribution of pairwise differences detected using a battery of restriction enzymes among 146 mtDNAs from individuals in five races. The mean difference was found to be approximately 0.57%. DIRIENZO and WILSON (1991) found similar patterns within some human populations but not in others. They plotted the distribution of pairwise differences of 6 populations (Sardinians, Middle Easterners, Japanese, American Indians, !Kung, and Pygmies) (DIRIENZO and WILSON 1991, Figure 3). The !Kung and Pygmies samples are clearly not similar to a Poisson but the others are. As we have emphasized, this similarity does not ensure that there has been exponential growth of these populations in the recent past but it does indicate that demographic events in the past have forced coalescent events into a narrow time window. DIRIENZO and WILSON (1991) applied TAJIMA's (1989) test of neutrality to the Sardinian and Middle Eastern samples. They found that the Sardinian samples (sample size 69) were not consistent with the neutral hypothesis but the Middle Eastern samples (sample size 42) were consistent.

CONCLUSIONS

We conclude that plotting frequency distribution of pairwise differences in sequence or equivalently pairwise divergence times of genes sampled provides an indication of the structure of the phylogenetic tree representing the history of those genes. It is difficult, however, to compare such a graph with a geometric distribution and reject the null hypothesis that the genes sampled were from a randomly mating population of constant size. The information in this kind of data is probably better extracted in other ways. J. FELSENSTEIN (personal communication) has suggested one test of constancy of effective population size.

Our results for an exponentially growing population suggest that the distribution of pairwise differences can provide useful information if the distribution is nearly a Poisson distribution. In that case there is a star-like gene tree with all the nodes clustered in time. That pattern would also be detected in the gene tree directly if branch lengths were known. It is possible then to use our analytic results to estimate the population growth rate under the assumption that the population has been growing exponentially for a long time. The observation of a nearly Poisson distribution of pairwise differences does not however imply that there has been exponential growth. That distribution would be consistent with other models of population growth that force most of the coalescent events into a narrow time period.

We thank J. C. AVISE, A. DIRIENZO, J. FELSENSTEIN, A. C. WILSON and two anonymous referees for helpful discussions of this topic and for useful comments on earlier versions of this paper. This research has been supported in part by National Institutes of Health grants GM40282 to M. S. and GM42447 to R. R. H.

LITERATURE CITED

AVISE, J. C., R. M. BALL and J. ARNOLD, 1988 Current versus historial 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.

BALL, R. M., J. E. NEIGEL and J. C. AVISE, 1990 Gene genealogies within the organismal pedigrees of random-mating populations. Evolution **44:** 360–370.

BROWN, W. M., M. GEORGE, JR. and A. C. WILSON, 1979 Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci. USA **76:** 1967–1971.

CANN, R. L., M. STONEKING and A. C. WILSON, 1987 Mitochondrial DNA and human evolution. Nature **325:** 31–36.

DIRIENZO, A., and A. C. WILSON, 1991 The pattern of mitochondrial DNA variation is consistent with an early expansion of the human population. Proc. Natl. Acad. Sci. USA **88:** 1597–1601.

GRADSHTEYN, I. S., and I. W. RYZHIK, 1965 *Tables of Integrals, Series and Products.* Academic Press, New York.

HUDSON, R. R., 1983 Testing the constant-rate neutral model with protein sequence data. Evolution **37:** 203–217.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process. Oxf. Surv. Evol. Biol. **7:** 1–44.

JOHNSON, N. L., and S. KOTZ, 1970 *Continuous Univariate Distributions.* Houghton & Mifflin, New York.

KAPLAN, N., R. R. HUDSON and C. H. LANGLEY, 1989 The "hitchhiking effect" revisited. Genetics **123:** 887–899.

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

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.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Communicating editor: A. G. CLARK