Estimation of Past Demographic Parameters From the Distribution of Pairwise Differences When the Mutation Rates Vary Among Sites: Application to Human Mitochondrial DNA

Stefan Schneider and Laurent Excoffier

Genetics and Biometry Laboratory, Department of Anthropology and Ecology, University of Geneva, CP 511 1211 Geneva 24, Switzerland

Manuscript received July 30, 1998
Accepted for publication March 19, 1999

ABSTRACT

Distributions of pairwise differences often called “mismatch distributions” have been extensively used to estimate the demographic parameters of past population expansions. However, these estimations relied on the assumption that all mutations occurring in the ancestry of a pair of genes lead to observable differences (the infinite-sites model). This mutation model may not be very realistic, especially in the case of the control region of mitochondrial DNA, where this methodology has been mostly applied. In this article, we show how to infer past demographic parameters by explicitly taking into account a finite-sites model with heterogeneity of mutation rates. We also propose an alternative way to derive confidence intervals around the estimated parameters, based on a bootstrap approach. By checking the validity of these confidence intervals by simulations, we find that only those associated with the timing of the expansion are approximately correctly estimated, while those around the population sizes are overly large. We also propose a test of the validity of the estimated demographic expansion scenario, whose proper behavior is verified by simulation. We illustrate our method with human mitochondrial DNA, where estimates of expansion times are found to be 10–20% larger when taking into account heterogeneity of mutation rates than under the infinite-sites model.

With the advent of the coalescent theory (Kingman 1982), people have become increasingly aware of the profound effect of demography on the amount of genetic variability maintained in a population (Hudson 1990; Donnelly and Tavaré 1997). Population expansions or contractions indeed leave recognizable signatures in the pattern of molecular diversity (reviewed in Harpending et al. 1998). For instance, sudden demographic expansions lead to star-shaped phylogenies and unimodal distributions of pairwise differences (Slatkin and Hudson 1991; Rogers and Harpending 1992), to a reduction of the number of segregating sites (Bertorelle and Slatkin 1995; Arsène-Brosou and Excoffier 1996; Tajima 1996), to a lower amount of linkage disequilibrium between linked loci (Slatkin 1994), or to the occurrence of a large proportion of very low frequency mutations (Fu and Li 1993; Fu 1997). Note that population bottlenecks usually have other recognizable effects, often opposed to those of population expansions (e.g., Tajima 1993; Cornuet and Luikart 1996; Harpending et al. 1998).

When exposed to the evidence of a past demographic expansion, one might want to estimate the parameters of the expansion, such as the time at which it occurred and its magnitude, but the choice of parameters to be estimated depends on a particular scenario of population growth one might choose, such as exponential growth, logistic growth, or an instantaneous stepwise population size change. These three models are obviously related but have rarely been compared (but see Polanski et al. 1998). The latter model has been favored in the literature due to its simplicity and because simulation studies have shown that it was a good approximation of rapid logistic growth (Rogers and Harpending 1992). Rogers and Harpending (1992) convincingly showed that under an infinite-sites model, the shape of the distribution of the number of observed differences between pairs of DNA sequences (often called the mismatch distribution) conveyed information on the timing and the amplitude of a stepwise expansion. They proposed a nonlinear least-squares approach (Rogers and Harpending 1992) or a method of moments (Rogers 1995) to find the theoretical mismatch distribution that would best fit the observations. Several authors noted that this approach could be inadequate (Lundstrom et al. 1992; Bertorelle and Slatkin 1995; Arsène-Brosou and Excoffier 1996; Wakeley and Hey 1997) because the infinite-sites model was not realistic, especially in the case of the mitochondrial genome where an important heterogeneity of mutation rates had been observed (Lundstrom et al. 1992; Wakeley 1993). In fact, the presence of a few fast mutating sites can lead to unimodal mismatch distributions even for moderate
expansions (Lundström et al. 1992; Aris-Brosou and Excoffier 1996) and would thus lead to an underestimation of the age of the expansion and to an overestimation of its magnitude. Rogers and his collaborators tried to address these concerns. They concluded that both the mean number of pairwise difference (Rogers 1992) and the confidence intervals around the estimated demographic parameters were relatively unaffected by slight to moderate amounts of rate heterogeneity (Rogers et al. 1996). They thus proposed to ignore any heterogeneity of mutation rates when estimating demographic parameters from the mismatch distribution and therefore to stick to the infinite-sites predictions. One cannot be entirely satisfied by these conclusions because simulation studies have shown that the mean of the mismatch distribution was indeed very sensitive to rate heterogeneity (Aris-Brosou and Excoffier 1996) and that the relative insensitivity of the confidence intervals to rate heterogeneity was mainly due to their very large size (Rogers et al. 1996).

We thus propose in this article to extend the model of Rogers and Harpending (1992) to explicitly take into account possible heterogeneity of mutation rates when estimating the demographic parameters. We also propose an alternative way to derive confidence intervals around the estimated parameters based on a simple bootstrap approach. We check by simulations the validity of those confidence intervals for a few test cases and show that only those associated with the timing of the expansion are approximately correctly estimated. As one can sometimes observe a poor fit between the data and the mismatch distribution predicted by the model, we also propose a test of the validity of the stepwise demographic expansion scenario. We finally illustrate our method by human data from the mitochondrial DNA control region.

**THEORY AND METHODS**

The mismatch distribution under the infinite-sites model: We assume that t generations ago, a population at equilibrium of size \( N_0 \) entered a demographic expansion phase to instantaneously reach a new size \( N_1 \) and that it remained at that size ever since. Under this demographic scenario described in Figure 1 and assuming that every new mutation occurs at a new site [the infinite-sites mutation model of Kimura (1969); Li (1977)] derived an expression for the probability of observing \( i \) differences between two genes taken at random from this population, as

\[
F_i(\theta_1, \theta_o, \tau) = F_i(\theta_1) + \exp\left(-\frac{\theta_1}{\theta_1 + 1}\right) \times \sum_{j=0}^{i-1} \tau_{i-j}^j [F_{i-j}(\theta_0) - F_{i-j}(\theta_1)],
\]

where \( \theta_0 = 2N_o\theta, \theta_1 = 2N_1\theta, \tau = 2ut, \) and \( u \) is the total mutation rate per generation per gene. Here, \( F_i(\theta) \) is also the probability of observing \( i \) differences between two genes in an equilibrium population as (Waterson 1975)

\[
F_i(\theta) = \frac{\theta^i}{(\theta + 1)^{i+1}}.
\]

Rogers and Harpending (1992) rederived Equation 1 and used it to describe the distribution of the number of pairwise differences between nonrecombining DNA sequences or RFLP haplotypes in a given sample, which they called the “mismatch distribution.” They proposed to estimate the demographic parameters \( \theta_o, \theta_1, \) as well as the expansion time \( \tau \) directly from this mismatch distribution.

The mismatch distribution under a finite-sites model: Under the finite-sites model, \( F_i^m(\theta_1, \theta_o, \tau) \) provides the distribution of the number of mutations having occurred during the evolution of a random pair of genes. Note that this number can be equal to or larger than the number of observed differences, depending on whether the same site has been hit several times by mutations or not. In this case, the expected mismatch probability distribution noted by \( F^m_i(\theta_1, \theta_o, \tau) \) can be obtained by taking into account those cases where \( j \) mutations \( (j \geq i) \) have led to exactly \( i \) differences, as

\[
F^m_i(\theta_1, \theta_o, \tau) = \sum_{j=i}^{\infty} F^m_j(\theta_1, \theta_o, \tau) H_m(i, j),
\]

where \( H_m(i, j) \) is the conditional probability of observing \( i \) differences given \( j \) mutations having occurred in the ancestry of two sequences of length \( m \). We now describe how to obtain these conditional probabilities \( H_m(i, j) \) starting with \( m = 1 \) and extending it to a sequence of arbitrary length.

Onesite case: We first solve the problem for one site \( (m = 1) \) assuming Kimura’s two-parameters model of mutation (Kimura 1980) with arbitrary relative transi-
tion (s) and transversion (v) rates subject to \( s + v = 1 \). Conditional on the number of mutations, we consider the mutation process as a random walk between the nucleotides A, C, G, and T. This Markov process has the single-step transition matrix

\[
M = \begin{pmatrix}
A & C & G & T \\
0 & v/2 & 0 & v/2 \\
0 & 0 & s/2 & 0 \\
v/2 & s/2 & 0 & 0 \\
v/2 & s/2 & 0 & 0 \\
\end{pmatrix}
\]

where the elements of each row add up to one. The \( j \)th power of the matrix \( M \) can be used to describe the impact of \( j \) mutations at that site. A closed-form expression for \( M^j \) can be conveniently obtained by a diagonalization of \( M \), as \( M = V D V^{-1} \), where \( D = \text{diag} \{ -s, -s, v + s, -v + s \} \) and \( V \) is a matrix where the columns are the eigenvectors of \( M \).

We thus obtain

\[
M^j = V D^j V^{-1} = \begin{pmatrix}
\alpha_1 & \beta_1 & \gamma_1 & \beta_1 \\
\beta_1 & \alpha_1 & \beta_1 & \gamma_1 \\
\gamma_1 & \beta_1 & \alpha_1 & \beta_1 \\
\beta_1 & \gamma_1 & \beta_1 & \alpha_1 \\
\end{pmatrix}
\]

where \( \alpha_i = \frac{1}{2}(1 + 2(-s) + (s - v)), \beta_i = \frac{1}{2}(1 - (s - v)) \), and \( \gamma_i = \frac{1}{2}(1 + 2(-v) + (s - v)) \). The diagonal terms of \( M^j \) here all equal to \( \alpha_i \) represent the probability of returning to the initial state after \( j \) mutations. It thus follows that

\[
H^j(0, j) := P(\text{no difference} | j \text{ mutations}) = \alpha_i,
\]

\[
H^j(1, j) := P(\text{one difference} | j \text{ mutations}) = 1 - \alpha_i.
\]

Multisite case, homogeneous mutation rates: Instead of deriving an explicit equation for \( H^n(i, j) \), when \( m > 1 \) we can compute these probabilities numerically using a recurrence equation, as shown below. Let us suppose that we have already derived the probability \( H^{n-i}(i, j) \) and that we want to study the case for an additional site and thus derive \( H^n(i, j) \). Suppose that \( l \) mutations have occurred at the \( m \)th site and that the \((j-1)\) remaining mutations have occurred at the \( m-1 \) other sites. The probability of observing overall \( i \) differences will depend upon whether we observe one or no difference at the \( m \)th site. With probability \( P_1 \), one difference will be observed at the \( m \)th site and \((i - 1)\) at the \((m - 1)\) other sites, and with probability \( P_2 \), all \( i \) differences will be observed among the \((m - 1)\) other sites. Therefore,

\[
P_1 = H^j(1, l)H^{n-i}(i - 1, j - l),
\]

\[
P_2 = H^j(0, l)H^{n-i}(i, j - l),
\]

where we admit that \( H^{n-i}(-1, j - l) = 0 \). Summing over all possible \( l \) values and multiplying by the probability that \( l \) mutations occur at the additional site, we finally have

\[
H^n(i, j) = \sum_{l=0}^{j} b(l, j, p)(P_1 + P_2)
\]

where \( b(l, j, p) = (j!/l!((j - l)!) \cdot p(1 - p)^{j - l} \) is the binomial probability with parameter \( p = 1/m \).

The mismatch distribution under a two-rates finite-sites model: Mutation rate heterogeneity arises when the mutation rates are not equal for all nucleotide sites. The simplest form of heterogeneity to be considered is a two-rates mutation model, where we make the distinction between fast and slow sites. As most mutations accumulate at a small number of fast sites, convergent or reverse mutations can become quite common. The consequence of this type of heterogeneity on the pattern of diversity has been studied in the case of the control region of human mitochondrial DNA (Wakeley 1993; Yang 1994, 1996; Bertorelle et al. 1995). Following Yang (1996), who inferred the number of segregating sites in a stationary population under the finite-sites model for two classes of mutation rates, we can modify Equation 3 by considering that \( m_1 \) of the \( m \) sites are fast and that \( m_2 \) are slow. In this case, we have to take into account all possible ways of partitioning the \( i \) differences among the slow and fast regions. If we assume that mutations occur independently, the probabilities of these partitions simply follow a binomial distribution. The expected mismatch probability distribution is thus given by

\[
F_{m_1 m_2}(\theta_r, \theta_o, \tau) = \sum_{l=0}^{j} F_r^-(\theta_r, \theta_o, \tau) \sum_{i=0}^{\infty} \sum_{k=0}^{\infty} b(k, j, p)
\times H^n(l, k) H^n(i, j-k),
\]

where \( b(k, j, p) \) is the same binomial probability function as in Equation 6 but with parameter \( p = m_1r/(m_1r + m_2) \) being the conditional probability that a mutation will hit one of the \( m_1 \) fast mutating sites, and \( \tau \) is here the ratio of fast and slow mutation rates. Note that Equation 7 is the equivalent to Yang’s (1996) Equation 39, derived for the case of a stationary population. Yang used an infinite-alleles mutation model, stipulating that once a site is hit by one or more mutations, we observe one difference. Due to the high transition bias, this model also tends to overestimate the number of differences because it does not allow back mutations at non-segregating sites.

Multisite case, \( m \)-rates mutation model: Suppose that we have a sequence of length \( m \) and that each nucleotide has a potentially different probability \( p_i \) \((i = 1 \ldots m)\) of being hit by a mutation, subject to the condition \( \sum_i p_i = 1 \). Under this \( m \)-rates model, \( H^n(i, j) \), noted here \( H^r(i, j) \), can be also obtained by the recurrence
relation given in (6) except that the parameter \( p \) is now equal to \( p_n / \sum_{i=1}^{n} p_i \). Here \( n \) is the index of the recursion step (1 \( \leq n \leq m \)), and \( p \), thus changes at every step of the recursion. It is important to note that, unlike the homogenous mutation rate case, the intermediate re-
currence matrices are here meaningless, and only the last matrix obtained by this recurrence is correct for the
required heterogeneity pattern. Under this \( m \)-rates mutation model, the expected mismatch distribution is not given in a more complex form than in the constant
mutation model, the expected mismatch distribution is a coalescent process with superimposed mutations.

\[
F^m_{ij}(\theta_i, \theta_0, \tau) = \sum_{i+j=n} F^0_{ij}(\theta_i, \theta_0, \tau) H^m(i,j).
\]

Note that the two-rates model is a special case of the present model and that it could be treated similarly. In
that case, the double summation of Equation 7 could be condensed into the last term of Equation 8.

The mismatch distribution for gamma-distributed mutation rates: A gamma distribution of mutation rates can be
seen as a special case of the \( m \)-rates mutation model. The parametric model that we used here is a coalescent
process with superimposed mutations. We adapted the coalescent simulation program from
Hudson (1990) to generate \( B \) samples of DNA se-
quences according to the estimated parameters \( \theta_i, \theta_0, \) and \( \tau \). For each of the \( B \) simulated data sets, we applied
our estimation procedure according to Equation 3, 7,
or 8 to evaluate \( B \) bootstrapped values \( \theta_i^*, \theta_0^*, \) and \( \tau^* \). For a given confidence level \( \alpha \), the approximate limits of the
confidence interval were obtained as \( \alpha / 2 \) and \( 1 - \alpha / 2 \) percentile values (Efron and Tibshirani 1993, p. 168).
It is important to emphasize that this form of parametric bootstrap assumes that the data are distrib-
uted according to the sudden expansion model.

Testing the validity of the sudden expansion model: We tested the hypothesis that the observed data fitted
the sudden expansion model defined by the estimated parameters using the same parametric bootstrap
approach as described above. We used here SSD defined in
Equation 9 as a test statistic. We obtained its distribution
under the hypothesis that the estimated parameters are
the true ones by simulating \( B \) samples around the estimated
parameters. As before, we reestimated each time new parameters \( \theta_i^*, \theta_0^*, \) and \( \tau^* \) and computed their associated
sums of squares SSD_{sim}. The \( P \) value of the test is there-
fore approximated by

\[
P = \frac{\text{number of SSD}_{\text{sim}} \geq \text{SSD}_{\text{obs}}}{B}.
\]

To check the accuracy of this procedure, we generated
1000 random data sets for the parameters \( \theta_i = 1, \theta_0 =
1000, \) and \( \tau = 3 \) under a two-rates model with 270 slow
sites and 30 fast sites mutating 20 times faster than the
slow sites, which corresponds to the simulation condi-
tions of the top of Table 1. The simulated distribution of the \( P \) values for these parameters was almost uniform
between 0 and 1 (data not shown), suggesting that the
SSD statistic provides a valid test of the sudden expansion model.

RESULTS

We show in Figure 2 the theoretical mismatch distributions and the demographic parameters estimated using different methods and mutation models for the two hypervariable segments of Mandenka mtDNA control region (Graven et al. 1995). Except for the curve based on Rogers’ method of moments (Rogers 1995), the shapes of the theoretical mismatch distributions are very similar to each other and provide no real indication of the validity of the underlying model. Note also that a much better fit is found for HV1 than for HV2. While the shape of the best-fitted mismatch does not seem to depend much on the mutation model, the values of the estimated parameters do change quite extensively, especially the expansion time $\tau$, which shows larger values for finite-sites models than for the infinite-sites model. This is of course due to the fact that several mutations can accumulate at a given site in finite-sites models and that a longer evolutionary time is necessary to lead to the same number of observed differences. The magnitude of the expansion is also found to be smaller in finite-sites models, in agreement with previous simulation results (Aris-Brosou and Excoffier 1996).

In Figure 3, we show the expected mismatch distributions fitted for the Turkana sample (Watson et al. 1996) according to a finite-sites two-rates model (Figure 3a) and to a finite-sites gamma distribution model (Figure 3b). We also report the average mismatch distributions obtained from 5000 simulations performed according to the estimated parameters. Simulated and expected mismatch distributions are found to be in very good agreement, motivating the use of simulations to get empirical confidence intervals around the parameters.

To check if these confidence intervals have good coverage properties (i.e., the true parameters should be included in the confidence interval with a probability $1 - \alpha$), we performed a series of simulations for a set of predefined parameters. For a given set of parameters $\theta_0$, $\theta_1$, and $\tau$, we simulated 1000 data sets from which we estimated the parameters $\hat{\theta}_0$, $\hat{\theta}_1$, and $\hat{\tau}$. For each set of estimated parameters, we simulated 1000 additional data sets from which new values $\hat{\theta}_0^*$, $\hat{\theta}_1^*$, and $\hat{\tau}^*$ were estimated. The distribution of these 1000 bootstrap values was used to evaluate the lower and upper limits of a $100(1 - \alpha/2)$% confidence interval around the $\theta_0$, $\theta_1$, and $\tau$ values as the $\alpha/2$ and $1 - \alpha/2$ percentiles of the distribution, respectively. The results of these analyses are shown in Tables 1 and 2 for different types of mutation rate heterogeneity. It can be seen that the only parameter for which the bootstrap confidence interval has a good coverage is $\tau$, as the proportion of the times the true value is outside the confidence interval is approximately equal to the significance level $\alpha$. Note, however, that the confidence interval is not well centered, as the true values outside the confidence interval are always found on the left of the distribution. The bootstrap confidence intervals for $\theta_0$ and $\theta_1$ are much too broad (the true value of the parameter is found too often within the empirical confidence interval). In Figure 4, we plot the distributions of the two statistics $x - x'$ for $x = \tau$ and for $x = \theta_0$. To generate true
confidence intervals, the bootstrap percentile method requires that these two distributions be identical (see, e.g., Rice 1995, p. 271). We can see that it is only approximately the case for \( \tau \) but not for \( \theta_0 \) or for \( \theta_1 \) (data not shown for \( \theta_1 \)). Moreover, the estimations of \( \tau \) and \( \theta_0 \) appear much less biased than that of \( \theta_1 \). Note that a possible explanation of the bias for \( \theta_1 \) is mentioned in Rogers and Harpending (1992), who rightly point out that it is difficult to distinguish between a large expansion and a very large expansion, thus generating an upward bias in \( \theta_1 \). Interestingly, the parameters of old expansions (\( \tau = 9 \)) seem more precisely recovered than those of relatively recent expansions (\( \tau = 3 \); Tables 1 and 2). These results confirm several points. First, the time of a sudden expansion (\( \tau \)) can be adequately recovered from the data with approximately valid confidence intervals. Second, the estimate of the initial population size appears quite well recovered, but with an overly conservative confidence interval due to a too large upper bound (see Figure 4). Finally, both the estimate of the population size after the expansion and its confidence interval cannot be adequately recovered from the mismatch distribution. However, as the bias in \( \theta_1 \) appears mainly due to an overly large upper bound, the lower bound for \( \theta_1 \) could be useful even though still underestimated.

For illustration purposes, we present in Figure 5 the expected mismatch distributions of a few human samples analyzed for HV1 or HV2, as well as the limits of a 95% confidence interval around the mismatch distributions. Despite an obvious lack of goodness-of-fit for some distributions, the adequacy of the sudden expansion model could only be rejected for the Ngoebe HV2 sample (SSD P value = 0.007). For the other samples, random mismatch distributions generated by simulations lead to SSD values larger than the observation in >5% of the cases, making the observed mismatch distributions compatible with the estimated parameters.

**DISCUSSION**

In this study, we extend the model of Rogers and Harpending (1992) to estimate the parameters of a sudden stepwise demographic expansion by explicitly taking into account a possible heterogeneity of mutation rates. Contrary to previous claims (Rogers 1992; Rogers et al. 1996), we find that the estimated values of the parameters and their confidence intervals are quite sensitive to departure from the infinite-sites model. For instance, the estimated values of the expansion time (\( \tau \)) shown in Figure 2 for the Mandenka population are found, respectively, 9 and 20% larger for HV1 and HV2 when using a model with gamma-distributed mutation rates than for the infinite-sites model. Even though our methodology appears computationally more intensive, it thus seems justified to take into account the known departures from the infinite-sites model. The present approach does not allow us to retrieve all the parameters of a demographic expansion with the same efficiency. As shown in Tables 1 and 2, the expansion time (\( \tau \)) and the initial population size (\( \theta_0 \)) are the only parameters that can be estimated without much bias and with reasonable precision, while the estimation of \( \theta_1 \) is clearly biased upward. The confidence intervals obtained from the parametric bootstrap approach are fairly estimated only for the expansion time \( \tau \), while those for the population sizes are clearly too large and thus overly conservative. This implies that the magnitude of the expansion cannot be precisely recovered by the present approach. This is understandable because once the expansion is sufficiently large, very few coalescent events (if any) will have occurred between the present time and the beginning of the expansion. As it is the accumulation of those coalescent events that can
TABLE 1
Proportion of noninclusion of the exact parameter value: two mutation rates model

<table>
<thead>
<tr>
<th>Parameters of the two-rates mutation model</th>
<th>Demographic parameters</th>
<th>True values</th>
<th>Average estimated value</th>
<th>Significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>m₁</td>
<td>m₂</td>
<td>r</td>
<td>τ</td>
<td>True values</td>
</tr>
<tr>
<td>270</td>
<td>30</td>
<td>20</td>
<td>3</td>
<td>3.726 (1.195)</td>
</tr>
<tr>
<td>290</td>
<td>10</td>
<td>50</td>
<td>3</td>
<td>3.854 (1.493)</td>
</tr>
<tr>
<td>270</td>
<td>30</td>
<td>20</td>
<td>9</td>
<td>9.253 (1.037)</td>
</tr>
<tr>
<td>290</td>
<td>10</td>
<td>50</td>
<td>9</td>
<td>9.286 (1.072)</td>
</tr>
</tbody>
</table>

* m₁ is the number of slow sites; m₂ is the number of fast sites; r is the ratio of mutation rates between the fast and the slow sites.

Note: Average value of the parameters estimated from 1000 coalescent simulations around the true values (standard deviations shown in parentheses).

provide some information on the present population size, there will often be too few of them to get a reliable estimate of the present size, which will also tend to be overestimated. The present parametric bootstrap approach for defining the confidence intervals differs somewhat from that described in previous studies (Rogers 1995; Rogers et al. 1996). The previous approaches consisted of finding a set of values of the demographic parameter compatible with the observed data. The “compatibility” criterion was a statistic of goodness-of-fit (mean absolute error) between the observed and the theoretical mismatch distribution. A set of demographic parameters θ₀, θ₁, and τ was declared compatible if the goodness-of-fit statistic fell within a 95% confidence interval obtained by simulation. While this approach seems valid, it requires much heavier computations than ours if one wants to adequately explore the space of possible parameters, as a series of simulations needs to be carried out for each set of parameters. Moreover, the potential impact of the chosen goodness-of-fit statistic on the results and the reliability of the confidence intervals has not been addressed. The fact that the effective popula-

TABLE 2
Proportion of noninclusion of the exact parameter value: gamma distribution of mutation rates

<table>
<thead>
<tr>
<th>Parameters of the mutation model</th>
<th>Demographic parameters</th>
<th>True values</th>
<th>Average estimated value</th>
<th>Significance level</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>α ²</td>
<td>τ</td>
<td>True values</td>
<td>Average estimated value</td>
</tr>
<tr>
<td>300</td>
<td>0.26</td>
<td>3</td>
<td>3.514 (0.920)</td>
<td>0.014</td>
</tr>
<tr>
<td>300</td>
<td>0.13</td>
<td>3</td>
<td>3.457 (0.897)</td>
<td>0.007</td>
</tr>
<tr>
<td>300</td>
<td>0.26</td>
<td>9</td>
<td>9.210 (1.073)</td>
<td>0.013</td>
</tr>
<tr>
<td>300</td>
<td>0.13</td>
<td>9</td>
<td>9.219 (1.146)</td>
<td>0.006</td>
</tr>
</tbody>
</table>

* Shape parameter of the gamma distribution.

Note: Average value of the parameters estimated from 1000 coalescent simulations around the true values (standard deviations shown in parentheses).
to small changes in the mismatch distribution (see also Rogers 1997). Conversely, small differences in the shape of the observed mismatch distribution will profoundly affect the values of the estimated parameters. As the estimation of the expansion time $\tau$ depends essentially on the mean of the mismatch distribution (Rogers and Harpending 1992), while the other parameters $\theta_0$ and $\theta_1$ depend on higher moments of the distribution, those latter two parameters are more likely to be affected by the stochasticity of the genealogical process than $\tau$. This is in keeping with our simulations, which show that the expansion time is usually quite well recovered from the mismatch distribution (Tables 1 and 2).

Even though we have refined the mutation model for mtDNA sequences, one can see that the theoretical mismatch distributions do not always perfectly fit with the observed distributions (Figures 2, 3, and 5). We can see two reasons explaining this discrepancy.

First, the single stepwise expansion model may be inadequate for some populations. Alternative population expansion models such as exponential growth or logistic growth could be more realistic that the stepwise growth used in this study (Polanski et al. 1998), but as long as the magnitude of the expansion is large and we start from a small population, they should lead to results very similar to those provided here (Rogers and Harpending 1992; Rogers 1997). It seems more likely that demographic scenarios very different from population growth may explain these discrepancies. Population contractions may indeed have occurred in some populations and could explain the rejection of the sudden expansion model. A population crash could have occurred in the Ngöebe population from Panama, as well as in other native Amerindian populations, where the hypothesis of sudden expansion is not supported, as in the Kuna from Panama ($P = 0.05, HV1$), the Hueter from Costa-Rica ($P = 0.026, HV2$), or the Mapuche from Argentina ($P = 0.009, HV2$). Additional evidence for a recent population contraction comes from the observation of large positive values for Tajima's $D$-statistics (Tajima 1989) in those populations (results not shown), which are expected in the case of a recent population bottleneck (Tajima 1993). Note also that other factors like admixture events, population substructure, or inbreeding could all affect the shape of the mismatch distribution but to an extent that has not yet been quantified.

Second, the probabilities derived in Equations 1 and 2 and their derivatives apply to a pair of genes chosen at random from the population, while they are applied here to a random pair chosen from the sample. However, pairs drawn from the sample are not independent due to the shared portions of their genealogy. In populations having gone through a recent and large expansion, the internal branches are very short due to the star-like structure of the tree (Slatkin and Hudson 1991; Fu 1997), and a very few mutations will accumulate...
Figure 5.—Empirical 95% "confidence intervals" for the mismatch distribution in four human populations analyzed for mtDNA control region. In all cases, a finite-sites mutation model was used, assuming gamma distribution of mutation rates ($\alpha = 0.26$ for HV1 and $\alpha = 0.13$ for HV2). For the Ngoebe sample (Kolman et al. 1995), the hypothesis of sudden expansion is rejected by the SSD test ($P = 0.007$). The SSD $P$ values for the other samples are as follows: English HV1 (Piercy et al. 1995), 0.764; Mandenka HV2 (Graven et al. 1995), 0.0652; !Kung HV1 (Vigilant et al. 1991), 0.1368.

on those branches. In that case, the correlation between the number of pairwise differences ($d_{ij}$) will only be due to shared external branches (i.e., $d_{12}$ and $d_{13}$ will be correlated due to the shared lineage leading to sequence 1, but $d_{12}$ and $d_{13}$ should be almost independent), and our derivations should better hold at the sample level. On the other hand, for stationary populations or relatively small or remote expansions, some coalescent events will occur before and after the expansion. The internal branches will be longer and have a large associated variance. Those equations, while still being correct for a single pair of genes, will thus not allow us to get the sample distribution of pairwise differences as they do not take into account the covariance of pairwise differences. Therefore, the present method is not expected to recover the parameters of a demographic expansion efficiently unless the expansion has been very large.

Although mismatch distributions carry some information on the shape of the underlying gene genealogy and coalescent process, other aspects of molecular diversity are not explicitly taken into account by this approach. It has been shown that demographic parameters recovered from the mismatch distribution did not allow the correct prediction of the number of observed polymorphic sites (Bertorelle and Slatkin 1995) or of the distribution of mutation frequencies (Wakeley 1993, p. 614) for human mtDNA. This could either be due to departure from the infinite-sites mutation model or from the proposed simple demographic model. Even after having introduced a more realistic mutation model, we still find that the observed number of segregating sites is not always in agreement with the distribution obtained from simulations based on the estimated demographic parameters. For instance, considering the mismatch distributions shown in Figure 5, even if we get a perfect fit for the English HV1 sample, the estimated parameters lead on average to far fewer segregating sites than observed, although not significantly so ($S_{\text{obs}} = 67$; $S_{\text{mean}} = 60.6$; $SD(S) = 6.2$; $P = 0.872$). Interestingly, the Mandenka sample presents a significant lack of segregating sites for HV2 as compared to the estimated expansion conditions ($S_{\text{obs}} = 27$; $S_{\text{mean}} = 38.0$; $SD(S) = 5.0$; $P = 0.014$). This discrepancy could be explained by a very large heterogeneity of mutation rates in HV2 for this population, but it seems difficult to understand how and why the structural and functional constraints that are supposed to shape the heterogeneity of mutation rates (Wakeley 1993, p. 614) could differ between populations. Additional studies on that matter would nevertheless be needed to exclude this possibility.

To get absolute values for the demographic parameters inferred using the present approach, one should get an estimation of the substitution rate at the nucleotide level. The real value of mutation rate in humans has recently been the subject of an intense debate between those advocating the use of a phylogenetic mutation rate ($\sim 3 \times 10^{-6}$ substitutions per site per generation of 20 yr) calibrated by the divergence between humans and chimpanzees (Jazin et al. 1998) and those studying...
the mutation process directly on pedigrees giving numbers $\sim 10$ times larger ($\sim 2.7 \times 10^3$ substitutions per site per generation; Howell et al. 1996; Parsons et al. 1997; Parsons and Holland 1998). For the present methodology to be fully beneficial, it thus seems highly necessary to get reliable estimates of mutation rates. Otherwise, the importance of taking into account more realistic mutation models would seem rather futile.

Even if the present approach is an improvement over previous methods, it seems that the use of the mismatch distribution as a summary statistic may not exploit the full potential of molecular data and that maximum-likelihood methods that take into account phylogenetic relationships between DNA sequences (e.g., Griffiths and Tavaré 1994; Kuhner et al. 1995; Tavaré et al. 1997; Kuhner et al. 1998; Weiss and von Haeseler 1998) would be needed to get more reliable estimates of demographic parameters. However, considering the fact that these methods are extremely computer-intensive when heterogeneity of mutation rates is considered, the present approach may still be useful in most practical purposes.

We are grateful to Ziheng Yang, Monty Slatkin, and two anonymous reviewers for their helpful comments on the manuscript and to Simon Tavaré for some useful advice. We thank Alan Rogers for providing us with C source code for generating gamma-distributed random variates. We are grateful to André Langaney for his support throughout this work. This study was made possible by Swiss National Fund grant nos. 32-04705.96 and 31-039847.93. A program for computing the estimated demographic parameters from the mismatch distribution and performing the tests described in this article is available from S.S. upon request. These programs are also integrated into the Arlequin software, available on http://anthropologie.unige.ch/arlequin/.

LITERATURE CITED


Tajima, F., 1996 The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. Genetics 143: 1457-1465.

Communicating editor: G. B. Golding