Theoretical Population Biology **53**, 256–271 (1998)
Article No. TP981363

TPB

Microsatellite Behavior with Range Constraints: Parameter Estimation and Improved Distances for Use in Phylogenetic Reconstruction

David D. Pollock[1]
Interval Research Corporation, Palo Alto, California 94304;
and Department of Mathematical Biology, National Institute for Medical Research,
Mill Hill, London NW7 1AA, United Kingdom

Aviv Bergman[2]
Interval Research Corporation, Palo Alto, CA 94304

Marcus W. Feldman
Interval Research Corporation, Palo Alto, California 94304;
and Department of Biological Sciences, Stanford Unviversity, Stanford, California 94304

and

David B. Goldstein
Interval Research Corporation, Palo Alto, CA 94304; and Department of Zoology,
University of Oxford, OX1 3PS, United Kingdom

Received April 1, 1997

A symmetric stepwise mutation model with reflecting boundaries is employed to evaluate microsatellite evolution under range constraints. Methods of estimating range constraints and mutation rates under the assumptions of the model are developed. Least squares procedures are employed to improve molecular distance estimation for use in phylogenetic reconstruction in the case where range constraints and mutation rates vary across loci. The bias and accuracy of these methods are evaluated using computer simulations, and they are compared to previously existing methods which do not assume range constraints. Range constraints are seen to have a substantial impact on phylogenetic conclusions based on molecular distances, particularly for more divergent taxa. Results indicate that if range constraints are in effect, the methods developed here should be used in both the preliminary planning and final analysis of phylogenetic studies employing microsatellites. It is also seen that in order to make accurate phylogenetic inferences under range constraints, a larger number of loci are required than in their absence.

Key Words: **microsatellite; evolution; phylogenetics; least squares.**

[1] Current address: Department of Integrative Biology, Valley Life Sciences Building, University of California, Berkeley, California 94720-3140. E-mail: d-polloc@nimr.mrc.ac.uk.
[2] Current address: Center for Computational Genetics and Biological Modeling, Stanford University, Stanford, California 94305-5020.

0040-5809/98 $25.00

1. INTRODUCTION

Microsatellites have fully realized their early promise in building genetic maps (e.g., Dib *et al.* (1996)) and in estimating relatedness among individuals (Strassman *et al.*, 1996). There is a striking absence, however, of successful application to phylogenetic reconstruction. This almost certainly results from a combination of two complicating factors: (1) the existence of range constraints limiting the size of microsatellite alleles, and (2) the degradation of microsatellite loci over time. Preliminary studies indicate that the latter can sometimes make it difficult to find microsatellites which are polymorphic in multiple species (Shriver *et al.*, 1995; Garza *et al.*, 1995; Goldstein and Clark, 1996), although in other instances polymorphic microsatellites can last over considerable phylogenetic divergences (Fitzsimmons *et al.*, 1995; Rico *et al.*, 1996). In order to facilitate the use of microsatellites in phylogenetic reconstruction, the dependence of the microsatellite degradation rate on microsatellite type and genomic location should be studied systematically. Our primary concern in this paper, however, is range constraints, whose effects may often be realized before those of degradation.

A number of authors have recently emphasized that range constraints critically influence the utility of microsatellite loci (Garza *et al.*, 1995; Feldman *et al.*, 1997; Nauta and Weissing, 1996; Slatkin, 1995b. Goldstein *et al.*, 1995a; Goldstein *et al.*, 1995b; Zhivotovsky *et al.*, 1997). For example, one recently introduced distance, $(\delta\mu)^2$, the squared separation between population allelic means, is extremely biased under range constraints (Goldstein *et al.*, 1995b). Feldman *et al.* (1997) provided an analytical characterization of stepwise mutations under range constraints and developed a less biased distance, D_L, under the assumption of no variation among loci in range constraints and mutation rates. They also show that in the general case the appropriate correction cannot be implemented due to statistical difficulties. Computer simulations were used, however, to demonstrate that D_L is not highly sensitive to range and rate variation. Nevertheless, D_L is formally incorrect and has a non-linear expectation under these conditions, and there may be statistically appropriate distance measures that perform significantly better. Here we apply least squares procedures to improve estimates in the more realistic general case when the range and mutation rate vary among loci, and show that linearity is always improved, while accuracy can be improved dramatically under conditions of mutation rate variation. In order to implement the method, and more generally to improve our understanding of properties of the stepwise process, information about range constraints and mutation rates are required. It could be particularly useful if these properties were associated with *a priori* characteristics of a microsatellite, such as its motif size and composition. Unfortunately, almost nothing is known about range constraints beyond their existence. Furthermore, estimates of mutation rates using pedigree analyses are averages across loci, and the data are too few to allow rates at different loci to be compared (Weber and Wong, 1993).

Here we provide a detailed evaluation of how range constraints may be estimated from population data. We also compare the estimation of mutation rates obtained from allelic variances and from an iterative least squares procedure based on population divergence. The latter approach is likely to be less sensitive to violations of mutation-drift equilibrium caused by changes in population size or by selection at linked sites (Feldman *et al.*, 1997). Beyond their use in implementing improved distance measures, such methods are necessary to partition microsatellite loci into sets appropriate for specific phylogenetic problems, a practice which must become routine if microsatellites are to have wide application in interspecific phylogenetic reconstruction. Good estimators of the range and mutation rate are also needed to develop a clear understanding of microsatellite evolution.

2. RANGE ESTIMATION CONSIDERATIONS

2.1. Expected Behavior of the Distance

A recently introduced measure of genetic distance for microsatellites, $(\delta\mu)^2$, the squared difference between mean allelic scores averaged over all loci, was shown to have an expectation proportional to the time, t, elapsed since the separation of two populations in the absence of range constraints (Goldstein *et al.*, 1995b). Feldman *et al.* (1997) modeled microsatellite evolution under the symmetric stepwise mutation model (SSM) with reflecting boundaries (range constraints) and maximum mutation size of one repeat unit. They showed that in this model the expectation of $(\delta\mu)^2$ for a locus with total mutation rate 2β (assumed constant for all allelic sizes) and range R will reach a maximum of

$$M(R) = \frac{(R^2 - 1)}{6} - D_0, \qquad (1)$$

where, with N chromosomes, D_0 can be approximated by $4\beta(N-1)(1-1/R)$ when $N\beta < 1$, or calculated numerically using their Eq. (11) when $N\beta$ is large. They also showed that the rate of approach to this maximum will be the leading non-unit eigenvalue of the squared mutation matrix, namely $(1-2\beta+2\beta\cos\pi/R)^2$. For small β it is convenient to use an approximation for the expectation of $(\delta\mu)^2$:

$$E[(\delta\mu)^2] = M(R)\{1-\exp[-(4\beta-4\beta\cos(\pi/R))t]\}. \quad (2)$$

The lower curve in Fig. 1 shows the analytical expectation of $(\delta\mu)^2$ with time for an example parameter set, along with mean values for simulated population pairs, and it is clear that the expectation closely matches simulated results. It is also clear that the use of this expectation in concert with the formula of Zhivotovsky and Feldman (1995) for the standard deviation gives accurate results early on, and is not off by more than 25% for larger separation times. In these and other simulations throughout, haploid populations of a given size, N (here $N=50$), were started from a random position within a given range, R (here $R=20$), allowed to equilibrate under the mutation model with range, R, and mutation rate β (here $\beta=0.01$). This population was then split into two identical populations, also of size N, which were allowed to evolve independently of each other for a given period of time, after which relevant estimators were calculated. If appropriate, further population divergences were handled in the same manner. The simulations were generally repeated 1000 times to get accurate estimates of the mean behavior of the relevant estimators.

2.2. Range Estimation

The utility of microsatellite loci in phylogenetic reconstruction will be strongly affected by their ranges, which will not be known at the outset and need to be estimated from population data. Here we discuss issues related to such estimation. In order to take advantage of some useful results from order statistics (Arnold 1992, p. 33), we first consider the case of $N\beta$ sufficiently small that each of the sampled populations maintain a single allele at any one time. For this case, we obtain a simple analytical result for predicting the true range from the observed range of alleles. Using computer simulations,

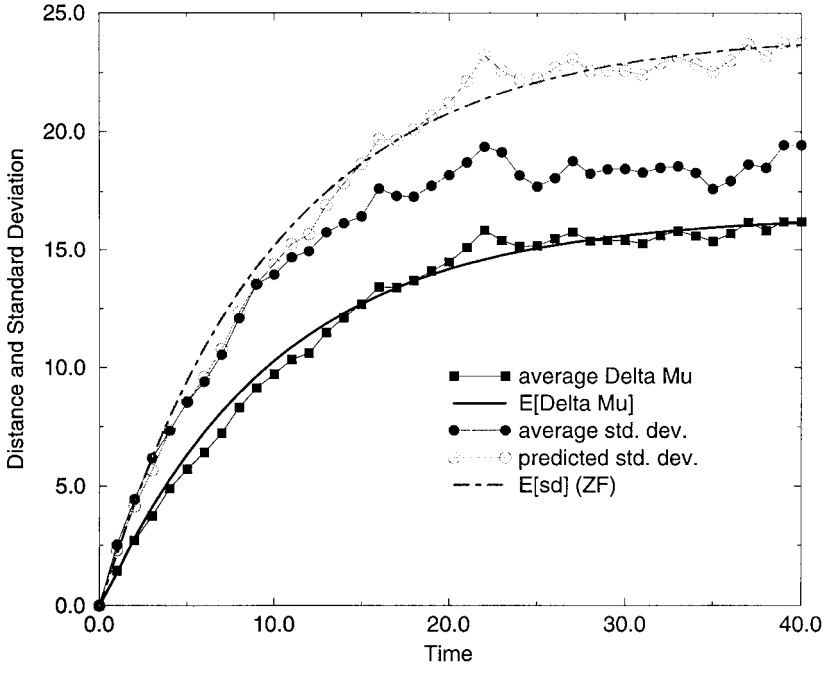


FIG. 1. Expectation and standard deviation of delta mu squared. The mean (squares) and standard deviation (filled circles) of $(\delta\mu)^2$ are shown for 1000 simulated replications where $N=50$, $R=20$, and $\beta=0.01$. The predictor of the standard deviation from the observed mean and the formula (ZF) in Zhivotovsky and Feldman (1995) is also shown (open circles). Continuous lines are the expectation of the mean as calculated from Eq. (2)(solid line), and the expectation of the standard deviation calculated from Eq. (2) and ZF (short/long dashed line). Time is measured in units of $4\beta t$, where β is the directional mutation rate and t is the number of generations. In this and all other simulations, populations of size N were allowed to equilibrate, and were then duplicated at $t=0$, after which they evolved independently according to the SSM model with reflecting boundaries and the parameters given.

we then find a simple correction in the presence of allelic variance, which can be large in microsatellites.

Let K be the number of sampled populations, and let the width, W_K, stand for the difference between the observed maximum and observed minimum scores among all the K populations. If the allelic score from each of the K populations is distributed as a continuous uniform random variable on $[a, R+a-1]$, where a is an unknown minimum, then the expectation of the width is

$$E[W_k \mid R] = (R-1)\frac{K-1}{K+1}. \quad (3)$$

Note that although the unknown minimum, a, may be physically bounded (for example, it cannot be less than one repeat), this information is not used in the above equation: the actual minimum is not estimated. From Eq. (3), a reasonable estimator for R in this case is

$$\hat{R} = \frac{K+1}{K-1}(W_k) + 1. \quad (4)$$

This assumes that the numbers of repeats (scores) are uncorrelated across populations (a reasonable assumption if populations have diverged sufficiently). The difference between this estimator and a more complex estimator assuming a discrete uniform distribution is negligibly small (data not shown), and in any case will be accounted for in the subsequent correction for allelic variation. The accuracy of Eq. (4) in estimating R was evaluated by 10,000 replicated computer simulations of points drawn from a discrete uniform distribution on $[1, R]$, and for a variety of R and K was generally found to have a lower bias and a smaller mean square error [$MSE = BIAS^2 + Var$; (Rice, 1995)] than an uncorrected estimate (data not shown).

Microsatellite mutation rates are often sufficiently high that the variance in allele size within populations must be taken into account in estimating the range. The expectation of the variance in allele size under the unconstrained SSM model in a population of N chromosomes is $2(N-1)\beta$ (Moran, 1975). The expectation under the constrained model can be evaluated numerically (Nauta and Weissing, 1996; Feldman *et al.*, 1997). For $N\beta/R$ sufficiently small the variance is still proportional to $(N-1)\beta \approx N\beta$ (Feldman *et al.*, 1997), and it is clear that when $N\beta$ is moderately large the biases of the uncorrected estimates of the range (M_k and W_k) will be less than would be expected from the analysis of the model assuming no variation within populations. Thus, the statistical corrections developed in the absence of variation are too large in the presence of variation, assuming that the difference between the grand minima and maxima is used to estimate the range. The discrepancy is small for $N\beta$ less than one, but as $N\beta$ becomes larger this discrepancy increases to the point where a correction based only on the uniform distribution of points would result in an estimate significantly greater than the true range.

Here we estimate the effect of variation using computer simulations. To do so, we compare the average observed width for populations which have significant allelic variation to the corresponding expectation based on a model without variation (Eq. (3)). We will henceforth call the difference between these the observed minus expected width, or Δ_W. Both the number of individuals sampled, n, as well as the allelic variance, V, influence Δ_W. Fortunately, computer simulations of populations with different amounts of allelic variation (1000 replicates per condition) show that Δ_W is independent of the range, only slightly dependent on K, and is linearly dependent on $\sqrt{N\beta}$ (an approximate numerical substitute for the square root of the expected allelic variance). Fig. 2 shows values of Δ_W for two different sample sizes and two values of R for a series of 14 values of $\sqrt{N\beta}$ between 0.0 and 2.23. Also shown are the regression lines. As long as R is sufficiently large, Δ_w is not dependent on R and there appears to be a linear relationship between Δ_w and $\sqrt{N\beta}$. The regression lines for different values of R are virtually identical (Fig. 2: for $n=20$, the slopes for $R=40$ and $R=50$ are 2.16 and 2.15, the intercepts are 0.68 and 0.70, and the correlation coefficients are 0.995, and 0.991; for $n=5$, the slopes for $R=40$ and $R=50$ are 1.51 and 1.49, the intercepts are 0.75 and 0.72, and the correlation coefficients are 0.983 and 0.980). Fig. 2 also illustrates how the slope of the relationship between $\sqrt{N\beta}$ and Δ_w is dependent on the sample size, n. Note that the intercept is independent of n since there is no benefit from increased sampling in the absence of allelic variation. The intercept is nonzero (but small) because of the assumption of a continuous rather than discrete distribution in the calculation of $E[W_K]$. The dependence of Δ_w on K is slight, but discernible (data not shown). Note that the linear relationship shown in Fig. 2 breaks down when the square root of the expected allelic variance is on the order of R, but these are precisely the conditions when a correction on the observed width is unnecessary.

The linear relation of Δ_w with $\sqrt{N\beta}$ allows for a simple empirical correction. The linear equations for various reasonable values of n and K derived from repeated sampling of 10,000 random populations are shown in Table 1. The slopes were obtained by linear regression of Δ_W on $\sqrt{N\beta}$ for values of $N\beta$ between 0.0 and 5.0, and

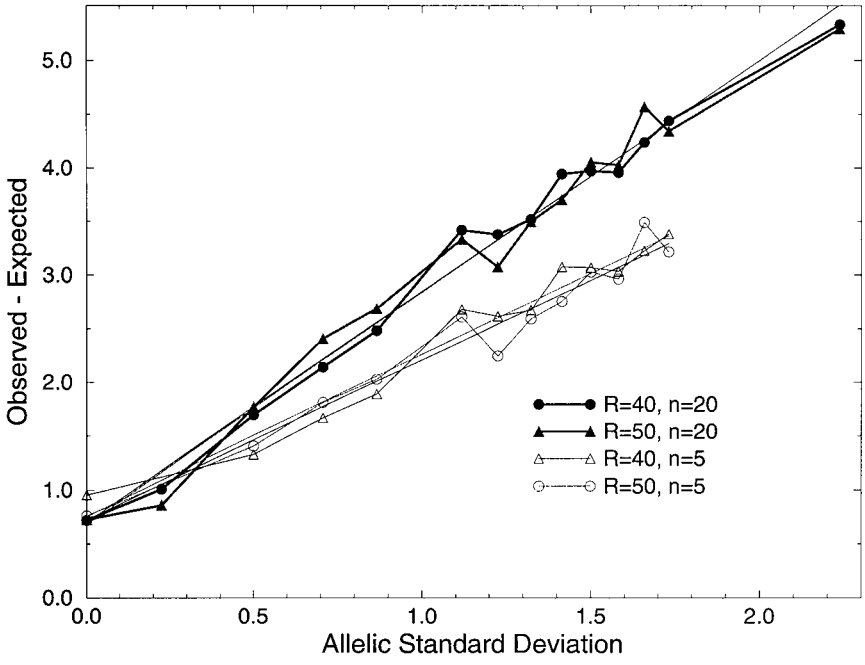


FIG. 2. Difference between observed and expected width (Δ_W). The difference between the average observed width and its expectation assuming no allelic variation (Δ_W), as a function of the square root of the product of population size and total mutation rate (which is approximately proportional to the allelic standard deviation). Points are plotted for ranges of 40 (circles) and 50 (triangles), and for 5 (open symbols) and 20 (filled symbols) individuals sampled per population. The number of pre-equilibrated populations ($N = 50$) sampled was 6, and each data point is the average of 1,000 independent simulations. Regression lines are plotted as solid lines.

TABLE

Displacement as Function of Allelic Variance

K:	1	2	3	4	5	6	8	10
				Intercept (ψ)				
	0.0	0.343	0.568	0.568	0.676	0.727	0.754	0.728
n				Slope (ω)				
2	0.912	0.913	0.817	0.920	0.840	0.798	0.758	0.738
3	1.37	1.34	1.23	1.34	1.24	1.20	1.12	1.08
4	1.65	1.61	1.50	1.59	1.50	1.43	1.34	1.28
5	1.84	1.81	1.68	1.75	1.65	1.58	1.49	1.41
6	1.96	1.95	1.83	1.88	1.79	1.69	1.60	1.52
8	2.19	2.13	2.02	2.07	1.95	1.86	1.75	1.64
10	2.32	2.27	2.13	2.19	2.07	1.97	1.83	1.74
12	2.43	2.36	2.23	2.27	2.15	2.05	1.90	1.80
14	2.50	2.44	2.30	2.34	2.21	2.11	1.95	1.85
16	2.57	2.50	2.35	2.39	2.26	2.16	2.00	1.88
18	2.612	2.55	2.39	2.43	2.31	2.20	2.03	1.91
20	2.651	2.58	2.43	2.47	2.34	2.23	2.06	1.93

Note. **Slopes and intercepts for linear relationship of Δ_W with allelic standard deviation**. Slopes ($\omega_{K,n}$) and intercepts (ψ_K) for use in Eq. (10). Predictors of the linear parameters were calculated for all combinations of the number of populations sampled ($K = 1$, 2, 3, 4, 5, 6, 8, 10) and the number of individuals sampled ($n = 2$, 3, 4, 5, 6, 10, 12, 14, 16, 18, 20). Slope and intercept were obtained by regression of Δ_W versus $\sqrt{N\beta}$ (used as a substitute for allelic standard deviation) in 1,000 repeated samples. The intercepts for each K are average of all intercepts obtained with different n and that K.

the intercept is the average intercept calculated for all values of n. This suggests that the expected value of the width, W_K, should be modified from Eq. (3) as follows.

$$E_A[W_K] = (R-1)\frac{K-1}{K+1} + \psi_K + \omega_{K,n}\sqrt{N\beta}, \quad (5)$$

where $\omega_{K,n}$ is the slope of Δ_w with allelic standard deviation, and ψ_K is the intercept, which is dependent only on K. This expression suggests the following estimator for the range,

$$\hat{R} = \frac{K+1}{K-1}(W - \psi_K - \omega_{K,n}\widetilde{\sqrt{N\beta}}) + 1. \quad (6)$$

A tilde is placed above $\sqrt{N\beta}$ in Eq. (6) because, in practice, $\widetilde{\sqrt{N\beta}}$ would be calculated from the observed mean allelic variance. Values for $\omega_{K,n}$, and ψ_K can be looked up in Table 1 for use in Eq. (6). In using Eq. (6), it is clear that when $N\beta$ is sufficiently large relative to the range no correction is needed; that is, the correction should be ignored if $\hat{R}$ is less than the observed width plus one. This neatly accounts for the conditions where the assumption of the linear relationship between the correction and $\sqrt{N\beta}$ breaks down.

Computer simulations show that an improvement in range estimation can still be achieved in this more complicated case of allelic variation. The behavior of Eq. (6) as an estimator of the range is shown in Fig. 3. When $N\beta = 0.5$ and 20 individuals are sampled per population, the reduction in MSE obtained by using Eq. (6) rather than the uncorrected width is substantial for all ranges over 10. For a larger allelic variance ($N\beta = 5.0$), the reduction becomes substantial for ranges over 30 (Fig. 3a). In Fig. 3b it can be seen that the decrease in MSE that results from reducing the sample size from 20 to 5 is moderate when $N\beta$ is 5.0, and minimal when $N\beta =$ 0.5. We may conclude: (1) it is possible to make a reasonable correction for the range estimate even when there is substantial allelic variation at microsatellite loci, and (2) extensive sampling within a population for the purpose of making such a correction appears worthwhile only when the allelic variance is much greater than one. It would be interesting to determine exactly how the sampling requirements increase with $N\beta$.

It was assumed above that populations have diverged sufficiently that their allelic scores are independent samples on the range. In reality, the non-independence of populations is an issue. One way to deal with it would be to use the correlation among allelic scores to define an "effective" K, but we have made no attempts to do so here. We suggest that when range estimation is used for fast selection of loci appropriate for a particular study, the methods outlined above should be adequate as long as an effort is made to ensure that the populations used are reasonably divergent. For phylogenetic studies, the

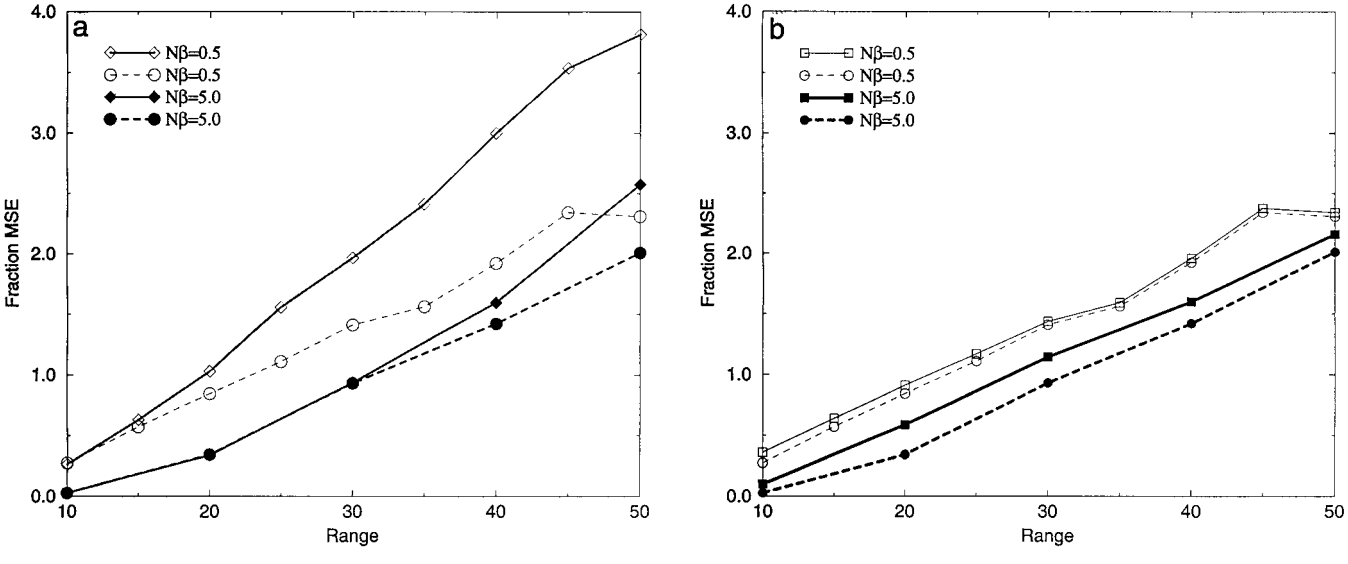


FIG. 3. Fractional mean square error in presence of allelic variation. The mean square errors (MSE) as a fraction of the range for uncorrected and corrected (using Eq. (9) and Table 1) range estimates are plotted against the range. Each point is the average of 10,000 simulations of 6 populations. In (a), the number individuals sampled was 20, and $N\beta$ was either 0.5 (open symbols) or 5.0 (filled symbols). The uncorrected averages are solid lines (squares), while the corrected versions are dashed lines (circles). In (b), only corrected averages are shown, and the number of individuals sampled was 5 (dashed lines, circles) or 20 (solid lines, squares), while the $N\beta$ was either 0.5 (upper set of lines, open symbols) or 5.0 (lower set of lines, filled symbols).

taxa of interest may well be highly correlated, and it will be necessary to include more divergent, and therefore less correlated, taxa in order to obtain accurate range estimates. The results discussed here indicate that the number of extra divergent taxa needed is not excessive. Possible interactions between the range estimates and distance estimates presented in the following sections will likely be complex, and should be the focus of future study. The performance and optimal use of these methods in evaluating real datasets are also important future subjects of research.

3. DISTANCE ESTIMATES AND MUTATION RATES

Feldman *et al.* (1997) showed that a useful distance measure when all loci have sufficiently similar ranges (R) and mutation rates (β) is

$$D_L = \log\left[\left(LM - \sum_{i=1}^{L} (\delta\mu)_i^2\right) \Big/ LM\right], \tag{7}$$

and that

$$E[D_L] = Ct, \tag{8}$$

where M is the maximum value of the distance, given in Eq. (1), L is the number of loci, C is the constant $-4\beta(1-\cos\pi/R)$, t is time, and the expectation is for replicates of the evolutionary process. As emphasized by these authors, it will often be difficult to find a large number of loci all having the same parameters. Loci with similar parameters could be clustered. Feldman *et al.* (1997) demonstrated, however, that the linearity of Eq. (7) is not, in fact, highly sensitive to variation in R and β. Nevertheless, it seems worthwhile to develop methods analytically appropriate in the case of rate and range variation. Here we introduce a least squares approach, which can improve both linearity and accuracy.

3.1. Least Squares Estimation of Distance

A simple least squares estimate of the separation time between two populations can be obtained by minimizing the sum of squares of the differences between the observed and expected values of $(\delta\mu)_l^2$ across all loci, with respect to time. That is, we minimize with respect to time

$$SS = \sum_l \{(\delta\mu)_l^2 - E[(\delta\mu)_{l,t}^2]\}^2, \tag{9}$$

with the sum over all loci and the expectation given in Eq. (2). For loci with different ranges (R_l), mutation rates (β_l), and maximal distances (M_l),

$$d(SS)/dt = 2\sum_l \left(\frac{C_l M_l}{e^{C_l t}}\right)[M_l(1-e^{-C_l t}) - (\delta\mu)_l^2], \tag{10}$$

where C_l is $[4\beta_l - 4\beta_l\cos(\pi/R_l)]$.

When all R_l, β_l, and thus M_l and C_l, are identical, $d(SS)/dt$ simplifies to

$$d(SS)/dt = 2\left(\frac{CM}{e^{Ct}}\right)\left[M\left(L - \sum_l e^{-Ct}\right) - \sum_l (\delta\mu)_l^2\right], \tag{11}$$

so that at $d(SS)/dt = 0$, we have

$$\sum_l (\delta\mu)_l^2 = LM(1-e^{Ct}). \tag{12}$$

Solution of this equation for Ct (using observed values of $(\delta\mu)^2$) leads to the distance suggested by Feldman *et al.* (1997) and shown in Eq. (7). Their distance, therefore, turns out to be the least squares estimate in the absence of rate and range variation. When the ranges and mutation rates differ across loci, an analytical expression for t is not apparent and numerical methods must be used. In practice, we iteratively increase or decrease t in Eq. (2) for use in Eq. (9) until a minimum sum of squares is reached. The incremental change in t is then reduced, and the process repeated until the change in value of Eq. (9) is less than some arbitrary small cutoff. The time value at that point is then the least squares distance, D_{LS}.

The Least Squares procedure is usually more accurate when the squared differences are weighted by the inverse variance-covariance matrix of the least squares estimates at each locus (Goldstein and Pollock, 1994; Pollock and Goldstein, 1995; Pollock, 1998). Since all loci are assumed to be independent here, the covariance between loci is zero, and the weights are simply the variances of the least squares distances at each locus σ_l^2. The individual estimates must also have the same expectation, and thus the generalized sum to be minimized with respect to time (as described above for D_{LS}) is

$$GSS = \sum_l \frac{\{(\delta\mu)_l^2 - E[(\delta\mu)_{l,t}^2]\}^2}{E[(\delta\mu)_{l,t}^2]^2\,\sigma_l^2}. \tag{13}$$

This equation is minimized as described above for D_{LS} to obtain a generalized least squares distance, D_{GLS}.

The variance at each locus can be calculated analytically by beginning with the observation that Eq. (12) can be solved for t rather than Ct, as in Eq. (7). Thus, the expectation of D_{LS} for an individual locus is

$$E[D_{LS}]_{l,t} = -\frac{\log[(M_l - (\delta\mu)_l^2)/M_l]}{4\beta_l(1-\cos[\pi/R_l])}. \qquad (14)$$

We can treat Eq. (14) as a function of $(\delta\mu)_l^2$, and taking the derivative with respect to $(\delta\mu)_l^2$ and using the formula for the variance of $(\delta\mu)_l^2$ from Zhivotovsky and Feldman (1995), we can employ the delta method (Rice, 1995, p. 149) to obtain

$$\sigma_l^2 \approx \frac{2\{E[(\delta\mu)_{l,t}^2]\}^2}{\{M_l - E[(\delta\mu)_{l,t}^2]\}^2 C^2}, \qquad (15)$$

where the expectation of $(\delta\mu)^2$ is given in Eq. (2) and C is $-4\beta[1-\cos(\pi/R)]$. In the process of iteration used to arrive at the generalized least squares estimate of time, we observed that it is better to use an initial estimate of time to calculate the expectation and variance of $(\delta\mu)^2$, which are then kept constant throughout the iterations. We use the estimate of time from D_{LS} for that purpose.

Least Squares estimation and Generalized Least Squares estimation of t were applied to a set of 100 loci with mutation rates distributed on the interval, [0.1, 0.001]. In an approach similar to Goldstein *et al.* (1995a), Pollock and Goldstein (1995), and Goldstein and Pollock (1994), two populations were allowed to diverge for a sufficiently long time that many of the loci would have been uncorrelated between the two populations, and mean distance measures and their variances for 200 replications were calculated at regular intervals during that time. D_{LS} and D_{GLS} were compared to $(\delta\mu)^2$ for all loci (an unbiased estimator of $2\beta t$ in the absence of range constraints), along with the allele-sharing distance (D_{AS}, which, in the absence of range constraints, is known to be more accurate early in the process of divergence. see Goldstein *et al.* (1995a) for definition of D_{AS}), and the log correction from Feldman *et al.* (1997), D_L. The linearity of the distances with time was assessed, and it is clear that $(\delta\mu)^2$, D_{AS}, and D_L all asymptote, while the least squares estimators are nearly linear with time (Fig. 4a). At large distances, D_{LS} may become slightly curved. This is similar to an effect previously noted in sequence-based distances, and is due to the small number of loci which are actually contributing information to the distance at these time points (Tajima, 1993). D_{GLS} is much less susceptible to this effect. With a larger number of loci, linearity is maintained for a longer time. This is the same effect demonstrated in Feldman *et al.* (1997), where adding more loci with the same R and β extends the period of linearity of their distance.

The accuracy index, $(dD/dt)\,\sigma_D^{-1}$, of a distance, D, with respect to time, t, (where σ_D is the standard deviation of the distance), is a good predictor of the utility of a distance measure in phylogenetic reconstruction (Tajima and Takezaki, 1994). For evaluation of simulation results, we may use $(\Delta D/\Delta t)\,s_D^{-1}$, where ΔD is the difference between the mean distances at one time point and another Δt later. The observed standard deviation, s_D, is calculated from data for the second time point in each accuracy calculation. For visual clarity, and because absolute accuracy is generally less important for deeper nodes, accuracy is weighted by time in the comparisons shown in Fig. 4b. As in the absence of range constraints (Goldstein *et al.*, 1995a), D_{AS} is accurate early on, but it quickly becomes the least accurate. Despite their improvements in linearity, the log distance, D_L, and the unweighted D_{LS} are less accurate than $(\delta\mu)^2$ over the entire range of conditions shown because of the number of loci used. D_{GLS}, however, is the most accurate distance after the first time interval (over which the accuracy of D_{AS} is not calculated since the distance does not start at zero), having 21–44% greater accuracy (average = 37%) than $(\delta\mu)^2$.

The same five distances were also applied to a similar set of simulations where R was evenly spaced from 10 to 48 across the 100 loci, while the mutation rate was held constant at 0.1. As in Fig. 4a and 4b where mutation rates varied, $(\delta\mu)^2$, D_{AS}, and D_L all begin to asymptote, while the least squares estimators are close to linear with time (Fig. 4c). As noted in Feldman *et al.* (1997), D_L will become increasingly linear with increasing numbers of loci, and with sufficient loci it should outperform $(\delta\mu)^2$. In terms of accuracy, however, the situation is slightly different than in the case where R varies (Fig. 4d). The relative accuracies of $(\delta\mu)^2$, D_{AS}, D_L and D_{LS} are similar to the case with varied mutation rates, but D_{GLS} is almost identical in accuracy to D_{LS}, which is slightly less accurate than $(\delta\mu)^2$ and D_L. The best explanation for this may be that when R varies, the loci which lose accuracy most quickly are those with the smallest expectation as time goes by. In contrast, when mutation rates vary, the loci with the highest mutation rates will approach their maximum more quickly than other loci, at which time they will be much less accurate and have a greater expectation than other loci. Thus, noisy loci should be given less weight at greater distances. The noisy loci in the case where R varies are already naturally and appropriately downweighted when $(\delta\mu)^2$ is taken as the average across all loci, and so the least squares approach cannot do

much better; in fact, it seems to suffer slightly due to the complexity of its calculation. It appears that when all mutation rates are known to be identical among loci, $(\delta\mu)^2$ is the most accurate distance. The accuracy of D_L, however, is very similar and we may expect that it will improve with increasing numbers of loci. D_{LS} and D_{GLS} have the best combination of accuracy and linearity. These results are not particularly sensitive to the number of loci; the same mutation rate and range parameter sets were also simulated for 20 loci (1000 replications each), and the results are almost identical except that all the curves are shifted downwards (by a factor of approximately $\sqrt{5}$) compared to the data for 100 loci (data not shown).

If both mutation rates and ranges vary among loci, the relative merits of each distance will likely depend on which parameters vary more. The increases in accuracy of D_{GLS} over a reasonable degree of mutation rate variation are somewhat greater than its relative decrease in accuracy for reasonable range variation. It is therefore likely, although not certain, to be both the most accurate and linear distance measure among those considered. The increase in the accuracy of D_{GLS} over other distances might also be improved in studies with many loci if loci can be clustered by the size of their ranges. Admittedly, the difficulty of implementation of a distance is also a consideration, and when the distances are expected to perform similarly, it might be preferable to use the simpler distances, $(\delta\mu)^2$ or D_L.

In the case where R and β are identical at all loci, all loci have the same bias and variance properties at all times. When R and/or β differ among loci, some loci will have lower accuracy than others at later time points, and will thus contribute disproportionately to the variance and any bias which might exist in the D_{LS} distance. The generalized least squares procedure compensates by downweighting those loci with higher variance, but the effective number of loci is reduced and the deeper distances have greater coefficients of variance. It is therefore preferable to select a larger number of loci with properties (large R, small β) that allow them to be informative for deeper separation times rather than to increase the assemblage randomly. It is known that a large number of loci are needed in order to obtain much resolving power even under a model without range constraints. With range constraints, resolving power will be less at deeper separation times, and even more loci will be needed to obtain the level of phylogenetic resolution available in their absence. Thus, it is important to ameliorate this effect as much as possible by preferentially selecting and devoting resources to those loci which will lose information content least rapidly.

Another complication, not directly considered here, is that if $N\beta$ is sufficiently large relative to R, then the high allelic variance will not allow divergence between populations. Although all distances will be affected to some extent, Nauta and Weissing (1996) have shown that they are differentially sensitive to variation in $N\beta$. In practice, it will be important to determine that the allelic variance is well below the maximum possible for the relevant R.

3.2. Mutation Rates

An obvious estimator for the mutation rate at a locus is the observed allelic variance ($V = D_0/2$), which is proportional to $N\beta$ if the populations are in equilibrium and $N\beta/R \ll 1$. Under these conditions the expected value of allelic variance (V) at a locus across all populations will be equal to $2\beta(1-1/R)(\bar{N}-1)$, where $\bar{N}$ is the mean (haploid) population size. Thus if R is known, $2\bar{N}\beta$ can be easily estimated from V as $2\bar{N}\beta \approx 2(\bar{N}-1)\,\beta = V/(1-1/R)$. The ratio of these averages for any two loci, $2\bar{N}\beta_i/2\bar{N}\beta_j$, will then be an estimate of the ratio of mutation rates, β_i/β_j. For larger $N\beta/R$, $N\beta$ can be estimated numerically from V as described in Feldman *et al.* (1997). Despite the attractive simplicity of this approach, it is not ideal for many reasons. First, it is heavily dependent on the assumption that all loci in all populations are at mutation-drift equilibrium. Most loci in a particular population may be out of mutation-drift equilibrium if there have been recent large changes in population size, and it can be seen from the results of Feldman *et al.* (1997) that the individual loci will only return to equilibrium at a rate approximately equal to

$$\left(1-\frac{1}{N}\right)(1-2\beta+2\beta\cos\pi/R)^2. \qquad (16)$$

Second, even if there have been no disturbances in population size, selective sweeps or balanced polymorphism will perturb allelic variances at closely linked microsatellite loci (Slatkin, 1995a). The distribution of slightly deleterious alleles throughout the genome may also influence microsatellite variation (Hudson, 1995; Charlesworth, 1994). Finally, it is known that in the absence of range constraints, the variance across loci in the allelic variance is high (Zhivotovsky and Feldman, 1995), and so the allelic variance would not be an accurate measure of relative mutation rates even if it were entirely unbiased.

Relative mutation rates at microsatellite loci may also be calculated using population divergences. A simple strategy commonly used in the study of DNA

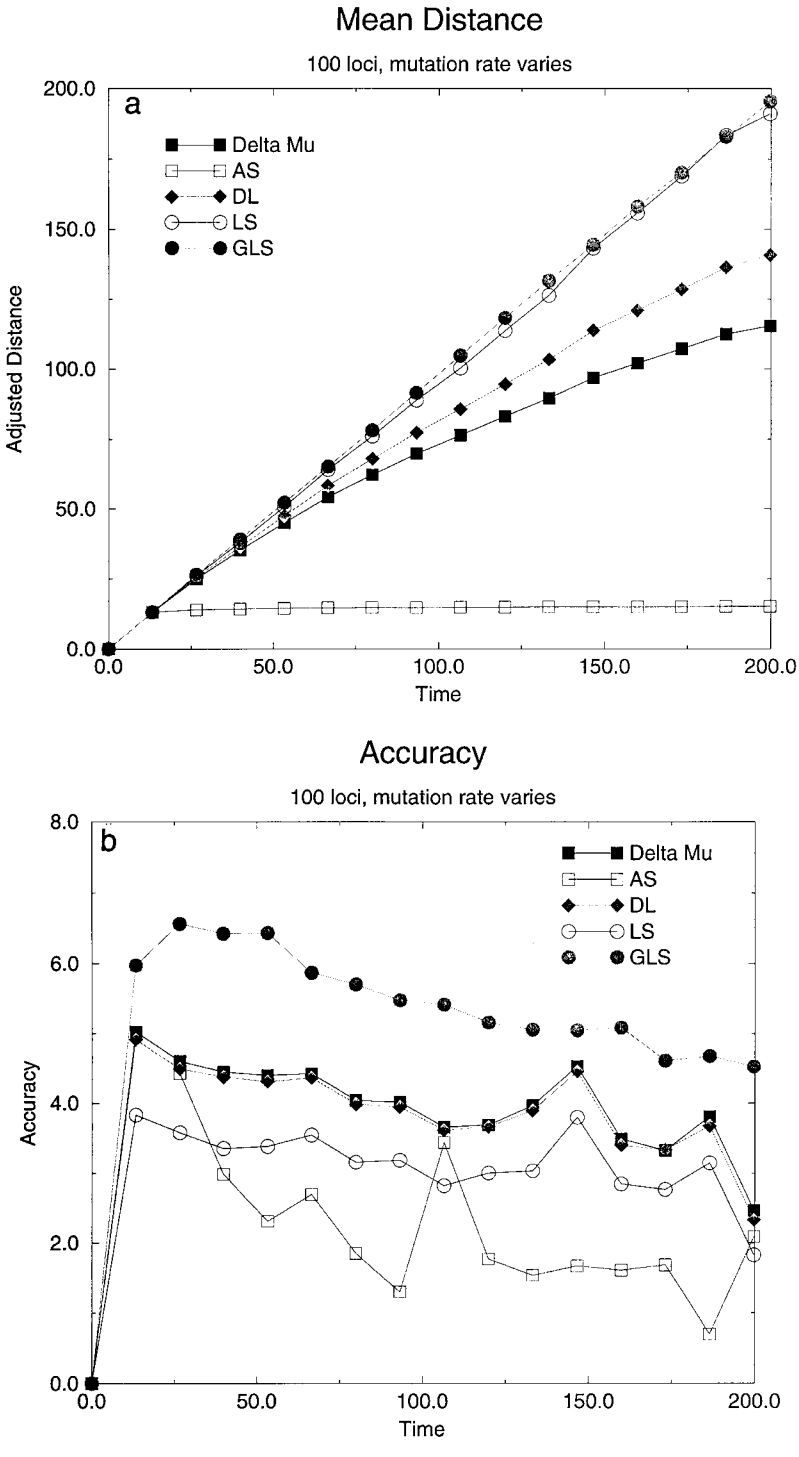
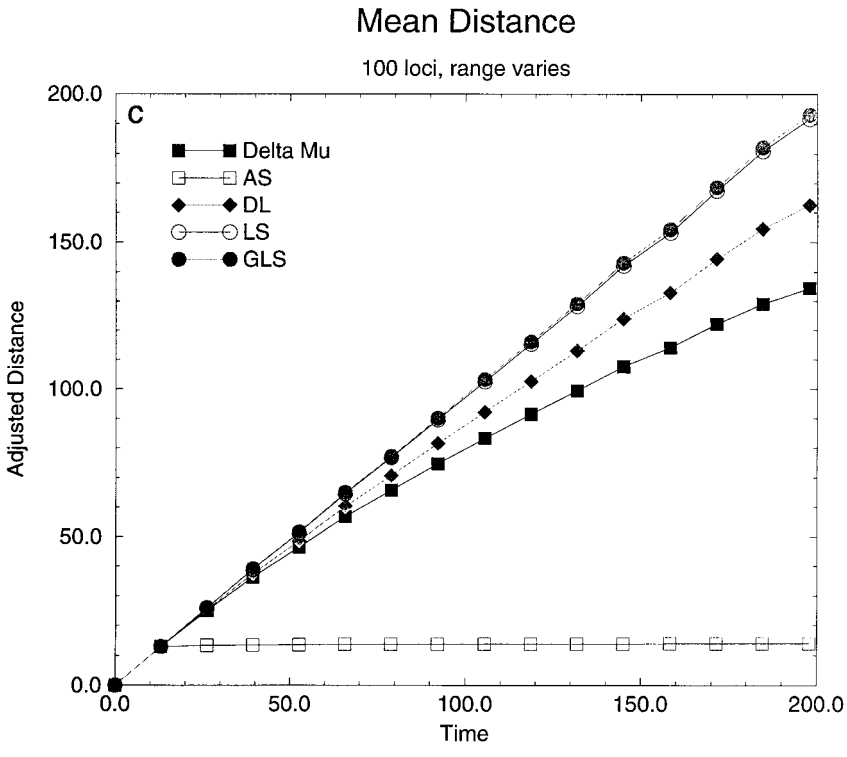


FIG. 4. Behavior of distances with time. The average behavior of $(\delta\mu)^2$ (Delta Mu), Allele Sharing (AS), the Log Distance correction (DL), and the unweighted (LS) and weighted (generalized) least squares estimate (GLS) are shown for 200 replicates of 100 loci in populations of size $N = 50$. In (a) and (b) mutation rates were distributed from 0.1 to 0.001, and the range was 50. In (c) and (d) locus ranges were evenly distributed from 10 to 48 and the mutation rate was 0.1. In (a) and (c), all distances are adjusted for comparative purposes such that their initial rate of increase over the first time interval is equal to $2t$. In (b) and (d), the accuracy (rate of increase/standard deviation of the distance) is weighted by time. Accurate knowledge of locus ranges and mutation rates is assumed. The time axes are measured in units of quadruple the median mutation rate (0.01 in (a) and (b), 0.1 in (c) and (d)) times generations (simulation cycles).



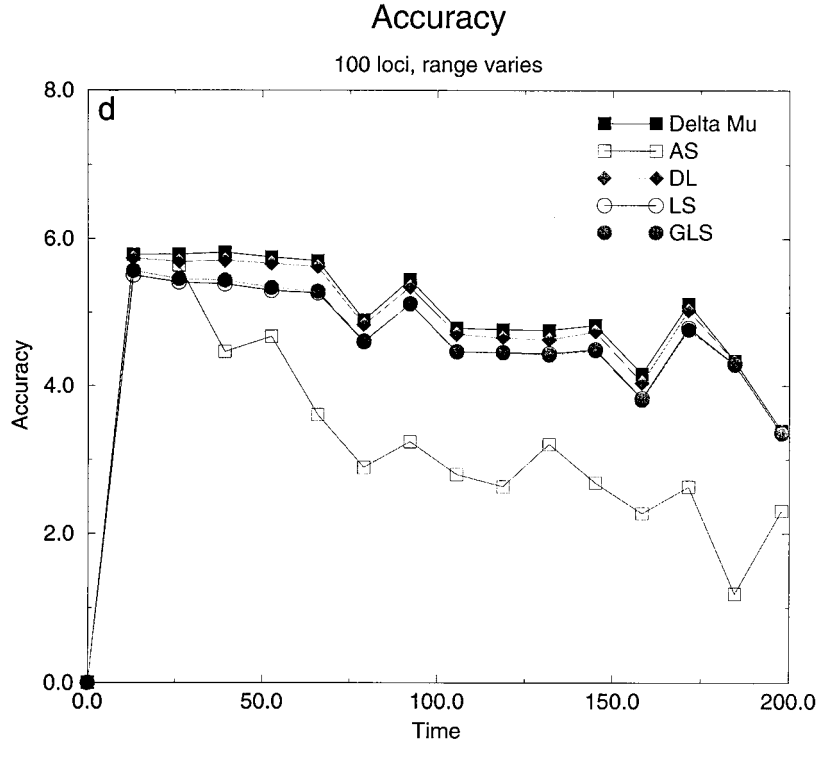


FIG. 4—*Continued*

sequence evolution to estimate the ratio of transition to transversion rates is to compare the ratio of uncorrected distances for closely related sequences. This strategy could also be applied to microsatellites, but it is highly inaccurate even for transition and transversion rates, which probably differ by no more than a factor of 10–20 (Wakeley, 1994). With microsatellites, where mutation rates may differ by a factor of 100, another approach is needed. The approach taken by Pollock and Goldstein (1995) for estimating rate ratios in DNA sequences is also unsuitable for microsatellites. They average the ratios of corrected distances (weighted by their variances) across all taxon pairs, but distance corrections for single microsatellite loci have large variances and may be undefined (Feldman *et al*, 1997). Here we use the least squares methodology iteratively to estimate separation times and mutation rates.

In the absence of perturbing factors, the rough estimates of the relative mutation rates obtained from the mean allelic variances will have a constant factor of N. When using the least squares or generalized least squares methods described above, this factor only affects estimates of the time of separation between population pairs, j, in that the estimates of time will be t_j/N rather than t_j. Thus, as usual, time cannot be estimated independently without further (generally unavailable) information on the magnitude of the mutation rate (or the population size). The proportional time estimates can be put into the sum of squares

$$SS=\sum_j \{(\delta\mu)_j^2 - E[(\delta\mu)_{j,\beta}^2]\}^2, \qquad (17)$$

where the sum is now over all population pairs, j, and again the expectation is given in Eq. (2). This sum is minimized with respect to β, and the process is repeated for all loci to obtain new estimates of the relative mutation rates. With the newly estimated mutation rates, the pairwise time estimates can then be recalculated. This minimization could be performed iteratively for a set number of times, or until convergence of the time and mutation rate parameters to a specified level of accuracy. As with least squares estimation of time, the weighted least squares solution can also be obtained by weighting each squared difference by the variance of the mutation rate estimate for the locus at the estimated time of separation of the population pair, j, obtained in a similar manner to the variance of the time estimate. Since t is the expectation of D_{LS}, if t is given then from Eq. (14) we may write

$$\hat{\beta}_l=\frac{\log[(M_l-(\delta\mu)_l^2)/M_l]}{4t(1-\cos[\pi/R_l]}, \qquad (18)$$

and

$$\sigma_{\hat{\beta}}^2 \approx \frac{2\{E[(\delta\mu)^2]\}^2}{(M-E[(\delta\mu)^2])^2\phi^2}, \qquad (19)$$

where $\phi=4t(1-\cos(\pi/R))$. Note that this is a weighted rather than generalized least squares solution as the correlation between population pairs is not taken into account.

Computer simulations were again used to evaluate mutation rate estimates based on the observed allelic variance, and on Eq. (17), β_{LS}, and its weighted equivalent, β_{WLS}. In order for β_{LS} and β_{WLS} to be reasonably accurate mutation rate estimates, a spread of divergence times between populations is required. Thus, for these simulations, a single population that had previously been iterated to equilibrium was split seven times at regular intervals, creating eight populations such that the phylogenetic tree relating these populations was maximally imbalanced. The range for all loci was 50, and the mutation rates ranged from 0.1 to 0.001 among the loci. The least squares mutation rate estimates were calculated using D_{GLS} as a measure of time, with no further iterations. Each data point was replicated 1000 times, and the mean and variance for all 20 loci were obtained for each of the three estimators of the mutation rate. The coefficients of variation (CV) for mutation rates obtained from the least squares methods are on the order of double those obtained from the allelic variance for the mutation rates in the range $\beta=0.05$–0.1 (Fig. 5). The slight upward trend with smaller mutation rates for the allelic variance method combined with the slight downward trend for the weighted least squares method, mean that they have more similar CVs for smaller mutation rates. β_{WLS} has a smaller CV than β_{LS} for $\beta<0.07$, and when $\beta=0.01$–0.001, it is only slightly greater than the CVs for the mutation rate estimates from the allelic variance. In deciding which method to use, consideration should be given to the sensitivity of allelic variance to other factors. Perturbations in the equilibrium distribution of the allelic variance will last longest for very small $N\beta$, while the CVs of the β_{WLS} mutation estimates are also most comparable to those from the allelic variance when mutation rates are small. Thus the least squares methodology for estimating the mutation rate appears most appropriate under the conditions modeled when the estimators are comparable and $N\beta$ is small, that is, when $N\beta<0.5$ ($\beta<0.01$). It should be noted that while all three methods will improve with increasing numbers of populations sampled, they may be differentially sensitive to the phylogenetic relationships among those populations. For example, the accuracy of

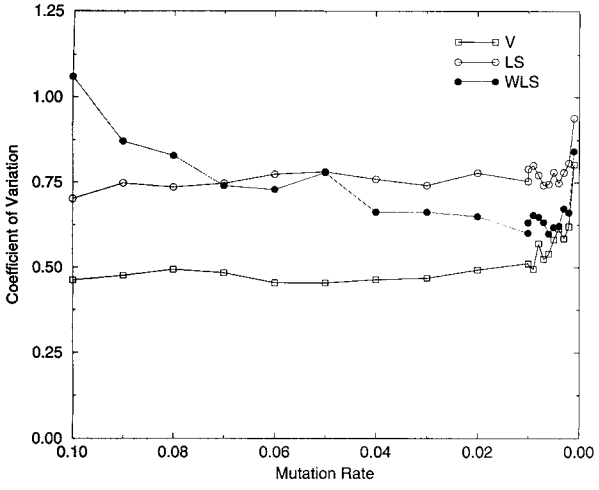


FIG. 5. Coefficients of variation for mutation rates. Mutation rates were estimated from the allelic variance (*V*, open squares), or via the least squares (*LS*, open circles) or weighted least squares (*WLS*, filled circles) procedure. Mutation rates were distributed from 0.1 to 0.001 across the 20 loci in 1000 replicate simulations. The range for all loci was 50. In each simulation, eight populations were created by splitting a single population seven times at regular intervals, so that the phylogenetic tree relating these populations was maximally unbalanced. The overall length of the simulation (seven times the length of the splitting interval) following the first split was 700 generations ($\beta t = 0.7$–70).

the least squares methods is likely to increase more than the variance method if larger numbers of closely related populations are added to the sample.

4. DISCUSSION

The results presented in this paper should be useful to researchers interested in extending microsatellite analysis of phylogeny to more divergent populations. Methodology is also provided for estimating parameters which should prove useful in comparing evolutionary models of microsatellite evolution. The estimation of parameters is a necessary preliminary step in choosing microsatellites for extended analyses, and will save the expenditure of resources on sampling loci which are unlikely to be informative. While large numbers of microsatellite loci are needed to achieve an unbiased and accurate distance over long periods of time, specific questions can be addressed with fewer, well-chosen loci. The model analyzed in this paper and in Feldman *et al.* (1997) provides a framework for defining what is most important in the selection of loci (Fig. 4). It also shows how to reduce the bias and increase the accuracy in estimating more distant relationships. Surprisingly, it appears from Fig. 4 that variation in the range is less of a problem in this regard than variation in the mutation rate.

The principal factors in determining the length of time during which a locus will provide accurate information are the range and the mutation rate. This useful lifetime of a locus can be defined as the length of time until the time-weighted accuracy drops below some cutoff. The cutoff is in some sense arbitrary, but can be envisaged as the point at which the accuracy of a distance using a given number of similar loci would drop below one; at that point such a distance would be nearly useless phylogenetically. The useful lifetime increases with the square of the absolute range within which allelic scores may wander (Fig. 6a). The effect of the mutation rate is only linear, but mutation rates potentially may vary more than ranges across loci (Fig. 6b). For moderate values of $N\beta$, the expected maximum distance between populations for a microsatellite locus decreases as $N\beta$ increases, but the initial rate of divergence decreases as well. Thus, a microsatellite with a range of 50 and a fast mutation rate of 10^{-2} would still last 2.5 times longer than a microsatellite with a range of 10 and a mutation rate of 10^{-3}. The absence of an independent effect of $N\beta$ will only hold up to a certain point, since for a sufficiently large $N\beta$ relative to R, equilibrium allelic variances may be too high to permit diversification, our primary concern.

It might be argued that the number of microsatellites necessary for accurate reconstruction of phylogeny is prohibitive due to the costs of selecting and screening the microsatellites. Recent genome projects, however, have been screening microsatellites in prodigious quantities. At last count, 5,264 (6,580) microsatellites dispersed across the human (mouse) genome have been characterized (Dietrich *et al.*, 1996; Dib *et al.*, 1996). Due to their usefulness in mapping, many more are, or soon will be, available from model organisms such as *Drosophila*, maize, and zebrafish, from pathogenic organisms such as *Plasmodium*, and from agriculturally useful organisms, such as cows, sheep, chickens, pigs, tomatoes, soybeans, and rice (Goldstein and Clark, 1995; Taramino and Tingey, 1996; Postlethwait *et al.*, 1994; Su and Willems, 1996; Ma *et al.*, 1996. Crawford *et al.*, 1995; Crooijmans *et al.*, 1996; Rohrer *et al.*, 1996; Broun and Tanksley, 1996; Akkaya *et al.*, 1995; Xiao *et al.*, 1994). With these large numbers available for sampling, along with recent improvements in rapid cloning of new microsatellites and more efficient typing of genotypes (Ostrander *et al.*, 1992), practical phylogenetics appears quite possible beyond the subspecies level, though it will entail a more sophisticated approach than is currently practised. A serious complication for evolutionary analysis is that microsatellites will degrade with time. There is reason to be optimistic, however, that a substantial window of time

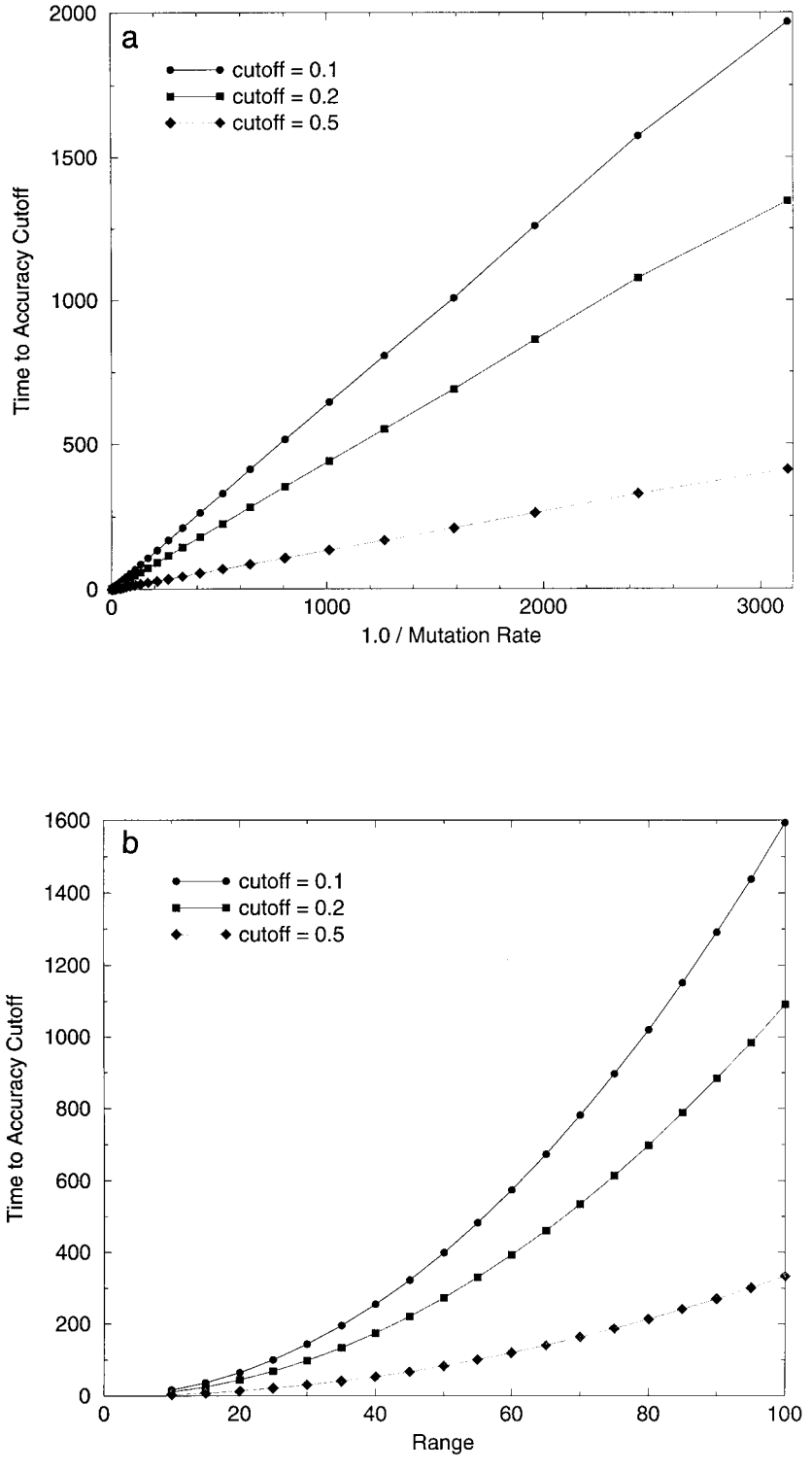


FIG. 6. Useful lifetime of microsatellite loci. The useful lifetime of a microsatellite was measured as the length of time until a particular accuracy cutoff was reached. The accuracy (slope/standard deviation) of D_L was weighted by the amount of time passed, and cutoffs were either 0.1 (circles), 0.2 (squares), or 0.5 (diamonds). These cutoffs correspond to a weighted accuracy of 1.0 for 100, 25, and 4 loci, respectively. In (a), the mutation rate varies from 0.64 to 0.0032, the range is 20. In (b), the mutation rate is 0.01, the range varies from 10 to 100. Time is measured in units of βt where $\beta = 0.01$.

may exist before this occurs. Many microsatellites originally typed in humans are polymorphic throughout the old world monkeys (Coote and Bruford, 1996), and recently published reports describe microsatellites lasting for hundreds of millions of years in sea turtles and fish (Fitzsimmons *et al*., 1995; Rico *et al*., 1996).

It cannot be overemphasized that the number of loci required to make phylogenetic estimates is large, even under the assumption of an infinite range. The requirement is even larger under range constraints. In the absence of range constraints, when the distance $(\delta\mu)^2$ is expected to grow linearly with time, the standard deviation of single distance estimates is slightly greater than the distance estimate itself. Since the standard deviation goes down with the square root of the number of loci, it is clear that one hundred loci are needed to get the coefficient of variation down to slightly greater than 0.1. Under a model with range constraints, the standard deviation of $(\delta\mu)^2$ is still slightly greater than its expectation, but the expectation is asymptotic with time. This further reduces the resolving power of the distance measure as time increases. Resolving power at deeper times can only be increased by increasing the number of loci. While adding loci with any R and β will somewhat reduce the degree to which the distances asymptote or are variable, resolving power will be increased most by selectively adding those loci with the greatest accuracy.

A feasible research plan would be to choose a small number of divergent species and type five to ten individuals from each for a large number of microsatellite loci (perhaps in the hundreds) in order to estimate the ranges of those loci. Preliminary estimates of the relative mutation rates would be made from the mean allelic variances at these loci. Further analyses on a large number of species or populations would then be restricted to those loci with acceptably large ranges and small mutation rates. As such datasets become available, patterns may emerge which would allow *a priori* selection of microsatellites for particular projects. For example, we anticipate that different motif sizes (e.g., dinucleotide, trinucleotide, and tetranucleotide repeats), and perhaps different motif types (e.g., CA vs. TA), may have tendencies towards particular range constraints or mutation rates. If so, specific types of microsatellites can be used for specific problems, obviating the need for prior characterization of each locus in every taxon.

The ideal distance measure for use with microsatellites will depend on the characteristics of the microsatellites and on the phylogenetic question being addressed. If range constraints are large, and the separation times between populations of interest are small, $(\delta\mu)^2$ may give good results both in terms of accuracy and linearity with time. Allele sharing can be more accurate for extremely short separation times, but will not be particularly linear. If separation times are larger, D_L is nearly as accurate as $(\delta\mu)^2$, and asymptotes more slowly. The least squares distances, however, will maintain linearity indefinitely as long as the number of loci is large enough. The least squares distances are slightly less accurate in those cases where there is variation in range sizes and no variation in mutation rate among loci, but when variation in mutation rate occurs in the presence of range constraints, D_{GLS} is considerably more accurate than the other distances for all but the shortest separation times.

Some of the assumptions made in the model analyzed here may not exactly mimic natural microsatellite behavior. Actual boundaries may be soft rather than reflecting, the mutation rate may change with size (Goldstein and Clark, 1995), or become directionally biased with size, and locus mutation and death should be considered. Range constraints have clear and dramatic effects on the dynamics of microsatellite evolution, however, and the analysis here has shown that some degree of compensation can be made for these constraints. Parameter estimations using these analyses can also be used to determine how well real-world microsatellite conform to the expectations of this model compared to any other established model.

ACKNOWLEDGMENTS

We thank M. J. Nauta and F. J. Weissing for making available an unpublished manuscript. D.D.P. is a Hitchings–Elion fellow of the Burroughs Wellcome Fund. This research is supported in part by NIH Grants GM 28428 and GM 28016 to M.W.F. The analytical techniques developed here are currently being incorporated into an existing computer program by Minch *et al*., available at http://lotka.stanford.edu, or by writing to E. Minch at minch@malecot.stanford.edu.

REFERENCES

Akkaya, M. S., *et al.* 1995. Integration of simple sequence repeat DNA markers into a soybean linkage map, *Crop Sci.* **35**, 1439–1445.

Broun, P., and Tanksley, S. D. 1996. Characterization and genetic mapping of simple repeat sequences in the tomato genome, *Mol. Gen. Genet*. **250**, 29–49.

Charlesworth, B. 1994. The effect of background selection against deleterious mutations on weakly selected, linked variants, *Genet. Res.* **63**, 213–227.

Coote, T., and Bruford, M. W. 1996. Human microsatellites applicable for analysis of genetic variation in apes and old world monkeys, *J. Heredity* **87**, 406–410.

Crawford, A. M., *et al.* 1995. An autosomal genetic linkage map of the sheep genome, *Genetics* **140**, 703–724.

Crooijmans, R. P. M. A., *et al.* 1996. Preliminary linkage map of the chicken (*Gallus domesticus*) genome based on microsatellite markers, *Poultry Sci.* **75**, 746–754.

Dib, C., *et al.* 1996. A comprehensive genetic map of the human genome based on 5,264 microsatellites, *Nature* **380**, 152–154.

Dietrich, W., *et al.* 1996. A comprehensive genetic map of the mouse genome, *Nature* **380**, 149–152.

Feldman, M., Bergman, A., Pollock, D. D., and Goldstein, D. B. 1997. Microsatellite genetic distances with range constraints: Analytic description and problems of estimation, *Genetics* **145**, 207–216.

Fitzsimmons, N. N., Moritz, C., and Moore, S. S. 1995. Conservation and dynamics of microsatellite loci over 300-million years of marine turtle evolution, *Mol. Biol. Evol.* **12**, 432–440.

Garza, J. C., Slatkin, M., and Freimer, N. B. 1995. Microsatellite allele frequencies in humans and chimpanzees with implications for constraints on allele size, *Mol. Biol. Evol.* **12**(4), 594–603.

Goldstein, D. B., and Clark, A. G. 1995. Microsatellite variation in North American populations of *Drosophila melanogaster*, *Nucl. Acids Res.* **23**, 653–662.

Goldstein, D. B., and Clark, A. G. 1996. unpublished results.

Goldstein, D. B., and Pollock, D. D. 1994). Least squares estimation of molecular distance—Noise abatement in phylogenetic reconstruction, *Theor. Popul. Biol.* **45**, 219–226.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L., and Feldman, M. W. 1995a. An evaluation of genetic distances for use with microsatellite loci, *Genetics* **139**, 463–471.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L., and Feldman, M. W. 1995b. Genetic absolute dating based on microsatellites and modern human origins, *Proc. Natl. Acad. Sci. USA* **92**, 6723–6727.

Hudson, R. R. 1995. Deleterious background selection with recombination, *Genetic* **141**, 1605–1617.

Ma, R. Z., *et al.* 1996. Isolation and characterization of 45 polymorphic microsatellites from the bovine genome, *Animal Genet.*, 43–47.

Moran, P. 1975. Wandering distributions and the electorphoretic profile , *Theor. Popul. Biol.* **8**, 318–330.

Nauta, M. J., and Weissing, F. J. 1996. Constraints on allele size at microsatellite loci: Implications for genetic differentiation, *Genetics* **143**, 1021–1032.

Ostrander, E. A., Jong, P. M., Rine, J., and Duyk, G. 1992. Construction of small insert genomic DNA libraries highly enriched for microsatellite repeat sequence, *Proc. Natl. Acad. Sci. USA* **89**, 3419–3423.

Pollock, D. D., and Goldstein, D. B. 1995. A comparison of two methods for constructing evolutionary distances from a weighted contribution of transition and transversion differences, *Mol. Biol. Evol.* **12**(4), 713–717.

Pollock, D. D. 1998. Increased accuracy in analytical molecular distance estimation, *Theor. Popul. Biol.*, in press.

Postlethwait, J. H., *et al.* 1994. A genetic linkage map for the zebrafish, *Science* **264**, 699–703.

Rice, J. A. 1995. "Mathematical Statistics and Data Analysis," 2nd ed., Duxbury, N. Scituate, MA.

Rico, C., Rico, I., and Hewitt, G. 1996. 470-million yeats of conservation of microsatellite loci among fish species, *Proc. Roy. Soc. London B* **263**, 549–557.

Rohrer, G. A., *et al.* 1996. A comprehensive map of the porcine genome, *Genome Res.* **6**, 371–391.

Shriver, M., Jin, L., Boerwinkle, E. 1995. R. Deka, R. E. Ferrel, and R. Chakraborty, A novel measure of genetic distance for highly polymorphic tandem repeat loci, *Mol. Biol. Evol.* **12**(5), 914–920.

Slatkin, M. 1995a. Hitchhiking and associative overdominance at a microsatellite locus, *Mol. Biol. Evol.* **12**, 473–480.

Slatkin, M. 1995b. A measure of population subdivision based on microsatellite allele frequencies, *Genetics* **139**, 457–462.

Strassman, J. E., Solis, C. R., Barefield, K., and Queller, D. C. 1996. Trinucleotide microsatellite loci in a swarm-founding neotropical wasp, *Parachartergus coloboptersu* and their usefulness in other social wasps, *Mol. Ecol.* **5**, 459–461.

Su, X. Z., and Willems, T. E. 1996. Toward a high-resolution *Plasmodium falciparum* linkage map-polymorphic markers from hundreds of simple sequence repeats, *Genomics* **33**, 430–444.

Tajima, M., and Takezaki, N. 1994. Estimation of evolutionary distance for reconstructing molecular phylogenetic tree, *Mol. Biol. Evol.* **11**, 278–286.

Tajima, M. 1993. Unbiased estimation of evolutionary distance between nucleotide sequences, *Mol. Biol. Evol.* **10**, 677–688.

Taramino, G., and Tingey, S. 1996. Simple sequence repeats for germplasm analysis and mapping in maize, *Genome* **39**, 277–287.

Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias, *Mol. Biol. Evol.* **11**, 436–442.

Weber, J. L., and Wong, C. 1993. Mutation of human short tandem repeats, *Hum. Mol. Genet.* **2**, 1123–1128.

Xiao, J. H., *et al.*1994. Saturated molecular map of the rice genome based on an interspecific backcross population, *Genetics* **138**, 1251–1274.

Zhivotovsky, L. A., and Feldman, M. W. 1995. Microsatellite variability and genetic distances, *Proc. Natl. Acad. Sci. USA* **92**, 11549–11552.

Zhivotovsky, L. A., Feldman, M. W., and Grishechkin, S. A. 1997. Biased mutations and microsatellite variation, *Mol. Biol. Evol.* **14**, 926–933.