Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method

JOSEPH FELSENSTEIN

Department of Genetics SK-50, University of Washington, Seattle, Washington 98195

Electronic mail address: joe@genetics.washington.edu

Summary

We would like to use maximum likelihood to estimate parameters such as the effective population size N_e, or, if we do not know mutation rates, the product $4N_e\mu$ of mutation rate per site and effective population size. To compute the likelihood for a sample of unrecombined nucleotide sequences taken from a random-mating population it is necessary to sum over all genealogies that could have led to the sequences, computing for each one the probability that it would have yielded the sequences, and weighting each one by its prior probability. The genealogies vary in tree topology and in branch lengths. Although the likelihood and the prior are straightforward to compute, the summation over all genealogies seems at first sight hopelessly difficult. This paper reports that it is possible to carry out a Monte Carlo integration to evaluate the likelihoods approximately. The method uses bootstrap sampling of sites to create data sets for each of which a maximum likelihood tree is estimated. The resulting trees are assumed to be sampled from a distribution whose height is proportional to the likelihood

surface for the full data. That it will be so is dependent on a theorem which is not proven, but seems likely to be true if the sequences are not short. One can use the resulting estimated likelihood curve to make a maximum likelihood estimate of the parameter of interest. This will be the effective population size N_e or, if the neutral mutation rate μ is not known, of $4N_e\mu$. The method requires at least 100 times the computational effort required for estimation of a phylogeny by maximum likelihood, but is practical on today's workstations. The method does not at present have any way of dealing with recombination.

1 Introduction

Given homologous nucleotide sequences sampled randomly from a population, how could we make estimates of parameters of evolutionary interest? If the sequences have diverged by neutral mutation, without recombination, from an ancestral gene according to a genealogy determined by genetic drift in an isolated population of constant effective size, there are only two parameters controlling the processes, the effective population number N_e and the neutral mutation rate per generation, μ. It will turn out that we can only estimate these by estimating their product, the most convenient parameter being

$$\theta = 4N_e\mu. \tag{1}$$

The method used here will be maximum likelihood. If a random sample of n nucleotide sequences is called S, the likelihood will be the probability of S given N_e and u,

$$L = Prob(S|N_e, \mu). \tag{2}$$

We can break up (2) into terms corresponding to the different possible genealogies G' on which evolution could have led to the observed sequences. As we assume that there is no recombination, each gene is descended from a single ancestral copy, and the genealogy of the genes is a branching tree with branch lengths that are scaled in generations. For each value of G' the contribution to (2) consists of the probability $P(G'|N_e)$ that genetic drift would have led to this genealogy G' of the sequences, times the probability $P(S|G', \mu)$ that these particular sequences would arise on that genealogy, given the neutral mutation rate. Therefore

$$L = \sum_{G'} Prob(G'|N_e) Prob(S|G', \mu) \tag{3}$$

This equation, given also in Felsenstein (1988), is in a sense the fundamental formula for likelihoods of sequence samples under neutral mutation in the absence of recombination. The two probabilities are each fairly straightforward to compute. The difficulty comes with the summation which must run over all possible genealogies, of which there are a vast number. In this paper I present a resampling method of carrying out this summation approximately, using maximum likelihood estimates of G' on bootstrap-resampled data sets to choose points for Monte Carlo integration.

Strobeck (1983) has previously considered the same problem for the case of the "infinite sites model" (Watterson, 1975) where each mutation occurs at a different site, so that no reversals or parallelisms occur. In this case, which is in effect the limit as $\theta \to 0$, the computations are not as daunting, as only a limited number of genealogies are compatible with the data. He gave recursion formulae for the likelihood in some cases with small numbers of different sequences observed. He did not give general formulas. Griffiths

(1989) has shown how to do the computations of the likelihood for the infinite sites model for any number of sequences, in the case where it is known which state is the ancestral one for each site. This is usually not known. He has made available a computer program to calculate likelihoods for that case. The present work proposes a method that can cope with all values of θ, and cases when there may be ambiguities about the genealogy.

One alternative to the present computation is the use of pairwise methods, which make a separate estimate of θ for each pair of sequences in a sample, and average these. Nei and Tajima (1981) have presented methods for making such an estimate. They have been used by Avise (1987). Watterson's (1975) estimation method, based on the number of segregating sites in a sample, is another alternative. Both of these are computationally far faster than the methods proposed here, but I have shown (Felsenstein, 1992) that in the limiting case of long sequences they also make far less efficient use of the data. It is not known to what extent this inefficiency will appear with sequences of finite length, but it must be applicable to sufficiently long sequences, and thus it seems at least worthwhile to try to find a way of making a maximum likelihood estimate of θ.

2 The coalescent prior

The quantity $Prob(G'|N_e)$ is the prior probability of the genealogy G' under genetic drift. We assume that there are no hitchhiking effects of selection at nearby loci. If there are, they will distort this probability, often making the observed sequences on average more closely related than would be the case under pure genetic drift.

Kingman (1982a, 1982b) has shown that, under a classical Wright-Fisher model of the reproduction of an idealized population of constant size N_e, where N_e is not small, the genealogy of the population is extremely well-approximated by the process he called the *n-coalescent*. This generates a genealogy by starting with n sequences, and successively joining lineages, going backwards in time. There are $n-1$ such coalescences. For each one, we pick two of the extant lineages at random to be the next two combined. The additional time to the next coalescence from the one which leaves us with i sequences is (in generations) drawn from an exponential distribution:

$$v_i \sim Exponential\ \left(\frac{4N_e}{i(i-1)}\right) \tag{4}$$

Note that v_i is the time between coalescence $i+1$ and coalescence i, not the total time from the present to the i-th coalescence. The first coalescence going backwards is number n and the coalescences further back in time are numbered $n-1$, $n-2$, ..., 2. The random tree constructed by this process is an outcome of the n-coalescent, and approximates a random genealogy of n genes produced by genetic drift.

These genealogies have a time scale of generations. We will find it more useful to express them in terms of a time scale which has one unit per $1/\mu$ generations (we shall call this genealogy G rather than G'), so that the rate of neutral mutation is one per unit time. This is helpful because our observations of time are all in terms of the time necessary to produce a given amount of sequence divergence. On this scale one generation is only μ units of the rescaled time so that the v_i in (3) are replaced by the rescaled variables

$$u_i \sim Exponential\ \left(\frac{\theta}{i(i-1)}\right). \tag{5}$$

Given the intervals u_i between coalescences, the probability density of the prior distribution of genealogies is a product of exponential densities, times the probability of the tree topology, giving:

$$\begin{aligned} Prob(G|\theta) &= P_T \prod_{i=2}^{n} \frac{i(i-1)}{\theta} \exp\left(\frac{i(i-1)u_i}{\theta}\right) \\ &= P_T \frac{n!(n-1)!}{\theta^{n-1}} \exp\left(\frac{1}{\theta}\sum_{i=2}^{n} i(i-1)u_i\right) \end{aligned} \tag{6}$$

where P_T is the probability of the particular pairs of lineages which were chosen to coalesce. At the i-th coalescence the probability of the particular pair of lineages which were chosen is one out of the number of possible pairs, $i(i-1)/2$. The product of these probabilities is

$$P_T = \prod_{i=2}^{n} \left(1 \Big/ \left(\frac{i(i-1)}{2}\right)\right) = \frac{2^{n-1}}{n!(n-1)!} \tag{7}$$

and when this is substituted into (6) we get

$$Prob(G|\theta) = \left(\frac{2}{\theta}\right)^{n-1} \exp\left(\frac{1}{\theta}\sum_{i=1}^{n} i(i-1)u_i\right) \tag{8}$$

3 Monte Carlo integration

The second probability in (2) is also readily calculated: when the time scale of G is taken to be in units of $1/\mu$ generations, as here, the probability of the sequences S given G no longer depends on μ. The quantity becomes $Prob(S|G)$ and is the likelihood of G for the data S. This can be calculated by standard methods for computing the likelihood of a set of sequences on a tree G (Felsenstein, 1981).

The genealogical trees G live in a space which has discrete tree topologies with ordered interior nodes ("labelled histories"), each of which has $n-1$

parameters, the u_i. A labelled history is a tree topology, with the additional specification of the ordering in time of the nodes. There are $n!(n-1)!/2^{n-1}$ of these (Edwards, 1970). For example for $n = 10$ there are 2,571,912,000, for $n = 20$ there are 6.96×10^{18} and for $n = 100$ there are 1.37×10^{284}.

For each of these tree topologies, there is an $(n-1)$-dimensional space of genealogies, corresponding to the possible values of the u_i, each of which varies from 0 to ∞. To evaluate the sum in equation (3) amounts to calculating (for $n = 10$), over 2×10^9 9-dimensional integrals. I have not been able to find explicit solutions for any of these individual integrals. Lacking a dramatic algebraic breakthrough, the straightforward numerical approximation would be to add up values of the summand in (3) over an $(n-1)$-dimensional grid of points. The high dimensionality of the individual integrals, and the large number of them, makes this impractical.

An alternative that is often used in such cases is Monte Carlo integration (Kahn, 1950; Hammersley and Handscomb, 1964). In Monte Carlo integration random points are drawn from the domain of the function, the function evaluated at each of these, and the sum used to approximate the average height of the function over the domain. This technique has its limitations. It needs a large sample size to accurately approximate the integral, and if the height of the function varies greatly over the domain the variance of the approximation can be large. This will almost certainly be true in the present case. Consider the function being integrated in the favorable case when the sequences are very long. In that case the sequences will define the genealogy G unambiguously, that is, concentrate the area under the likelihood curve (2) almost entirely within one of the $n!(n-1)!/2^{n-1}$ tree topologies. As this is a very large number, when genealogical trees are sampled at random

most of them will contribute very little to the integral (3) so that most of the sampling effort is wasted.

The efficiency of Monte Carlo integration can be greatly improved by the use of *importance functions* (Kahn, 1950; Hammersley and Handscomb, 1964). These are distribution functions whose density is concentrated in the regions which are expected to contribute most to the integral. If we sample genealogies from a density function $g(G)$, and if the function being integrated is $f(G)$, the approximation to the integral weights the function $f(x)$ for each sampled point x by the size of the interval which it represents. This will be inversely proportional to the density $g(x)$. In fact, if m points are sampled, the weight assigned to point x will be $1/(mg(x))$. The approximation to the integral of $f(x)$ will be:

$$\int f(x)dx \simeq \sum_{i=1}^{m} \frac{f(x_i)}{mg(x_i)} \tag{9}$$

If an appropriate importance function is found, the sampling variance of the numerical integral can be greatly reduced.

4 The bootstrap sampler

In order to find an appropriate importance function, let us look at the function (3) being integrated. It has two factors, the prior and the likelihood. Of these the prior is the simpler, and its density at all points in the space of genealogies is easily calculated. But it is a poor candidate for an importance function. It assigns equal weight to all possible labelled histories, and we have seen that in the cases of long sequences this will fail to concentrate sampling on the trees that actually contribute to the integral.

The obvious alternative would be the other term, the likelihood of the tree $P(S|G)$. This is certainly concentrated in the regions of interest, but has the disadvantage that it is not obvious how to sample from a distribution whose density is proportional to this likelihood.

I suggest that there is a procedure that samples from a distribution whose density is nearly proportional to the likelihood function $P(S|G)$, the accuracy of the approximation increasing with sequence length. This is to draw a bootstrap sample (a sample with replacement) S^* of the p sites in the sequences, and to find for this sample the genealogy $\hat{G}^*$ that maximizes the likelihood for this bootstrap sample:

$$P(S^*|\hat{G}^*) = \max_G P(S^*|G) \tag{10}$$

Bootstrap sampling (Efron, 1979) is a resampling method that involves drawing from the original sample, with replacement, a sample of the same size. It produces a sample whose estimate of a parameter has approximately the same distribution as the true distribution of the parameter. The proportionality of this distribution to the likelihood curve for the parameter is closely related to this property. The boootrap has been applied to phylogenies (Felsenstein, 1985) where the bootstrap sample involves sampling sites. If the data table has sequences as rows, the bootstrap sampling samples whole columns to make up a new table of the same length as the original data, without altering the order of sequences within each column when it is copied.

The method of this paper depends on the assertion that $\hat{G}^*$ is drawn from a distribution whose density is proportional to the likelihood $P(S|G)$.

If this were the case, so that when $\hat{G}^*$ is drawn this way,

$$g(G) = cP(S|G) \tag{11}$$

where c is the constant needed to make $g(G)$ integrate to 1. The formula for computing the approximation to the integral from m such bootstrap samples is from (9)

$$\begin{aligned} \int_G P(G|\theta)P(S|G) &\simeq \sum_{i=1}^{m} P\left(\hat{G}_i^*|\theta\right) P\left(S|\hat{G}_i^*\right) / \left(mcP\left(S|\hat{G}_i^*\right)\right) \\ &\simeq \sum_{i=1}^{m} P\left(\hat{G}_i^*|\theta\right) /(mc) \end{aligned} \tag{12}$$

where $\hat{G}_i^*$ is the maximum likelihood genealogy for the i-th bootstrap sample drawn. Note that neither the constant c nor the number of bootstrap replicates m depends on the unknown θ. Thus (12) can be used to compute mc times the desired integral:

$$mc \int_G P(G|\theta)P(S|G) \simeq \sum_{i=1}^{m} P(\hat{G}_i^*|\theta). \tag{13}$$

Once we have computed the sum on the right-hand-side of (13) it gives us the likelihood up to the unknown constant mc. We can use that curve to find an approximate maximum likelihood estimate of θ, the value that maximizes this sum, so that:

$$\sum_{i=1}^{m} P\left(\hat{G}_i^*|\hat{\theta}\right) = \max_{\theta} \sum_{i=1}^{m} P\left(\hat{G}_i^*|\theta\right). \tag{14}$$

Note that we need not compute maximum likelihood genealogies $\hat{G}_i^*$ separately for each value of θ that we evaluate. We can compute one set of estimated genealogies $\hat{G}_i^*$ and, as evaluation of the prior $P(\hat{G}_i^*|\hat{\theta})$ is rapid, we can then trace out (13) which is the sum of m curves, each a function of θ.

5 A theorem-free assertion

What assurance do we have that the maximum likelihood genealogies computed from the bootstrap samples do in fact have the required approximate density? There is at present no general theorem guaranteeing this. We may at best regard the statement as a Theorem-Free Assertion. The most I will do here is to establish that the result is a plausible one for sequences that are not too short. It needs to be investigated whether the theorem is actually true under the regularity conditions that apply to maximum likelihood estimates of genealogies from nucleotide sequences.

(a) The limit with long sequences

It should be apparent that, when the sequences are very long, the estimates of the genealogy from bootstrap samples will be concentrated near the true genealogy, given that the probabilistic model of change of the sequences model is correct. This is as it should be, since we hope to concentrate the integration on the true tree topology, which will contribute almost all of the likelihood.

Asymptotically, as the number p of sites becomes large, the likelihood curve as a function of the genealogy will more and more closely approach in shape a multivariate Gaussian density whose mean is the true genealogy. The parameters of the genealogy will be the node times (on the rescaled time scale) of the genealogy, the topology not being at issue. The covariance matrix of the estimates of these parameters will be the inverse of the matrix of curvatures of the expected likelihood at the true genealogy.

If we bootstrap the sequences and estimate genealogies from each boot-

strap sample, in this case the tree topologies will almost always be the same as the true tree. The variances and covariances of the branch length parameters will approach the true values, being asymptotically $(p-1)/p$ of the true values as $p \to \infty$. It can also be argued that the asymptotic distribution of the branch lengths is the same multivariate normal. An informal sketch of the argument is given by Efron (1982, pp. 34-35) who says that "If $Q(.,.)$ is a well-behaved function, as described in Efron (1979a, Remark G), then the bootstrap distribution of Q^* is asymptotically the same as the true distribution of Q."

In the present case we also want to argue that this distribution is also asymptotically the same as a rescaled version of the likelihood function. That the likelihood function is asymptotically proportional to the same multivariate normal distribution which is the distribution of the branch length parameters (and hence to the distribution of bootstrap estimates of those parameters as well) is well-known (see, for example, Kendall and Stewart (1973, p. 240) who cite a proof by Wald). Thus the bootstrap will approximate well the asymptotic normal distribution of genealogies.

(b) The case of two sequences

We can investigate whether the result is confined to cases where asymptotic behavior guarantees normality by looking more carefully at the simplest case, which is that of $n = 2$. In that case there is only one possible genealogy, a simple two-sequence rooted tree, whose only parameter is the time t of the common ancestor of the sequences, rescaled in units of nucleotide substitutions per site. The coalescent prior is the simple exponential density

for t:

$$t \sim \text{Exponential } (\theta/2). \tag{15}$$

If we assume, for purposes of the example, that the sites are evolving according to a simple Jukes-Cantor (1969) model, then the likelihood for a pair of sequences of length p will be

$$L = P^d(1-P)^{p-d} \tag{16}$$

where d is the number of sites differing between the two sequences, and P is the probability that a site will differ between the two sequences, which is

$$P = \frac{3}{4}\left(1 - e^{-\frac{8}{3}t}\right). \tag{17}$$

This differs from the usual Jukes-Cantor expression only because the length of branches separating the two species is $2t$ rather than t.

The likelihood function for t for a given value of d will be given by substituting (17) into (16), obtaining

$$L(t,d) = \frac{3^d}{4^p}\left(1 - e^{-\frac{8}{3}t}\right)^d \left(1 + e^{-\frac{8}{3}t}\right)^{p-d} \tag{18}$$

The central issue is whether the likelihood curve (18) is a good approximation to the density function of estimates of t. That it cannot be a perfect approximation is seen by considering $t = \infty$. That is the estimate that will be made whenever $m > (3/4)p$, which will happen with finite probability (less often, the larger is p). When t is large L will approach the value $3^d/4^p$. This is tiny but finite and nonzero so there should at that value be an infinite area near $t = \infty$ under the density that is proportional to $L(t)$. But that fact, although disquieting, need not be fatal to the proposition. We are

interested in using the values of $L(t, d)$ for values of t near those generated by the observation d, and we can generally ignore the region near $t = \infty$.

When we bootstrap sample the sequences, the number of differences between them will be distributed as a binomial variate with p trials and probability d/p. If there are m differences observed between two bootstrap sampled sequences, the estimate of P will be m/p and the estimate of t will be obtained by solving (17) for t as a function of $P = m/p$, obtaining

$$\hat{t} = -\frac{3}{8} \ln \left(1 - \frac{4}{3} \frac{m}{p} \right) \tag{19}$$

For long sequences, the distribution of m/p will asymptotically be normal with expectation d/p and variance $p(d/p)(1-d/p)$. Even after the nonlinear transformation of m/p into $\hat{t}$ by (19), we will expect asymptotically to see a normal distribution with expectation

$$E[\hat{t}] = -\frac{3}{8} \ln \left(1 - \frac{4}{3} \frac{d}{p} \right) \tag{20}$$

and variance, using the standard delta-method approximation, becomes (Kimura and Ohta, 1972):

$$\text{Var } [\hat{t}] = \frac{9d(p-d)}{4p(3p-4d)^2} \tag{21}$$

Figure 1 shows for $d = 10$ and $p = 100$ the histogram of the bootstrap estimates of t, the normal density with mean (20) and variance (21), and a density which is the likelihood curve (18) rescaled so that it sums to 1. Since the number of differences between the sequences with bootstrapping follows a binomial distribution, whose classes are evenly spaced on the P scale, they form classes unevenly spaced on the t scale. The likelihood and normal curves can be integrated over these intervals so that they too predict

the same histogram: the curves shown here for them are not the original ones but lines connecting the centers of their histogram bars.

Note that the likelihood curve and the density of bootstrap estimates do a reasonable job of approximating each other, but not much more closely than either is approximated by the asymptotic normal curve.

(c) The case of three species

For three species we can do less analytically, but a numerical example is instructive. For the Kimura 2-parameter model with transition/transversion ratio 2.0, using the tree in Figure 2 as the true tree, I simulated three sequences of 150 base-pairs of DNA. We can use those data to compute the likelihood at a grid of points that have different tree topologies and different branch lengths. Tables 1a, 1b, and 1c show the heights of the likelihood surface for the three tree topologies shown in Figure 3. The likelihoods have been rounded to integers after being scaled so that their total over the points on the grid was 1000. This is done to facilitate comparison with Tables 2a, 2b, and 2c. These show a histogram of the distribution of trees found in 1000 bootstrap replicates. The correspondence is not perfect, but the two distributions are generally similar. Figure 4 shows the likelihood curves for θ that one would obtain using the true tree, the bootstrap approximation shown in Table 2, and a distribution proportional to the likelihood surface. The general impression one gets is that the method is not working perfectly but is an acceptable approximation. It is certainly spreading itself over the three tree topologies, which the asymptotic normal approximation cannot do, as it is undefined on other tree topologies. The proportions in the three tree topologies are, however, not quite right – too few estimates are

in the second topology. However the likelihood curves estimated by the bootstrap are essentially correct. The likelihood curves from the bootstrap approximation and from the density proportional to the likelihoods are very similar. Both differ from the likelihood curve for the true tree but that is expected, as both are based on the same data set of 150 sites, while it is in effect based on a data set of an infinite number of sites.

6 The computation in practice - examples

We can get some feel for the process by carrying it out on a simulated data set of moderate size. In the first, the tree in Figure 5 was used to simulate 10 sequences, the branch lengths being scaled so that one unit of time is one expected change per site. Data sets of length 100, 250, 500, and 1000 sites were simulated, and for each 100 bootstrap replicates were made, the maximum likelihood trees estimated, and the priors for these summed according to equation (14). The resulting approximate likelihood curves are shown in Figure 6, together with the likelihood curve for $\theta = 4N_e\mu$ based on the true tree, which is in effect what would be found if an infinite number of sites were observed. Note that the value of $4N\mu$ is 0.4, which is far greater than is realistic in most populations. Conventional estimates of $4N\mu$ are in the vicinity of 0.1 *per locus* which is likely to be 100-fold less than the present value, which is per site.

The curves are reasonably similar, and the true value of $4N\mu = 0.4$ is close to the maximum value in all cases. As the number of sites becomes large the curves approach the curve for the true tree. Note that for smaller numbers of sites, the curve is expected to be different, and wider, as the error

based on finiteness of the number of sites becomes substantial compared to the error based on the finiteness of the number of sequences. The curves shown in this Figure seem to show this behavior, and all of them seem compatible with the true value of $4N\mu$. There is some possibility that the curves for 100 and perhaps 200 sites are systematically displaced leftwards.

In the second example, the tree in Figure 7 was used, which was generated with a coalescent with a much smaller value, $\theta = 0.004$, which is closer to the values that might be found in samples from actual populations. recall that the θ used here is computed *per site*, not per locus as in most previous papers. Thus $\theta = 0.1$ per locus might typically translate into $\theta = 0.0001$ per site. Sequences of length 100, 200, 500, and 1000 were simulated and 500 bootstrap replicates used. The resulting approximate likelihood curves are shown in Figure 8. The curve for 1000 sites approximates the curve for an infinite number of sites, but not very closely. The curves for fewer sites are not well-behaved. This suggests that one may need many sites to make reasonable estimates of θ.

In the third example, the tree in Figure 9 was used, which was generated with $\theta = 0.4$. Sequences of length 100, 200, 500 and 1000 were simulated, but this time 500 bootstrap replicates were used. The resulting curves (in Figure 10) are satisfactory: all of them have peaks at approximately the same value of θ, which is near the true value.

This picture is necessarily anecdotal. Until someone has the computing power to carry out a computer simulation of the behavior of this method, we will not know whether the patterns seen in these examples are general, or are specific to the 12 data sets (four in each of the three cases) that happened to be generated in these simulations, or whether the outcomes would have

been different with a different set of bootstrap samples.

With these qualifications, the pattern that these tests seem to find is that the method works if there are a sufficient number of informative sites. Thus it fails when θ and the lengths of the sequences are both small but seems to succeed when there are larger numbers of sites or a larger value of θ.

The method can be proven to work when the number of sites is so large that the estimates of trees after bootstrap sampling are all of the same tree topology, with branch lengths varying around the true tree in a multivariate normal distribution (see Efron's 1979a Remark G). There seems some sign in the above examples that the method fails with sequences with few informative sites, which should give wide variation in the topology of the bootstrap estimates of the genealogy. This raises the suspicion that perhaps the method does not work if there is any variation in the tree topologies of the bootstrap estimates. To check this it is helpful to examine the bootstrap estimates for 100 sites for the first example and for 200 sites for the second example. These are the smallest sequence lengths for which it can be argued that the method did work adequately.

Figures 11 and 12 show majority rule consensus trees (Margush and McMorris, 1975) for these two sets of bootstrap estimated trees. Each of these shows those groups that occurred in more than 50% of the bootstrap estimates. They have in the first case been further resolved by inclusion of a group which did not occur in more than half of the trees, but was the next most frequently occurring group compatible with the others (in this sense we should call these "plurality rule" consensus trees). The consensus trees give us a rough sense of the amount of variation in the topologies of the

bootstrap estimates. They show that in cases where the method worked, there was a modest but noticeable variation in tree topology. Thus the bootstrap Monte Carlo integration method appears to be able to cope with data sets that are not able to estimate the genealogy with precision, and to integrate the likelihood over a number of different tree topologies.

7 Limitations

The method has some assumptions and properties that limit its usefulness:

(1) *No recombination.* It assumes that the genealogy of the sequences is a tree, which in effect means that no recombination has occurred within the sequences during the time since they were all descended from an ancestral copy. It will thus apply only to sufficiently short nuclear gene sequences, and there is as yet no clear picture of how short such sequences must be.

(2) *Computationally slow.* The method requires that on the order of 100 maximum likelihood estimates of trees be made. If the number of sequences is larger than tested here, the maximum likelihood estimation will be correspondingly slower. The program used here (DNAMLK from PHYLIP version 3.4) has execution speed proportional to the cube of the number of sequences. This means that for a data set twice as large the method would take eight times as long. It might be noted in passing that the Cann, Stoneking, and Wilson (1987) data set is over 13 times as large as the ones we have tested, and the tests reported here took 1-2 days on a DECstation 5000/200 to produce one likelihood curve. Thus the method is currently at the borderline of practicality, but this borderline will of course move as computing power becomes cheaper and as algorithms are improved.

(3) *Independence of sites.* The bootstrap sampling of sites assumes that evolutionary processes at different sites occur at the same rate and have independent outcomes. The maximum likelihood method can be corrected (Felsenstein, in preparation) for inequalities of rates between sites without much difficulty, and this less restrictive model will still allow us to make the maximum likelihood estimates that the bootstrap Monte Carlo integration method needs. But the assumption that sites, even consecutive sites, evolve independently is an unrealistic one, and one that is necessary for the bootstrapping process. However, it is possible to modify the bootstrapping process (Künsch, 1989) to one that resamples blocks of consecutive sites. If the nonindependence of evolution at different sites largely involves nearby sites affecting each other, then it may be possible to use block-bootstrapping to correct for this.

8 Alternatives

As discussed above, pairwise methods are a flawed alternative to the present method because of their low power for sufficiently long sequences.

For the limiting case when $\theta \simeq 0$, other alternatives to the present method may be possible. Griffiths (1989) has developed and distributed a computer program to compute likelihoods by a recursive algorithm for computing likelihoods for samples under the infinite sites model. It requires that we know what is the ancestral state at each site, something we do not usually know. Strobeck (1983) showed how to compute the likelihood for the infinite sites model in the more realistic case where the ancestral state at each site is not known, but he did not show how to generalize the com-

putations to cases with larger numbers of different sequences. Strobeck's approach can be effective when there are no incompatibilities between the information at different sites, as is expected when $\theta \simeq 0$. When θ is large enough to have some sites that have had more than one mutation, the data will not fit their assumptions. It is possible that by eliminating these sites by finding the largest "clique" of mutually compatible sites, one could apply Strobeck's method, although the general pattern of computations is not yet known, and there would seem to be some chance of biased results.

There is one other possible alternative method that might escape from limitations (2) and (3). This is the Metropolis-Hastings Sampler (Hastings, 1970; Metropolis *et. al.*, 1953). This random sampling method progressively alters an estimate (in this case it would be an estimate of the tree) in a biased random walk, in such a way that the time spent in each neighborhood of the space of trees is guaranteed to be proportional to the likelihood of the trees there, if one continues long enough. If this method can be applied, it will guarantee the properties we need. It would sample from the space of trees in proportion to the likelihood of those trees, allowing us to take advantage of the cancellation in equation (12). One would then only have to make this random walk through the space of all genealogies, sample from the resulting sequence of trees and average the priors for each tree sampled, just as we do in the present method.

The difficulty is that, unlike the bootstrap case, successive trees in the Metropolis-Hastings sequence are far from independent, so one must sample many more of them. This is not as difficult as it sounds, as for each tree one need only evaluate its likelihood, and it is not necessary to carry out a full maximum likelihood estimate for each tree in the sequence. As there is

then far less computation per sampled tree, this might compensate for the greater number of trees sampled and result in a computationally more effective method. The Metropolis-Hastings method, like other Markov Chain Monte Carlo methods, might get stuck in an isolated peak of the likelihood function for long periods of time, if such peaks existed. The present method does not suffer from this potential problem, as it samples independently from this distribution. This makes it an excellent candidate for a method of periodically choosing new starting points for the Metropolis-Hastings method. It will be interesting to see if the Metropolis-Hastings Sampler can carry us further in the direction of computing the likelihood for population samples of sequences. Preliminary tests indicate that it will.

9 Computer programs

Bootstrap Monte Carlo integration estimation of the likelihood for sequence samples of modest size can be done using three programs in versions 3.5 and later of the PHYLIP phylogeny inference package (this version will be released in mid-1992). The program SEQBOOT can be used to resample sites and prepare a file with multiple bootstrapped data sets. The program that computes a DNA maximum likelihood tree for a clock model, DNAMLK, can then be used to make maximum likelihood estimates for each of these. The resulting "tree file" containing these estimates can then be read as an input file by a new program, COALTREE, which calculates and averages the priors for those estimates for a range of values of $4N_e\mu$.

PHYLIP is distributed free in C and Pascal source code, and in precompiled executable versions for generic PCDOS systems, for 80386/80387

PCDOS systems, and for Macintoshes. For information on distribution media and policies contact the author (most easily done by electronic mail) or use anonymous ftp over Internet to fetch the appropriate archives from directory pub/phylip of evolution.genetics.washington.edu (128.95.12.41).

I am grateful to Elizabeth Thompson, Monty Slatkin, Charles Geyer and Kermit Ritland for helpful discussions, and to reviewers for this journal for catching errors and suggesting improvements. This work was supported by National Science Foundation grant number BSR-8614807, and by National Institutes of Health grant number 1 R01 GM 41716.

References

Avise, J. C. (1989). Gene trees and organismal histories: a phylogenetic approach to population biology. *Evolution* **43,** 1192-1208.

Cann, R. L., Stoneking, M. & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature* **325,** 31-36.

Edwards, A. W. F. (1970). Estimation of the branch points of a branching diffusion process. *Journal of the Royal Statistical Society, Series B* **32,** 155-174.

Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Annals of Statistics* **7,** 1-26.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Philadelphia: Society for Industrial and Applied Mathematics.

Felsenstein, J. (1981). Evolutionary trees from gene frequencies and quantitative characters: finding maximum likelihood estimates. *Evolution* **35,** 1229-1242.

Felsenstein, J. (1985). Confidence limits on phylogenies with a molecular clock. *Systematic Zoology* **34,** 152-161.

Felsenstein, J. (1988). Phylogenies from molecular sequences: inference and reliability. *Annual Review of Genetics* **22,** 521-565.

Felsenstein, J. (1992). Estimating effective population size from samples of sequences: inefficiency of pairwise and segregating sites as compared to phylogenetic estimates. *Genetical Research*, in press.

Griffiths, R. C. (1989). Genealogical tree probabilities in the infinitely-many-site model. *Journal of Mathematical Biology* **27,** 667-680.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57,** 97-109.

Hammersley, J. M. & Handscomb, D. C. (1964). *Monte Carlo methods.* London: Methuen.

Jukes, T. H. & Cantor, C. (1969). Evolution of protein molecules. pp. 21-132 in *Mammalian Protein Metabolism*, ed. M. N. Munro. NewYork: Academic Press.

Kahn, H. (1950). Random sampling (Monte Carlo) techniques in neutron attenuation problems - I. *Nucleonics* **6, No. 5,** 27-37.

Kendall, M. G. & Stewart, A. (1973). *The Advanced Theory of Statistics. Vol. 2. 3rd Edition.* New York: Hafner.

Kimura, M. & Ohta, T. (1972). On the stochastic model for estimation of mutational distance between homologous proteins. *Journal of Molecular Evolution* **2,** 87-90.

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. *Journal of Applied Probability* **19A,** 27-43.

Künsch, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Annals of Statistics* **17,** 1217-1241.

Margush, T. & McMorris, F. R. (1981). Consensus n-trees. *Bulletin of Mathematical Biology* **43,** 239-244.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E.1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* **21,** 1087-1092 1953

Nei, M. & Tajima, F. (1981) DNA polymorphism detectable by restriction endonucleases. *Genetics* **97,** 145-163.

Strobeck, C. (1983). Estimation of the neutral mutation rate in a finite population from DNA sequence data. *Theoretical Population Biology* **24,** 160-172.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* **7,** 256 -276.

Tables

Table 1. Likelihood surface for a set of simulated 150 base sequences for the three tree topologies shown in Figure 3, for various values of the length of the longest branch (rows in the table) and the short interior branch (columns in the table). Likelihoods have been discretized so that they sum to 1000 over all three parts of the Table. Dots indicate zeros.

Table 2. Histogram of the distribution of trees estimated from 1000 bootstrap samples from the data used to calculate the likelihood surface shown in Table 1. The three parts of the table correspond to the three tree topologies shown in Figure 3. The rows and columns show various values of the length of the longest branch (rows in the table) and the short interior branch (columns in the table). Dots indicate zeros.

Figure captions

Fig. 1. Histograms of the bootstrap estimates of the divergence time of two sequences, for sequences of length 100 sites which differ at 10 sites. The solid curve shows the distribution from the bootstrap estimates, the narrowly dashed curve the asymptotic normal approximation based on equations (20) and (21), and the more loosely dashed curve a distribution proportional in height to the likelihood curve (18). The curves pass through the centers of the tops of histogram classes.

Fig. 2. The tree used to simulate evolution at 150 sites for the three-sequence calculations.

Fig. 3. The three tree topologies used in Tables 1 and 2.

Fig. 4. Estimated log likelihood curves for the three-sequence case used to calculate Tables 1 and 2. The solid curve is the likelihood we would get for sequences of infinite length, the other two for the particular simulated sequences of length 150.

Fig. 5. One tree generated by the coalescent with $4N\mu = 0.4$, and used to simulate evolution of the sequences analyzed in Figure 6.

Fig. 6. Estiumated log likelihood curves for sequences simulated on the tree in Figure 5. The solid curve shows the likelihood curve that would be obtained with an infinite number of sites, which would result in perfect estimation of the tree. The bootstrap Monte Carlo integration estimates are based on trees from 100 bootstrap replicates.

Fig. 7. A tree generated by the coalescent with $4N\mu = 0.0004$, and used to simulate evolution of the sequences analyzed in Figure 8.

Fig. 8. Estimated log likelihood curves for sequences simulated on the tree in Figure 7. The solid curve shows the likelihood curve that would be obtained with an infinite number of sites, which would result in perfect estimation of the tree. The bootstrap Monte Carlo integration estimates are based on trees from 500 bootstrap replicates.

Fig. 9. Another tree generated by the coalescent with $4N\mu = 0.4$, and used to simulate evolution of the sequences analyzed in Figure 10.

Fig. 10. Estimated log likelihood curves for sequences simulated on the tree in Figure 9. The solid curve shows the likelihood curve that would be obtained with an infinite number of sites, which would result in perfect estimation of the tree. The bootstrap Monte Carlo integration estimates are based on trees from 500 bootstrap replicates.

Fig. 11. Majority-rule consensus tree (fully resolved by a plurality-rule criterion) for the 100 bootstrap estimates of the tree for Figure 5 from the 100-site case in the data used to calculate Figure 6.

Fig. 12. Majority-rule consensus tree for the 500 bootstrap estimates of the tree for Figure 7 from the 1000-site case in the data used to calculate Figure 8.