Maximum-Likelihood Estimation of Population Divergence Times and Population Phylogeny in Models without Mutation

Rasmus Nielsen; Joanna L. Mountain; John P. Huelsenbeck; Montgomery Slatkin


Stable URL:
http://links.jstor.org/sici?sici=0014-3820%28199806%2952%3A3%3C669%3AMEOPDT%3E2.0.CO%3B2-5

*Evolution* is currently published by Society for the Study of Evolution.

Your use of the JSTOR archive indicates your acceptance of JSTOR’s Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR’s Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ssevol.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is an independent not-for-profit organization dedicated to creating and preserving a digital archive of scholarly journals. For more information regarding JSTOR, please contact support@jstor.org.
MAXIMUM-LIKELIHOOD ESTIMATION OF POPULATION DIVERGENCE TIMES AND POPULATION PHYLOGENY IN MODELS WITHOUT MUTATION

RASMUS NIELSEN,1 JOANNA L. MOUNTAIN, JOHN P. HUELSNENBECK, AND MONTGOMERY SLATKIN
Department of Integrative Biology, University of California, Berkeley, California 94720-3140

Abstract.—In this paper we present a method for estimating population divergence times by maximum likelihood in models without mutation. The maximum-likelihood estimator is compared to a commonly applied estimator based on Wright’s $F_{ST}$ statistic. Simulations suggest that the maximum-likelihood estimator is less biased and has a lower variance than the $F_{ST}$-based estimator. The maximum-likelihood estimator provides a statistical framework for the analysis of population history given genetic data. We demonstrate how maximum-likelihood estimates of the branching pattern of divergence of multiple populations may be obtained. We also describe how the method may be applied to test hypotheses such as whether populations have maintained equal population sizes. We illustrate the method by applying it to two previously published sets of human restriction fragment length polymorphism (RFLP) data.

Key words.—$F_{ST}$, maximum likelihood, phylogeny, population divergence, population history, population size, population subdivision.

Received July 31, 1997. Accepted February 18, 1998.

The estimation of population divergence times and genetic distances has many applications in the demographic and genetic analysis of population history. Several estimators of genetic distances have been proposed (Cavalli-Sforza and Edwards 1964; Nei 1972; Reynolds et al. 1983). In this article we propose a new maximum-likelihood estimator of population divergence time based on the coalescent model (Kingman 1982) with no mutation. We describe applications of this estimator, most notably the inference of population phylogenies and show that the maximum-likelihood estimator has desirable statistical properties. To evaluate the performance of the new estimator, we compare it to the estimator based on Wright’s $F_{ST}$ statistic described by Reynolds at al. (1983). $F_{ST}$ is the most commonly applied divergence estimator based upon a model of genetic drift without mutation.

$F_{ST}$ measures the reduction in heterozygosity in a subpopulation relative to the heterozygosity of the total population. It is a standard measure of population differentiation and is applied in the estimation of population divergence times (see for example Weir 1996). Under a model of neutral divergence of two populations from a common ancestral population, $F_{ST}$ can be calculated directly into an estimate of the time since population divergence. In this case, divergence is measured in terms of the scaled time, $T = t/2N$, where $t$ is the number of generations since the two populations diverged and $N$ is the effective size of each of the two populations. $T$ can then be estimated by

$$
\hat{T} = -\log(1 - \hat{F}_{ST})
$$

(Cavalli-Sforza 1969), where $\hat{F}_{ST}$ is an estimate of $F_{ST}$. In many cases, it is assumed that the divergence time between the populations is so small that mutation is of little importance compared to genetic drift. Such an assumption may be appropriate, for short time scales, for genetic markers such as allozymes, RFLPs, SNPs, and amplified fragment length polymorphisms (AFLP's). For this type of data, $F_{ST}$ is commonly estimated assuming that no mutations occurred since the time of divergence between the populations. A commonly applied estimator of $F_{ST}$ is

$$
\hat{F}_{ST} = \frac{1}{2} \sum \frac{\sum (\hat{P}_{ij} - \hat{P}_{jj})^2}{2(n - 1)} \left(2 - \sum \left(\frac{\hat{P}_{ii} + \hat{P}_{jj}}{\hat{P}_{ij}}\right)\right)
$$

$$
= \frac{1}{2} \sum \left(1 - \sum \hat{P}_{ij} \hat{P}_{ij}\right)
$$

(Reynolds at al. 1983), where $n$ is the sample size and $\hat{P}_{ij}$ is the frequency at the $i$th locus of the $j$th allele in the $k$th population sample.

In many studies the quantity of interest is not the time since the divergence of a pair of populations, but the branching pattern of several populations (i.e., the population phylogeny). Recent examples of cases in which population phylogenies have been estimated from allelic counts are the studies of Andean birds by Brumfield and Capparella (1996), Acanthocephalans by Beehler et al. (1996), tephritid flies by Eber and Brandl (1997), and Hawaiian Alsinoidae by Weller (1996). Typically, divergence times (genetic distances) are estimated for each pair of populations and the population phylogeny is then estimated from the pairwise distances by a clustering algorithm, such as UPGMA (Sokal and Michener 1958) or neighbor joining (Saitou and Nei 1987). Information is lost, however, by first reducing the data to a set of pairwise differences and then estimating the population phylogeny. Alternatively, the population phylogeny is estimated directly, assuming that the observed allele frequencies in the sample are equal to the true population frequencies, by parsimony (Swoford and Berlocher 1987), or by maximum likelihood (Felsenstein 1993). Using these methods, information is lost by reducing the allelic counts to frequencies. Furthermore, these methods may not take full advantage of knowledge of the underlying population genetic process. Below we describe, after introducing a maximum-likelihood method for estimating population divergence times, a statistical popu-
Fig. 1. The divergence of two populations from an ancestral population. T is the divergence time and \( n_1 \) and \( r_1 \) are, respectively, the observed and ancestral configurations of the sample in the two populations.

**Derivation of the Likelihood Function**

In the following we consider a model in which two populations have diverged from an ancestral population (Fig. 1) \( T \) generations in the past. We will assume that the effect of mutation can be ignored. This assumption is reasonable for values of \( t/N \) where drift plays the dominant role. Moreover, we assume that the effect of drift in the two populations can be described by a neutral coalescent model (Kingman 1982). This implies selective neutrality and random mating within populations. However, unequal population sizes can be accounted for in this model. The expression for the likelihood function depends on the assumptions made regarding the ancestral population. We consider three different approaches. (1) We derive the likelihood function without making any assumptions regarding the ancestral population. This implies that we must estimate the frequencies of the alleles in the ancestral population. (2) We examine a pseudolikelihood estimator in which a method of moments estimate of the ancestral frequencies is substituted into the likelihood function for the allele frequencies in the ancestral population. (3) We examine a likelihood estimator that assumes that the ancestral population evolves according to an infinite alleles model (Kimura and Crow 1964).

To derive the first estimator we must obtain an expression for \( L(F,T) = P(n_1,n_2|F,T,n_1,n_2) \), where \( n_1 \) is the configuration of the sample obtained from the first population, \( n_2 \) is the configuration of the sample obtained from the second population, \( n_1 \) and \( n_2 \) are the respective sizes of the two samples, \( T \) is the divergence time scaled by the effective population size, and \( F \) is the vector of the ancestral frequencies (here, and in the following, symbols in bold imply vectors). For example, if we obtained a sample of three copies of one allele, four copies of another allele and 12 copies of a third allele from the first population then \( n_1 = (3,4,12) \) and \( n_1 = 19 \). In other words, \( n_1 \) is the sum of the elements in \( n_1 \). Subscripts will always denote populations. Absence of subscripts indicates that only one population is considered. Superscripts denote alleles, for instance, in the example above \( n_1^1 = 4 \).

To obtain the likelihood function we condition on the number of coalescence events occurring since the time of divergence in the ancestry of the samples. Let \( C \) denote the number of coalescence events in the first population and \( C_2 \) denote the number of coalescence events in the second population. Then, by conditioning on the number of coalescence events occurring in the ancestry of the two samples, we obtain

\[
L(T, F | n_1, n_2) = \sum_{C,C_2:C_1=n_1-1, C_2=n_2-1} P(n_1, n_2 | C_1, C_2, F) P(C_1, C_2 | T)
\]

(3)

because the probability of obtaining the sample, conditional on the number of coalescence events, is independent of \( T \), and the distribution of the number of coalescence events is independent of \( F \).

Now let \( r_1 \) and \( r_2 \) denote the ancestral samples at a time immediately after \( T \). Then, by conditioning on all possible values of \( r_1 \) and \( r_2 \), noting that \( C_1 \) is completely specified by \( n_1 \) and \( r_1 \), and that the distribution of \( n_1 \) and \( n_2 \) does not depend on \( F \) when \( r_1 \) and \( r_2 \) are known, we obtain:

\[
L(T, F | n_1, n_2) = \sum_{r_1,r_2:r_1=n_1-1, r_2=n_2-1} P(n_1, n_2 | r_1, r_2, C_1, C_2) \times P(C_1, C_2 | T, r_1, r_2) P(r_1, r_2 | F).
\]

(4)

Equation (4) can also be expressed as

\[
L(T, F | n_1, n_2) = \sum_{r_1,r_2:r_1=n_1-1, r_2=n_2-1} P(n_1 | r_1, T) P(r_1 | F) P(n_2 | r_2, T) P(r_2 | F),
\]

(5)

where \( P(n_1 | r_1, T) = P(n_1 | r_2, C_3) P(C_3 | T, r_2) \). In other words, the likelihood function is obtained by summing over all possible configurations of the ancestral samples at a time immediately after the populations diverged and by conditioning on the number of coalescence events.

The probabilities on the right-hand side of equation (5) can be obtained analytically. \( P(C | T, r) \) is the probability that \( C \) coalescence events occur in the ancestry of a sample of size \( C + r \) in time \( T \). Tavaré (1984) showed that the probability of observing \( j \) ancestors of a sample of size \( i \) after a time \( T \) (going backward) is

\[
g_j(T) = \sum_{k=0}^{i} p_k(T) \frac{(2k - 1)(-1)^{j+k} i_j k_{i-j} i_{i-k}}{j! (k-j)! i_{i-k}}, \quad 2 \leq j \leq i
\]

(6)

\[
g_0(T) = 1 - \sum_{k=2}^{i} p_k(T) \frac{(2k - 1)(-1)^{j+k} i_{i-k}}{i_{i-k}}
\]

(6)
where \( p_k(T) = e^{-\theta K} K^{T-1/2} \), \( a_{ij} = a(a+1) \ldots (a+j-1) \), and \( a_{i1} = a(a-1) \ldots (a-j+1) \). In the present notation \( g_{(j+1)r}(T) = P(C(T,r), P(r|F) \) is simply given by the multinomial sampling probability:

\[
P(r|F) = \frac{r!}{r_{i1}! r_{i2}! \ldots r_{ik}!} \prod_{i=1}^{k} F_i^{r_i},
\]

where \( r_i \) is the count of the \( i \)th allele among a total of \( k \) different alleles in the sample and \( F_i \) is the frequency of the \( i \)th allele in the ancestral population.

We can also derive an expression for \( P(n|r,C) \). Slatkin (1996) showed that in a two-allele model, the probability of obtaining a sample \( n \) at the present, conditional on the ancestral sample \( r \) and the sample size, is

\[
P_2(n|r,C) = \frac{(n^1 - 1)(n^2 - 1)}{(r^1 - 1)(r^2 - 1)},
\]

where \( n^i \) is the number of copies of allele \( i \) in \( n \) and \( r^i \) is the number of copies of allele \( i \) in \( r \). This expression follows from the result of Kingman (1982) that all configurations will be equally likely. The derivation of the probability then has been reduced to a balls-in-urns problem. The extension to the \( k \)-allele case follows trivially from this result. \( P(n|r,C) \) is given by the probability that \( n \) balls are distributed into \( r \) nonempty urns. \( n^i \) are in the first \( r^i \) urns, \( n^2 \) are in the next \( r^2 \) urns, and so forth:

\[
P(n|r,C) = \frac{\prod_{i=1}^{k} (n^i - 1)}{(r^i - 1)}. \tag{9}
\]

Hence, all expressions in equation (5) can be easily calculated, and the joint maximum-likelihood estimate of the divergence time \( T \) and the allele frequencies in the ancestral population \( F \) can be obtained by numerically maximizing equation (5). We denote this estimator \( DL_1 \).

In general, likelihood estimators in which the number of parameters increases linearly with the amount of data do not have desirable properties. In certain cases it is preferable to substitute estimates obtained by other means for some of the parameters in the likelihood function. As long as these estimates are consistent, the likelihood estimator will also be consistent, where consistency means that the estimate converges to the true value as the sample size increases. In our case, we increase the number of loci sampled, the number of parameters (ancestral allele frequencies) increases. However, the observed allele frequencies provide an unbiased estimate of the allele frequencies in the ancestral population. By substituting these moment estimates of the allele frequencies into the likelihood function, a new estimator is obtained. We denote this estimator \( DL_2 \).

The last estimator we consider assumes that the ancestral population has evolved according to an infinite alleles model. This alleviates the need for estimating the ancestral allele frequencies. We assume that the long-term process that has generated the genetic diversity can be described by the infinite alleles model, but that the recent divergence between populations is dominated by genetic drift. In a population evolving according to an infinite alleles model, the sampling distribution, conditional on the number of alleles, is provided by Ewens’ formula:

\[
P(r|\theta) = \frac{r!}{\theta^r (r-1)!} \prod_{j=1}^{r} \left( \frac{\theta}{j} \right)^{n_j} \frac{1}{\alpha_j}, \tag{10}
\]

(Ewens 1972), where \( \alpha_j \) is the number of elements in \( r \) equal to \( j \) and \( \theta = 4N\mu \), where \( N \) is the effective population size and \( \mu \) is the mutation rate. Notice that the number of alleles in the ancestry of the sample immediately after the time of population divergence must equal the observed number of alleles because we do not allow for mutations. We can therefore apply Ewens formula to obtain

\[
P(r_1, r_2|\theta) = P(r_1 + r_2|\theta) \prod_{i=1}^{k} \left( \frac{r_i!}{r_i} \right), \tag{11}
\]

where \( k \) is the total number of alleles in the two samples and the second part of the expression is the probability of dividing a sample \( r^1, r^2, r^3, \ldots, r^k \) into two sets of \( r_1, r_2, \ldots, r_1 \) and \( r^2, r^3, \ldots, r^2 \). This expression can be used to obtain a new maximum-likelihood estimator of divergence time based on the assumption of an infinite alleles model in the ancestral population:

\[
L(T, \theta | n_1, n_2) = \sum_{r_1, r_2} P(n_1|r_1) P(n_2|r_2) P(r_1, r_2). \tag{12}
\]

We denote this estimator \( DL_3 \). Notice that the application of this estimator requires the simultaneous estimation of \( \theta \). An example of a likelihood surface for \( DL_3 \) is shown in Figure 2.

Under the assumption of free recombination, the joint maximum-likelihood estimate for many independent loci is obtained by combining the likelihood surfaces for all loci.

**IS \( F_{ST} \) A SUFFICIENT STATISTIC?**

The availability of an expression for the likelihood function enables us to analyze properties of the \( F_{ST} \)-based estimator of population divergence. It can, for example, be determined whether \( F_{ST} \) is a sufficient statistic for the divergence time. A sufficient statistic is one that contains all the information in the data regarding the parameter of interest. In our case, we want to know whether the \( F_{ST} \) estimator contains all of the information in the data regarding the population divergence time.

A simple method to show that an estimator is not sufficient is to show that there are two samples that provide the same value of \( F_{ST} \) but different values of the likelihood. For example, consider the very simple case where three loci have been tested in each of the two populations: \( (2, 0, 1) \) and \( (0, 2, 1) \). In this case, \( F_{ST} = 0.4 \), applying the estimator by Reynolds et al. (1983), and the log likelihood for \( T = 0.1 \) and \( F = \{1/3, 1/3, 1/3\} \) is \(-4.42\). Now consider the two samples \( (3, 0, 0) \) and \( (1, 1, 1) \). In this case, the estimate of \( F_{ST} \) is also \( F_{ST} = 0.4 \), but the log likelihood value for \( T =
0.1 and $F = \{1/3, 1/3, 1/3\}$ is $-4.63$. In other words, there are two sets of samples that provide the same $F_{ST}$-value but two different likelihood values. Noting that the ratio of the likelihood values will change with $T$, we have shown that $F_{ST}$ is not a sufficient statistic for $T$. This result will also hold for other estimators of $F_{ST}$. Information is lost if $F_{ST}$ is used instead of the likelihood function and this suggests that the maximum-likelihood estimator of divergence time should have lower variance than the $F_{ST}$-based estimator, at least for large samples.

**Performance of the Methods**

In this section we evaluate the performance of the four estimators of divergence time: $DF_{ST}$, $DL_1$, $DL_2$, and $DL_3$. However, because evaluation of the likelihood function may be time consuming for large samples, we initially consider only samples of size $n = 10$. Estimation of one maximum likelihood estimate requires the numerical optimization of a function that sums over all possible configurations of the ancestral sample and typically takes several hours for datasets containing more than 50 individuals. Also, because the assumption of no mutations is violated when the divergence time is large, we only evaluate the performance of the estimators for small divergence times (e.g., $T \leq 0.1$).

The first issue we consider is the bias of the estimators. For this purpose two sets of simulations were performed. In the first set of simulations, we assume that the ancestral population has evolved according to an infinite alleles model (Fig. 3a) with $\theta = 4N_\mu = 1.0$, where $N$ is the effective population size and $\mu$ is the mutation rate. In the second set of simulations, we assume that there are four alleles in the ancestral population, each at frequency 0.25 (Fig. 3b).

The graphs in Figure 3 demonstrate that $DL_1$ and $DL_2$ are both strongly biased. In other words, estimation of the ancestral frequencies leads to an underestimate of the divergence time. This is true for both sets of simulations even though they represent opposite extremes. The reason for this bias is that the estimated frequencies deviate from the true values in a direction toward the observed frequencies at time $T = 0$. This tends to bias the estimate of divergence time toward smaller values of $T$. For example, when the pseudomaximum-likelihood ($DL_3$) method is applied, the estimate of the ancestral allele frequencies is simply the observed frequencies. By estimating the ancestral allele frequencies by the current allele frequencies, we minimize the effect of genetic drift, which, in turn, will bias the estimate of $T$. In the graphs in Figure 3, the bias of $DL_1$ appears to persist for larger numbers of loci. One of the reasons for this may be
Table 1. The variance in the estimate of divergence time for $T = 0.1$ as estimated by 1000 simulations.

<table>
<thead>
<tr>
<th>$n = 10, 10$ loci</th>
<th>$F$ = (1/4, 1/4, 1/4, 1/4)</th>
<th>Var($DF_{ST}$)</th>
<th>$F$ = (1/4, 1/4, 1/4, 1/4)</th>
<th>Var($DL_3$)</th>
<th>$F = 1.0$</th>
<th>Var($DF_{ST}$)</th>
<th>$F = 1.0$</th>
<th>Var($DL_3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00259</td>
<td>0.00124</td>
<td>0.000405</td>
<td>0.00207</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00445</td>
<td>0.00238</td>
<td>0.00863</td>
<td>0.00382</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00223</td>
<td>0.00093</td>
<td>0.00136</td>
<td>0.00083</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

that the number of parameters increases linearly with the number of loci considered. Likewise, the bias for $DL_2$ does not disappear for larger numbers of loci (Fig. 3). In contrast, $DL_3$ appears to be only moderately biased for the values explored in this set of simulations. In the following, therefore, we concentrate on comparing the variance of $DF_{ST}$, and $DL_3$.

The variances of $DF_{ST}$ and $DL_3$ were evaluated by simulation (Table 1). For all conditions examined, the variance of $DL_3$ is lower than that of $DF_{ST}$. This result is expected because the likelihood estimator applies more of the information in the data than does the $F_{ST}$-based estimator. Also, the variance of $DL_3$ is lower than that of $DF_{ST}$ even when the assumption of an infinite allele model in the ancestral population is violated. This suggests that $DL_3$ may be relatively robust to violations of the assumption of an infinite allele model. Because $DL_3$ appears less biased when a small number of individuals has been sampled and has a lower variance than the $F_{ST}$-based estimator, $DL_3$ is preferable to $DF_{ST}$ as an estimator of population divergence.

**Estimation of Population Phylogenies**

A statistical, population genetic approach to the estimation of population phylogenies is to integrate over all possible coalescence trees for each locus. However, the computational burden of performing this integration directly would be phenomenal. Fortunately, the integral can also be evaluated using the present approach. Notice that $P(n|r,T)$ provides the transition probability along a branch in the population phylogeny, that is, the probability of observing state $n$ at the present given that $r$ was observed a time $T$. The likelihood for a given population phylogeny can therefore be evaluated by summing over all possible states at each node while multiplying the relevant transition probabilities, $P(n|r,T)$, for each branch of the phylogeny. For example, imagine that the configurations $n_1$, $n_2$, and $n_3$ were observed for three populations. Let $r_1$ and $r_2$ be the ancestral configurations at a time right after population 1 and 2 diverged and $r_2^*$ be the ancestral configuration at a time right before the divergence of the populations leading to $n_1$ and $n_2$. Finally let $r_{12}$ and $r_3$ be the ancestral configurations at the root of the phylogeny (Fig. 4). Then the likelihood function is provided by

$$L(T_1, T_2, T_3, T_4, \theta | n_1, n_2, n_3) = \sum_{r_1, r_2} P(n_1 | r_1)P(n_2 | r_2)P(r_1, r_2 | r_{12}^*) \times \sum_{r_{12}, r_3} P(r_{12}^* | r_{12})P(n_3 | r_3)P(r_{12}, r_3).$$

In other words, the likelihood is obtained by summing over all possible states at the internal nodes of the phylogeny. This approach is conceptually similar to the approach applied in phylogenetic inference when estimating a single gene genealogy (Felsenstein 1981). However, there are some notable differences. First, in the present approach the parameters of interest are not the gene genealogies, but the integral of the likelihood of the underlying population phylogenies evaluated over all possible gene genealogies. This method accounts for phenomena such as ancestral polymorphism and the discrepancy between gene trees and species trees. Second, the state space may change through the population phylogeny. Nodes deeper in the phylogeny may include more states than nodes near the tip of the phylogeny. Thus the state space changes at each node when the topology of the phylogeny changes, calling for a more dynamic implementation of the computational algorithm than is normally required for problems in phylogenetic inference. Third, the underlying stochastic process is not reversible, so that only rooted phylogenies can be evaluated. This means that a unique root of the tree can be inferred without relying on midpoint rooting or other heuristic methods.

---

**Fig. 4.** A graphical representation of the model assumed for the divergence of three populations: $n_1$, $n_2$, and $n_3$ are the observed samples; $r_1$ and $r_2$ are the ancestral configurations at a time right after population 1 and 2 diverged and $r_{12}^*$ is the ancestral configuration at a time right before the divergence of the populations leading to $n_1$ and $n_2$. $r_{12}$ and $r_3$ are the ancestral samples right after the divergence of the common ancestral population.
In many studies inference of the root is central to the demographic or population biological analysis.

**Hypothesis Tests**

For any set of three or more populations, at least two hypotheses can be tested. One can consider, for example, the null hypothesis that three or more of the populations are related by a star phylogeny. The alternative hypothesis relaxes the constraint of a multiway split among the populations. A second null hypothesis to be considered is that the populations have equal effective population sizes. The alternative hypothesis relaxes this constraint by allowing each population to have a unique population size (and hence a unique rate of genetic drift). In both cases, the likelihood is maximized under both the null hypothesis \( (L_0) \) and the alternative hypothesis \( (L_1) \). Because the models are nested, \(-2(\log L_0 - \log L_1)\) is approximately chi-square distributed with \( q \) degrees of freedom, where \( q \) is the difference in the number of parameters between the two models. The test of a trifurcating phylogeny has one degree of freedom, for example, while the test of equal sizes among three populations has two degrees of freedom.

**Illustration of the Method**

The inference of population branching patterns is central in evolutionary biology. We have developed a coalescence-based framework for such inference. Human population genetic datasets are currently the largest in terms of the number of loci tested. These data provide, therefore, the best opportunity for illustration of the maximum-likelihood approach to population phylogeny. Here we illustrate with two cases, analyzing datasets wherein population samples have been examined for a relatively large number of independently evolving loci (45) of the nuclear genome.

We first consider the relationships among three sub-Saharan African populations: the Biafra of the Central African Republic, the Mbuti of the Republic of Congo, and the Mandenka of Senegal. We then examine the relationships among three more geographically distant populations: the Japanese, the Italians of Trino, and the Mandenka of Senegal. Each population is represented by a sample of 10 individuals (20 chromosomes) tested for each of 45 restriction fragment length polymorphisms (RFLPs). These RFLPs are a subset of those previously published; for each locus only the polymorphism generated by one restriction enzyme was considered. Only those loci were considered for which at least 10 individuals of each of the three populations had been examined. Sample processing and DNA analyses have been described elsewhere (Bowcock et al. 1987; Matullo et al. 1994; Poloni et al. 1995).

For each set of three populations we obtained the maximum likelihood of the three possible rooted topologies by optimizing branch lengths and other parameters. In addition, we considered two null hypotheses for each dataset. The first null hypothesis we considered is that the populations are related by a trifurcation. The second null hypothesis we considered is that the three populations have maintained equal effective population sizes. In all cases the results should be interpreted with caution because the effect of migration has not been considered explicitly.

The Mbuti, elsewhere known as eastern Pygmies, are from the Ituri forest in the northeastern part of the Republic of Congo (formerly Zaire) and speak Nilo-Saharan languages. The Biaka, also known as the Aka, Babinga, or western Pygmies, speak languages of the Niger-Kordofanian family. Samples were obtained from villages in southwestern Central African Republic. There is currently no genetic exchange between these two Pygmy groups, which are geographically isolated (Cavalli-Sforza et al. 1994). Samples from these two populations were collected by L. Cavalli-Sforza and B. Hewlett (Cavalli-Sforza 1986). The third sample consists of Niokolokone, members of the Mandenka population of eastern Senegal, and was obtained by A. Langaney and colleagues (Tiercy et al. 1992; Poloni et al. 1995). The Mandenkalu speak a language of the Niger-Kordofanian family.

The relationships among these three sub-Saharan populations have been examined previously with other approaches. Poloni et al. (1995) used the neighbor-joining method to infer relationships among nine human populations based on RFLP allele frequencies (of which the data considered here are a subset). They found that the two Pygmy populations clustered most closely among the three African groups. Previous analysis had also suggested that the Biaka are more highly mixed with non-Pygmy populations than are the Mbuti (Wijsman et al. 1984). However, no statistical test of the hypothesis has been performed. Here we have maximized the likelihood for the three possible rooted topologies (Fig. 5). Because the units of the branch lengths are given by \( t/2N_e \), this type of analysis estimates and accommodates differences in the effective population sizes among populations, including ancestral populations. The population phylogeny with the highest likelihood is shown in Figure 5a. The Biaka and Mbuti form a cluster on this tree. The other two trees (Figs. 5b,c) are essentially identical: in each case one internal branch has collapsed, leading to a trifurcation. Branch lengths are also similar. While the difference between the likelihood of the first topology and that of the trifurcating tree is not significant at the \( P < 0.05 \) level, it is close to significance \( (0.05 < P < 0.1) \). Forty-five markers may provide insufficient power in this case. The weights for the branches of Figure 5a vary greatly, suggesting deviation from a constant and equal population size model. Assuming that the Biaka have received more gene flow from non-Pygmy groups (Wijsman et al. 1984), we might expect this group to have a relatively short branch. Based on the likelihood ratio test, however, we fail to reject the hypothesis that population sizes are equal.

As a second illustration we considered three more geographically distant populations. The Japanese sample, consisting of individuals born in Japan and living in the San Francisco Bay area, was collected by L. Wang and A. A. Lin. The Mandenkalu sample is described above. The Italian sample was collected from the town of Trino in northern Italy by A. Piazza and colleagues. This sample is described in greater detail elsewhere (Matullo et al. 1994).

The maximum-likelihood trees for the three possible topologies relating the Italian, Japanese, and Mandenka samples are shown in Figures 5d–f. The tree of Figure 5d has the highest likelihood, and is significantly more likely than
the trifurcating tree. As in previous studies, the European (TRO) sample appears to be intermediate between the Asian (JPN) and African (MAN) samples, with a short branch. Numerous genetic studies have indicated that European populations are genetically intermediate between African and Asian populations and/or have a short branch (Cavalli-Sforza et al. 1988; Nei and Livshits 1989; Bowcock et al. 1991, 1994; Poloni et al. 1995). Unlike other methods that allow for variable evolutionary rates, this approach locates a root for each tree. The root of the maximum-likelihood tree of Figure 5d indicates that the greatest degree of isolation has been between the African group on the one hand, and the Asian and European groups on the other. The tree of Figure 5d is also significantly more likely than a tree inferred assuming equal population sizes. It is possible that these populations have maintained very different long-term effective population sizes, the bifurcating model is oversimplified, or both.

**Discussion**

Simulation studies presented here demonstrate that the maximum-likelihood estimator of divergence time has less bias and lower variance than the $F_{st}$-based estimator. This difference between the estimators may arise because $F_{st}$ is not a sufficient statistic for divergence time. Therefore, the maximum-likelihood estimator is preferable for studies in which the divergence time between populations is of interest. The likelihood estimator may also be preferable for studies in which the subject of interest is not the divergence time per se, but rather the relative levels of divergence among populations.

We have illustrated the latter case with two examples. We applied the maximum-likelihood approach ($DL_2$) to the estimation of divergence among three sub-Saharan African samples and among three more globally distributed samples. Both migration and shared ancestry are important in explaining the genetic relationship among populations. Most methods fail to distinguish between these two phenomena underlying genetic similarity, and the present method is no exception. Interpretation of any phylogeny relating human populations, including the illustration presented here, must be carried out with this in mind.

The tree of African populations with the highest likelihood (Fig. 5a) groups the two Pygmy populations. We did not find strong evidence, however, that this tree is significantly more likely than a trifurcating tree. The two Pygmy populations do not appear to be much more closely related to one another than either is to the Mandenka sample. This result reflects the isolation that has been maintained between the two Pygmy populations (Cavalli-Sforza et al. 1994). Our analysis also suggested that the difference in population sizes are so small that they cannot be detected using 45 loci. For the three
populations from Africa, Asia, and Europe, we find strong evidence for an internal branch indicating that the closest relationship is between the Italian and Japanese samples. We also find clear evidence of differences in evolutionary rates: the European sample appears to be derived from a population with a very large effective population size. This probably reflects a bias in the ascertainment of RFLPs (Mountain and Cavalli-Sforza 1994). Also, the bifurcating model is clearly oversimplified. The large estimate for European population size may reflect major genetic contributions, over an unknown period of time, from both Africa and Asia. The European population may have originated as an admixture of ancestral Africans and ancestral Asians or it may have received high levels of gene flow from these populations more recently (Bowcock et al. 1991; Sokal et al. 1991; Cavalli-Sforza et al. 1994). Alternatively, the tree of Figure 5d may reflect differences in long-term effective population sizes and rates of gene flow rather than a bifurcating history (Relethford 1995; Harpending et al. 1996; Jorde et al. 1997). This example illustrates the shortcomings of the method in failing to account explicitly for gene flow.

The coalescent-based estimation of population phylogenies has several advantages over previous approaches. It applies more of the information in the data, provides rooted phylogenies, and provides a framework for hypothesis testing. However, the computer time required for carrying out the estimation increases dramatically with the addition of even one population. The method should be applicable to larger datasets, given sufficient computing resources. We envision this new coalescent-based method applied in many studies where the phylogeny of populations is of interest. Moreover, in the future, migration between populations may be incorporated into the model, allowing joint maximum-likelihood estimation of population sizes, migration rates, and divergence times. The relative roles of shared history and migration in the generation of current genetic diversity might thereby be directly examined.

ACKNOWLEDGMENTS

This work was supported in part by National Institute of Health grant GM40282 to M. Slatkin and by a personal grant to RN from the Danish Research Council. We thank T. Speed for suggesting the method for demonstrating that \( F_{ST} \) is not a sufficient statistic, and E. Williamson and C. Muirhead for comments on an earlier version of the manuscript. We thank the NOV Project at University of California at Berkeley, and especially E. Brewer, for providing computer time for the simulation and maximum-likelihood analysis.

LITERATURE CITED


Corresponding Editor: M. Hammer