Microsatellite Genetic Distances With Range Constraints: Analytic Description and Problems of Estimation

Marcus W. Feldman,*,† Aviv Bergman,† David D. Pollock† and David B. Goldstein†,‡

**Department of Biological Sciences, Stanford University, Stanford, California 94305, †Interval Research Corporation, Palo Alto, California 94305 and ‡Department of Zoology, University of Oxford, Oxford 0X1 3PS, United Kingdom*

Manuscript received April 8, 1996
Accepted for publication September 5, 1996

ABSTRACT

Statistical properties of the symmetric stepwise-mutation model for microsatellite evolution are studied under the assumption that the number of repeats is strictly bounded above and below. An exact analytic expression is found for the expected products of the frequencies of alleles separated by k repeats. This permits characterization of the asymptotic behavior of our distances D_1 and $(\delta\mu)^2$ under range constraints. Based on this characterization we develop transformations that partially restore linearity when allele size is restricted. We show that the appropriate transformation cannot be applied in the case of varying mutation rates (β) and range constraints (R) because of statistical difficulties. In the special case of no variation in β and R across loci, however, the transformation simplifies to a usable form and results in a distance much more linear with time than distances developed for an infinite range. Although analytically incorrect in the case of variation in β and R, the simpler transformation is surprisingly insensitive to variation in these parameters, suggesting that it may have considerable utility in phylogenetic studies.

MICROSATELLITES are a special class of tandemly repeated DNA in which a specific motif of 2–6 bp is repeated up to ~100 times (TAUTZ 1993). Microsatellite loci with > ~10 repetitions of the basic motif are highly variable in taxa ranging from plants to vertebrates (LAGERCRANTZ *et al.* 1993). It is common to observe heterozygosities of 0.8 and as many as 20 or more alleles at a locus (BOWCOCK *et al.* 1994; MACHUGH *et al.* 1994; DEKA *et al.* 1995; GOLDSTEIN and CLARK 1995). Microsatellites are also very easily scored using PCR-based methods and tend to be reliably variable in all populations of a given species. Another important advantage is that microsatellite analyses provide information about the state of specific loci, facilitating a number of population-genetic inferences.

For these reasons, microsatellites are rapidly replacing allozymes and newer markers (*e.g.*, randomly amplified polymorphic DNAs) in studies attempting to estimate demographic and evolutionary parameters of natural populations (MACHUGH *et al.* 1994; ESTOUP *et al.* 1995; GOLDSTEIN *et al.* 1995c; SLATKIN 1995a,b). They are also beginning to be used to estimate phylogenetic relationships among populations and closely related species (BOWCOCK *et al.* 1994; MACHUGH *et al.* 1994; DEKA *et al.* 1995; ESTOUP *et al.* 1995; GOLDSTEIN *et al.* 1995a,b; PEPIN *et al.* 1995), but success here has been limited both by the availability of variable microsatellites in multiple species, and by uncertainty as to which genetic distance measure is most appropriate for microsatellites (GOLDSTEIN *et al.* 1995a).

The variability of microsatellite loci is due to their exceptionally high mutation rate, which seems to average ~0.0001 (WEBER and WONG 1993). This high mutation rate also guarantees that isolated populations diverge rapidly, but an exact description of this process of divergence is elusive since it depends on the precise details of the mutation process. Direct studies of microsatellite mutation mechanisms, based both on artificial constructs in yeast (HENDERSON and PETES 1992) and analyses of human pedigrees (WEBER and WONG 1993) have shown that most mutations involve the addition or subtraction of a small number of repeat units. This contradicts the assumptions of the infinite alleles mutation model, in which all new mutations are to alleles not previously represented in the population. To account for this aspect of the mutation process, a number of authors have recently studied the stepwise mutation model, which was first developed to describe the evolution of the charge state of proteins as inferred from electrophoretic mobility (OHTA and KIMURA 1973).

Using different methods, GOLDSTEIN *et al.* (1995a) and SLATKIN (1995a) both demonstrated that an unbiased estimator of separation time can be obtained by taking the squared differences of all pairs of alleles drawn one from each of the two populations. Subsequently, GOLDSTEIN *et al.* (1995b) showed that a related distance, given by the squared difference between the means of the two populations, is also linearly related to time. This distance, called $(\delta\mu)^2$, has the further advantage of being independent of population size when pop-

Corresponding author: Marcus W. Feldman, Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020.
E-mail: marc@charles.stanford.edu

ulations are internally at mutation-drift equilibrium. Although analytically unbiased, sampling effects may inflate $(\delta\mu)^2$ in practice unless sample sizes are relatively large. The linear relationship with time of these stepwise distances is in contrast to traditional distances (*e.g.*, F_{ST}, Nei's distances), which are well known to possess an asymptote under stepwise mutations (NEI and CHAKRABORTY 1973; CHAKRABORTY and NEI 1977; NEI 1987).

The analytic descriptions in GOLDSTEIN *et al.* (1995a) and SLATKIN (1995a), however, were highly idealized. Most importantly, both assumed that microsatellite alleles can have an arbitrarily large number of repeats. Noting that allelic sizes are in fact tightly constrained, GOLDSTEIN *et al.* (1995a) emphasized that their distance would also asymptote when applied in practice. While lacking a formal model including range constraints, GOLDSTEIN *et al.* (1995a) presented a heuristic argument suggesting that the value of their stepwise distance at maximal divergence is about $(R^2 - 1)/6$, where R is the number of alleles possible. They also provided a very rough estimate of the duration of linearity by calculating how long it would take to reach this value in the unconstrained model.

The exact value of this asymptote and the exact details of the approach of the distance to its maximal value are of critical importance to the evaluation of stepwise distances. For the unconstrained model, both SLATKIN (1995a) and GOLDSTEIN *et al.* (1995a) demonstrated that nonstepwise distances are more accurate for closely related taxa, but that stepwise distances become superior beyond some critical level of divergence. Intuitively, the stepwise distances become superior after the nonstepwise distances have lost linearity and no longer accurately reflect separation times. With range constraints, however, stepwise distances also asymptote and should work better only within a certain window of separation times. The lower boundary of this window is related to the time at which nonstepwise distances asymptote, and the upper boundary is related to the time at which the stepwise distance reaches its asymptote of $(R^2 - 1)/6$. It is also important to appreciate that this window in which stepwise distances are superior will only exist if R is sufficiently large. GOLDSTEIN *et al.* (1995b) used polymorphisms in three primate species to investigate experimentally whether the level of divergence among closely related primates falls within this window. They showed that $(\delta\mu)^2$ allows the three possible rooted trees relating humans, chimps and gorillas to be distinguished, while Nei's distances, for example, do not.

For a more rigorous assessment of the reliability of different distances under range constraints, and to develop statistical corrections to recover linearity, it is necessary to have an analytic description of the dynamics of genetic distances under range constraints. Here we introduce an analytic framework that allows an exact description of the expected dynamics of loci undergoing stepwise mutations on a restricted set of R alleles. GARZA *et al.* (1995) recently proposed a model of microsatellite evolution that incorporates bias in the mutation process in the form of a "restoring force" such that small alleles tend to mutate upward and large alleles downward. In the present study, we incorporate explicit range constraints for the following reasons. First, maximal and minimal allele size would seem much easier to estimate than a parameter governing the degree of asymmetry in the mutation. This difference becomes especially important in connection with attempts to improve genetic distances by applying corrections based on parameters that must be estimated. A second motivation is that an explicit upper ceiling seems to be a closer representation of the known behavior of certain microsatellites (*e.g.*, trinucleotides) in which the mutation process is more symmetric and the rate moderate for small alleles, while for larger ones the rate is extremely high and biased upward. If we assume that the large alleles are severely disadvantageous, a fixed range becomes an appropriate representation. For simplicity, we consider only the strict stepwise mutation model.

Dealing with an infinite number of allele sizes, previous models (OHTA and KIMURA 1973; MORAN 1975) have described the evolution of the expected products of allele frequencies separated by k units, given by $C_k = \Sigma_i \nu_i \nu_{i+k}$, where ν_i and ν_{i+k} are the frequencies of alleles with i and $i + k$ repeats. In the case of finite R, however, it is straightforward to show that a closed form recursion for the C_k cannot be obtained independent of the underlying allele frequencies. Studying variation within a single population NAUTA and WEISSING (1996) approximated the frequencies of boundary terms in each of the C_k. Here we instead describe directly the evolution of the matrix of expected products of allele frequencies. The traditional C_k can be obtained from such a matrix by summing diagonal rows. This analysis confirms the numerical results obtained by NAUTA and WEISSING (1996) and allows us to obtain analytic expressions for the expectations of measures of variation within a population. More relevant to our purposes here, we also introduce the interpopulational sum C_k^*, which is the sum of products of allele frequencies drawn one each from two isolated populations. We show how this is related to our distances $(\delta\mu)^2$ and ASD, a distance we introduced earlier (GOLDSTEIN *et al.* 1995a), that is the average of the squared differences of alleles drawn one each from two isolated populations. To describe the evolution of the C_k^*, we derive recursions for the matrix of products of allele frequencies, drawn one from each of two isolated populations.

We use this approach to show formally that the expected value of ASD between two maximally diverged populations converges to $(R^2 - 1)/6$, as first suggested by GOLDSTEIN *et al.* (1995a). Similarly, $E(\delta\mu)^2$ converges to $(R^2 - 1)/6 - E(D_0)$, where D_0 is the average squared difference between pairs of alleles both drawn from a

single population (GOLDSTEIN *et al.* 1995a). In practice, D_0 may be calculated as twice the within-population variance in repeat sizes. We also show that the rate of convergence to this maximum is given by $(1 - 2\beta + 2\beta \cos \pi/R)^2$, where β is the stepwise per locus mutation rate. On the basis of these results, we outline the form that any correction must take to provide a statistically unbiased estimate of separation times. As we shall see, the required form ensures that such corrections will not be applicable in the general case in which β and R vary across loci. Nevertheless, in the special case of sets of loci with the same β and R, a simple analytic correction can lead to less biased estimation of separation times without the statistical complications that arise in the general case. We use computer simulations to assess the reliability of the correction in this special case and compare the performance of the new estimator to existing estimators like $(\delta\mu)^2$. We also test the sensitivity of the simplified distance to variation in β and R and find that it continues to behave well, despite being formally inappropriate in this case.

METHODS AND RESULTS

The following analysis is based on that of MORAN (1975) who derived a number of important results for the stepwise mutation model in which there is no constraint on the number of repeats occurring at a locus. Here we specify a range R of repeat scores, and for convenience these are represented as $1, 2, \ldots, R$. Thus, R is the possible number of alleles. Symmetric one-step mutation is postulated so that for any allele i the rate of mutation to each of $i - 1$ and $i + 1$ is β. At time t the number of (haploid) genes carrying i repeats is $n_i(t)$ with $\Sigma_i\, n_i(t) = N$, the population size. After mutation, the population frequency of allele i is

$$\pi_i(t) = (1 - 2\beta)\frac{n_i(t)}{N} + \beta\left[\frac{n_{i-1}(t)}{N} + \frac{n_{i+1}(t)}{N}\right], \qquad 2 \le i \le R - 1 \quad (1a)$$

with

$$\pi_1(t) = (1 - \beta)\frac{n_1(t)}{N} + \beta\frac{n_2(t)}{N}, \quad (1b)$$

$$\pi_R(t) = (1 - \beta)\frac{n_R(t)}{N} + \beta\frac{n_{R-1}(t)}{N}. \quad (1c)$$

Multinomial sampling then takes place to produce the next generation of alleles. Writing E_{t-1} as the expectation operator in generation t given the frequencies at time $t - 1$, we have for $i = 1, 2, \ldots, R$,

$$E_{t-1}[n_i(t)n_j(t)] = N(N - 1)\pi_i(t - 1)\pi_j(t - 1), \quad i \ne j \quad (2a)$$

$$E_{t-1}[n_i^2(t)] = N(N - 1)\pi_i^2(t - 1) + N\pi_i(t - 1). \quad (2b)$$

In our earlier studies of microsatellite evolution, two functions were used to study the evolution of allele frequencies within and between populations. These were D_0, the average squared difference in repeat numbers for two alleles drawn from the same population, and D_1, the same average when the alleles are drawn one each from different populations. (Note: We have used the average square distance symbols ASD and D_1 interchangeably in our earlier papers. Here we shall use D_1 for brevity.) The time-dependent behavior of these functions when there is no restriction on the range of repeat numbers can be studied directly (GOLDSTEIN *et al.* 1995a). In particular, the difference $D_1 - D_0$, denoted $(\delta\mu)^2$, provides a useful distance that is linear in the time since the separation of the two populations and removes the effect of population size (GOLDSTEIN *et al.* 1995b; ZHIVOTOVSKY and FELDMAN 1995).

In the absence of range restriction, MORAN (1975) analyzed the model in terms of the moments of the quantities

$$C_k(t) = C_{-k}(t) = \sum_i \frac{n_i(t)}{N}\frac{n_{i+k}(t)}{N}, \quad (3)$$

where we may write

$$D_0 = 2\sum_i i^2 C_i. \quad (4)$$

A similar sum of products with the alleles chosen from each of two populations gives rise to D_1:

$$D_1 = 2\sum_i i^2 C_i^* \quad (5)$$

with

$$C_k^*(t) = C_{-k}^*(t) = \frac{1}{2}\left[\sum_i \frac{n_i^{(a)}(t)}{N}\frac{n_{i+k}^{(b)}(t)}{N} + \sum_i \frac{n_{i+k}^{(a)}(t)}{N}\frac{n_i^{(b)}(t)}{N}\right] \quad (6)$$

and the superscripts to refer to the two populations (a) and (b).

The quantities C_i and C_i^* appear to be more difficult to analyze directly in the case of restricted range and we have chosen an approach that uses each of the summands in (3) and (6).

Analysis: Denote by $\mathbf{B}$ the $R \times R$ matrix with elements $B_{11} = B_{RR} = 1 - \beta$; $B_{ii} = 1 - 2\beta$ for $i \ne 1, R$; $B_{i,i+1} = B_{i-1,i} = \beta$. This tridiagonal matrix represents the one-step mutation process with forward and backward rates each β. Denote by $\boldsymbol{\nu}_t$ the (row) vector of R allele frequencies $n_1(t)/N$, $n_2(t)/N \cdots n_R(t)/N$ at time t. Then we may rewrite the relations (Equation 2) using the symmetry of $\mathbf{B}$ as

$$E_{t-1}(\boldsymbol{\nu}_t^T\boldsymbol{\nu}_t) = \left(1 - \frac{1}{N}\right)\mathbf{B}\boldsymbol{\nu}_{t-1}^T\boldsymbol{\nu}_{t-1}\mathbf{B} + \frac{1}{N}\sum_{i=1}^{R}\mathbf{A}_i(\mathbf{B}\boldsymbol{\nu}_{t-1}^T)\mathbf{e}_i, \quad (7)$$

where $\mathbf{A}_i$ is the $R \times R$ matrix with 1 in the (i, i) position and 0 elsewhere, $\mathbf{e}_i$ is the $1 \times R$ vector with 1 in position i and 0 elsewhere, and the superscript T represents the transpose. Relation 7 reports the expectation after a generation of multinomial sampling conditioned on generation $t-1$. Upon iteration we obtain the expectation given the initial frequencies

$$E_0(\boldsymbol{\nu}_t^T \boldsymbol{\nu}_t) = \frac{1}{N} \sum_{j=0}^{t-1} \left(1 - \frac{1}{N}\right)^j \mathbf{B}^j \left\{ \sum_{i=1}^{R} \mathbf{A}_i (\mathbf{B}^{t-j} \boldsymbol{\nu}_0^T) \mathbf{e}_i \right\} \mathbf{B}^j + \left(1 - \frac{1}{N}\right)^t \mathbf{B}^t \boldsymbol{\nu}_0^T \boldsymbol{\nu}_0 \mathbf{B}^t. \quad (8)$$

It is clear that $\mathbf{B}^t \boldsymbol{\nu}_0^T$ converges to the uniform probability vector $(1/R, 1/R, \ldots, 1/R)^T$. The rate of convergence is the second largest eigenvalue of $\mathbf{B}$, namely

$$\lambda_1 = 1 - 2\beta + 2\beta \cos \frac{\pi}{R}, \quad (9)$$

(BARNETT 1990, p. 350; the largest is obviously $\lambda_0 = 1$). This enables us to write

$$\lim_{t \to \infty} E_0(\boldsymbol{\nu}_t^T \boldsymbol{\nu}_t) = \lim_{t \to \infty} \frac{1}{N} \sum_{j=0}^{\infty} \left(1 - \frac{1}{N}\right)^j \mathbf{B}^j \left[\sum_{i=1}^{R} \mathbf{A}_i (\mathbf{B}^{t-j} \boldsymbol{\nu}_0^T) \mathbf{e}_i \right] \mathbf{B}^j$$

$$= \frac{1}{RN} \sum_{j=0}^{\infty} \left(1 - \frac{1}{N}\right)^j \mathbf{B}^{2j} \quad (10)$$

$$= \frac{1}{RN} \left[\mathbf{I} - \left(1 - \frac{1}{N}\right) \mathbf{B}^2 \right]^{-1} = \mathbf{Z}, \quad (11)$$

say, where $\mathbf{I}$ is the identity matrix.

The entries of the matrix $\mathbf{Z}$ give us the equilibrium expectations of products of frequencies $(n_i/N)(n_j/N)$ $(i, j = 1, 2, \ldots, R)$. Equilibrium values of C_i (as defined in Equation 3 above) are obtained by summing the appropriate diagonal entries of $\mathbf{Z}$. In principle, we may then compute D_0 as in (4).

First, let us examine the series in (10). The entries of the powers $\mathbf{B}^{2j}$ are polynomials in β. Only the diagonal and the first super- and subdiagonals have linear terms in β. The (1, 1) and (R, R) elements of $\mathbf{B}^{2j}$ are $(1 - 2\beta j)$, while the (i, i) elements $(i \neq 1, R)$ are $1 - 4\beta j$, neglecting terms $0(\beta^2)$. Upon summation, however, these terms in β^2 actually turn out to be $0(\beta N)^2$. If $\beta N < 1$, then to order $(\beta N)^2$,

$$C_0 = \frac{1}{RN} \sum_{j=0}^{\infty} \left(1 - \frac{1}{N}\right)^j [2(1 - 2\beta j) + (R - 2)(1 - 4\beta j)] \quad (12)$$

and, on summation, we have

$$C_0 = 1 - 4\beta(N - 1)\left(1 - \frac{1}{R}\right). \quad (13)$$

In the same way, we may compute $C_1 = C_{-1}$,

$$C_1 = \frac{1}{RN} \sum_{j=0}^{\infty} \left(1 - \frac{1}{N}\right)^j 2\beta(R - 1)j = 2\beta(N - 1)\left(1 - \frac{1}{R}\right), \quad (14)$$

where again terms $0(\beta N)^2$ are neglected. To this order of accuracy, we may write

$$E_0(D_0) = E_0(C_1) + E_0(C_{-1}) = 4\beta(N - 1)\left(1 - \frac{1}{R}\right). \quad (15)$$

Remark: As $R \to \infty$ in (15) we obtain the result of GOLDSTEIN *et al.* (1995) derived from MORAN (1975). Note also that for large N and $N\beta$ small, C_0 in (13) approximates $(1 + 8N\beta)^{-1/2}$, the result obtained by KIMURA and OHTA (1973) and MORAN (1975). It is important to stress that the approximation made by neglecting $(N\beta)^2$ and higher powers may not be good, especially when R is small. It appears that in computing D_0, terms of the form $(\beta N)^k$ vanish, leaving only terms of the form $c_k(\beta N)^k/R$, where c_k are constants. For βN large and R small, these may be important. As an example, retaining the terms in $(\beta N)^2$ and $(\beta N)^3$ gives the extension of (14):

$$E_0(D_0) = 4\beta(N - 1)\left(1 - \frac{1}{R}\right) - \frac{4\beta^2}{R}(4N^2 - 7N + 3) + \frac{16}{R}\beta^3(N - 1)^2(2N - 1) + 0(\beta N)^4. \quad (16)$$

We conjecture that all higher powers of (βN) will also occur only with a divisor of R. Since $(N\beta)$ is typically quite large (GOLDSTEIN and CLARK 1995; GOLDSTEIN *et al.* 1995b), it will usually be necessary to determine $E_0 D_0$ by numerical evaluation of the matrix $\mathbf{Z}$ in (11).

To study the problem of differentiation between populations, suppose that a population has reached the equilibrium defined by (10), that two populations are formed from this progenitor population and that initially each has the same statistical configuration as the ancestral group. Subsequently, the evolution of the two groups, labeled a and b, occurs independently. Write ${}_a\boldsymbol{\nu}_t$ and ${}_b\boldsymbol{\nu}_t$ for the (row) vectors of allelic frequencies in the two groups. Then, as before, we have

$$E_{t-1}({}_a\boldsymbol{\nu}_t^T) = \mathbf{B}\,{}_a\boldsymbol{\nu}_{t-1}^T, \; E_{t-1}({}_b\boldsymbol{\nu}_t^T) = \mathbf{B}\,{}_b\boldsymbol{\nu}_{t-1}^T,$$

and, after sampling in the two populations,

$$E_{t-1}[{}_a\boldsymbol{\nu}_t^T \, {}_b\boldsymbol{\nu}_t] = \mathbf{B}\,{}_a\boldsymbol{\nu}_{t-1}^T \, {}_b\boldsymbol{\nu}_{t-1}\mathbf{B},$$

so that

$$E_0[{}_a\boldsymbol{\nu}_t^T\,{}_b\boldsymbol{\nu}_t] = \mathbf{B}^t E[{}_a\boldsymbol{\nu}_0^T\,{}_b\boldsymbol{\nu}_0]\mathbf{B}^t. \quad (17)$$

Here, $E[{}_a\boldsymbol{\nu}_0^T\,{}_b\boldsymbol{\nu}_0]$ represents the state of the initial population given by $\mathbf{Z}$ in (11). Hence

$$\lim_{t\to\infty} E_0[{}_a\boldsymbol{\nu}_t^T\,{}_b\boldsymbol{\nu}_t] = \frac{1}{RN}\sum_{j=0}^{\infty}\left(1-\frac{1}{N}\right)^j \lim_{t\to\infty}\mathbf{B}^{2(t+j)}. \quad (18)$$

Now each summand $\mathbf{B}^{2(t+j)}$ converges to the matrix $\hat{\mathbf{B}} = \|B_{ij}\|$ with $B_{ij} = 1/R$ at the rate $(1 - 2\beta + 2\beta\cos\pi/R)^2$, which is the leading nonunit eigenvalue of $\mathbf{B}^2$. We conclude that for each entry of the matrix on the left of (18),

$$\lim_{t\to\infty}\{E_0[{}_a\boldsymbol{\nu}_t^T\,{}_b\boldsymbol{\nu}_t]\}_{ij} = 1/R^2, \quad i, j = 1, 2\cdots R. \quad (19)$$

We may now compute the equilibrium value of $E_0(D_1)$,

$$\begin{aligned} E_0(D_1) &= 2\sum_{i=1}^{R} i^2 E_0(C_i^*) \\ &= 2\sum_{i=1}^{R} i^2 \sum_{j=1}^{R-i} 1/R^2 \\ &= 2\sum_{i=1}^{R} i^2(R-i)/R^2 \\ &= (R^2-1)/6. \end{aligned} \quad (20)$$

This is the value postulated originally by GOLDSTEIN *et al.* (1995a) and used by NAUTA and WEISSING (1996); it can be obtained directly by noting that the allele frequencies within a population approach a uniform distribution.

In populations a and b, denote the average repeat scores at time t by $r_a(t) = \Sigma\, i_a\nu_t^i$ and $r_b(t) = \Sigma\, i_b\nu_t^i$. GOLDSTEIN *et al.* (1995b) defined the distance between populations a and b by $(\delta\mu_t)^2 = (r_a(t) - r_b(t))^2$ at time t. It is easy to see (GOLDSTEIN *et al.* 1995b)

$$E_{t-1}(\delta\mu_t)^2 = E_{t-1}(D_1(t)) - V_a(t) - V_b(t),$$

where $V_a(t)$ and $V_b(t)$ are the variances in populations a and b at time t, and therefore $E_0(\delta\mu_t)^2$ approaches the limit

$$\tilde{\delta}(\tilde{\delta\mu})^2 = \frac{(R^2-1)}{6} - E_0D_0 \quad (22)$$

as $t\to\infty$. For βN small, E_0D_0 may be estimated as in (15); otherwise, it must be calculated numerically. Recall that without restriction on the range of repeat scores, $E_0(\delta\mu_t)^2$ grew as $4\beta\tau$, where τ is the time since populations a and b split from a common ancestral population. Under the assumption that the ancestral population had reached equilibrium (*i.e.*, $E_0(D_0)$ was at its equilibrium value), then $E_0(\delta\mu)^2$ changes at the same rate as $E_0(D_1)$, namely $(1 - 2\beta + 2\beta\cos\pi/R)^2$. When t is small and R is large, therefore, $(\delta\mu)^2$ will grow approximately as in the unconstrained case.

Statistical issues: We showed above that the rate of convergence of $E(D_1)$ and $E(\delta\mu)^2$ are governed by the second largest eigenvalue of the matrix $\mathbf{B}^2$, namely $(1 - 2\beta + 2\beta\cos\Pi/R)^2$. Here we focus on $(\delta\mu)^2$, since that seems to be the preferred distance, and write

$$E(\delta\mu)_i^2(t) = M(R_i)[1 - (1 - 2\beta_i + 2\beta_i\cos\pi/R_i)^{2t}], \quad (23)$$

where the subscript i represents a mutation rate and a range specific to locus i, and $M(R_i)$ is the maximum value of the distance. For β small, this equation can be approximated as

$$E(\delta\mu)_i^2(t) = M(R_i)\{1 - \exp[-(4\beta_i - 4\beta_i\cos\pi/R_i)t]\}. \quad (24)$$

The asymptotic behavior of the stepwise distance guarantees that after sufficient time has passed, it will provide a biased estimate of separation times. In such cases, it is sometimes useful to find a correction that results in a linear distance, although this often entails too high a price in terms of the variance of the distance (GOLDSTEIN and POLLOCK 1994). To obtain a distance whose expectation is linear with time in the general case where R and β vary across loci, it is necessary to take the product of the differences between the observed and maximal distances across loci. For convenience we will represent each of these differences as a fraction, P_i, of maximal distance. That is, $P_i = (M(R_i) - E(D_{1,i}))/M(R_i)$. To obtain a function linear with time using all of the loci, we take the product across loci of the P_i:

$$\prod_i P_i = \prod_i \frac{M(R_i) - E(D_{1,i}(t))}{M(R_i)} \quad (25)$$

$$= \prod_i \exp[-(4\beta_i - 4\beta_i\cos\pi/R_i)t]. \quad (26)$$

Taking the logarithm of the products of the P_i results in a distance that is linear with time.

$$\log\left[\prod_i P_i\right] = -\sum_i (4\beta_i - 4\beta_i\cos\pi/R_i)t \quad (27)$$

$$= C_1 t, \quad (28)$$

say, where $C_1 = -\Sigma_i(4\beta_i - 4\beta_i\cos\pi/R_i)$ is a constant given by the sets of β_i and R_i. When mutation rates and range constraints are known, time can be obtained directly using the reciprocal of C_1. When the mutation rates are not known, a linear function can be obtained as long as the R_i are known so that P_i can be obtained. Since C_1 is negative, the distance in this case would be obtained as $-\log[\prod_i P_i]$.

It is tempting to suggest $-\log[\prod_i P_i]$ as a genetic distance for microsatellites with range constraints, since its expectation is linear with time. Unfortunately, though (analytically) unbiased, the statistical properties

of this distance are highly undesirable. The basic problem is that P_i must be calculated for each locus separately. Since stepwise distances have a high variance (GOLDSTEIN *et al.* 1995a; ZHIVOTOVSKY and FELDMAN 1995), many of the P_i will be negative, even when the expectation of $(\delta\mu)^2$ is well below $M(R_i)$. When one or more of the P_i is negative, the argument of the logarithm is negative and the distance cannot be calculated. Consequently, all loci for which $P_i < 0$ must be either discarded from the analysis or set to an arbitrary value.

We used computer simulations (described below) to evaluate the behavior of $-\log [\prod_i P_i]$ as a distance. As might be expected, discarding atypically large values results in a substantial bias, causing $-\log [\prod_i P_i]$ to asymptote even earlier than $(\delta\mu)^2$ (data not shown). It is conceivable that some method that truncates large distances at an arbitrary value would work better, but we are not optimistic that the behavior can be substantially improved.

For completeness it should be noted that even in the unlikely case that $M(R_i)$ is sufficiently greater than the expected distance to ensure that no domain errors occur, the distance still would not be likely to perform well. Since the P_i vary stochastically, and the function required to restore linearity is highly nonlinear, the correction has a substantial bias. To determine how the linearized distance $-\log [\prod_i P_i]$ is influenced by variation in the P_i let $g(P_1, P_2, \ldots, P_L) = \log [\prod_i P_i]$. Then, the expectation of $g(P_1, P_2, \ldots, P_2)$ can be estimated using the delta method (see for example RICE 1995, p. 149):

$$E[g(P_1, P_2, \ldots, P_L)] \simeq g[E(P_1), E(P_2), \ldots, E(P_L)] + \frac{1}{2}\sum_i \mathrm{Var}\,(P_i)\frac{\partial^2}{\partial P_i^2} g[E(P_1), E(P_2), \ldots, E(P_L)]. \quad (29)$$

[To obtain (29), covariances have been neglected. ZHIVOTOVSKY and FELDMAN (1996) showed that these covariances are indeed negligible when the recombination between all pairs of genes is large relative to $1/N$. Our simulations indicate that these covariances are much less than the variances of P_i.] Because $\partial^2 g/\partial P_i^2 = -1/[E(P_i)]^2$ is negative, the expectation of the linearized distance is always a smaller negative number than $g[E(P_1), E(P_2), \ldots, E(P_L)]$. That is, the distance has the form $-(C_1 t - k)$ and this is larger than it should be (recall C_1 is negative). Furthermore, the magnitude of the error is proportional to the variance of P_i, which is known to be very large (ZHIVOTOVSKY and FELDMAN 1995). Although this bias might be correctable, statistically it seems hardly worthwhile to make the effort.

In summary, there are two difficulties in employing this correction to recover linearity. (1) Some of the P_i will be negative, causing a log domain error that cannot be eliminated without introducing a substantial bias, and (2) taking products results in very biased estimates even when a domain error does not occur, the bias being proportional to the variance of the distance at a single locus. We are not optimistic about the prospect of overcoming these problems with any analytic estimator.

A solution to these problems is straightforward, however, in the special case in which $\beta_i = \beta$ and $R_i = R$ for all i: that is, mutation rates and range constraints do not vary across loci. In this case we may work with the arithmetic average of the distances across loci as follows.

$$E\sum (\delta\mu)_i^2 = \sum M(R)(1 - \exp[-(4\beta - 4\beta\cos\pi/R)t]) \quad (30)$$

$$= LM(R)(1 - \exp[-(4\beta - 4\beta\cos\pi/R)t]), \quad (31)$$

where L is the number of loci. Then a linear distance based on an arithmetic average is obtained as,

$$\log\left\{\left[LM - \sum (\delta\mu)_i^2\right]/LM\right\} = \log\left[1 - \sum (\delta\mu)_i^2/LM\right] = -C_2 t. \quad (32)$$

That is, the proposed distance, D_L, is constructed by summing $(\delta\mu)^2$ across loci, dividing by L times the predicted maximum (Equation 22) and, subtracting this from 1, taking logarithms and dividing by $-C_2$. The improvement can readily be seen with a procedure similar to that used in (29). In this case, the arithmetic averaging results in a bias that falls with the square of the number of loci. After discounting for the difference between C_1 and C_2 in (28) and (32), the bias of the summed distance is $1/L$ times that of the product distance. More importantly, combining observed distances arithmetically greatly reduces the chance of a log domain error. Finally, since the variance of $\Sigma D_{1,i}$ as a proportion of LM decreases with the number of loci, the probability of a domain error goes down as the number of loci increases. This suggests that for a sufficiently large number of loci, all with the same mutation rate and range constraint, the log correction shown above might have practical use.

Computer evaluation: The properties of D_L as a genetic distance were tested using computer simulations similar to those described in GOLDSTEIN *et al.* (1995a). A single population is brought to equilibrium under stepwise mutation and drift and an identical copy is then made. The independent evolution of these two populations is simulated, and at set intervals of time, distances are calculated. Figure 1 shows the average behaviors of D_L and $(\delta\mu)^2$ for simulations using 40 and 80 loci. In the case of 40 loci, D_L is somewhat more linear than $(\delta\mu)^2$ after ~750 generations, but the difference is slight. Although in expectation D_L is linear with t for all t, in actuality when the expectation of $\Sigma(\delta\mu)_i^2$ is sufficiently great there is a high probability that the sum of the distances across loci will exceed the average maximal

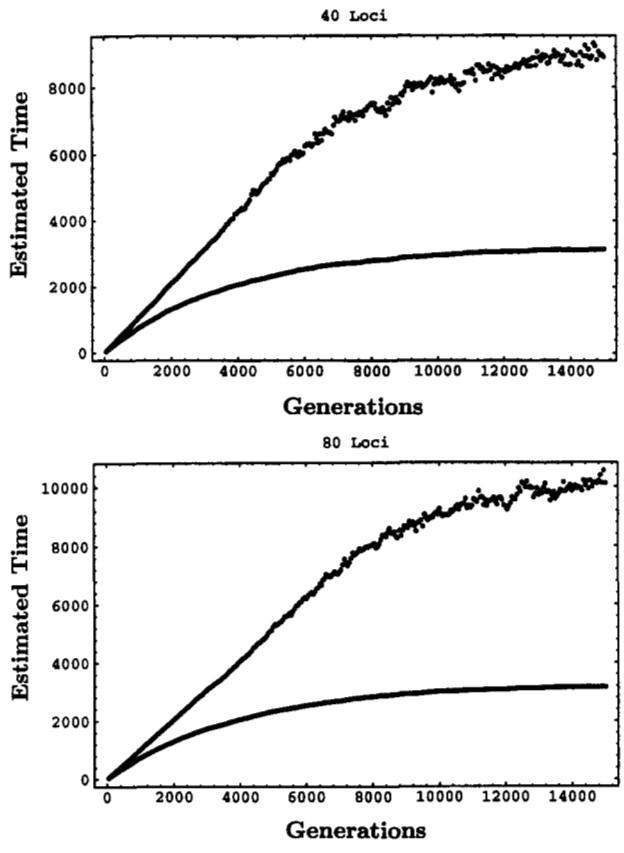

FIGURE 1.—Simulation results of two populations starting at the same equilibrium state. Each population has 100 individuals with either 40 or 80 loci each. The mutation rate per locus per generation was 0.01. Results reported here are averages over 500 replicates with different initial conditions. (Top) $(\delta\mu)^2$ and linearized distance (D_L) computed as averages over 40 loci (per individual), all having the same range constraint ($R = 20$). $(\delta\mu)^2$ has been divided by 4β to make it comparable to D_L. The x axis is the number of generations, and the y axis is the estimated time calculated from the observed distance. [Upper curve D_L, lower curve $(\delta\mu)^2$.] (Bottom) Same as A but with the averages over 80 loci (per individual) of $(\delta\mu)^2$ and D_L.

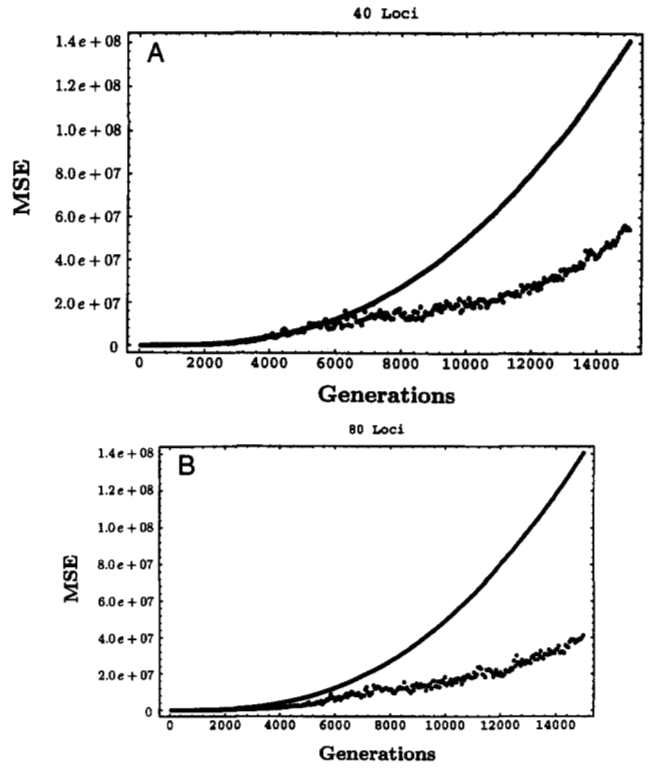

FIGURE 2.—Mean square errors. (A) The mean square error (MSE) (see text) of $(\delta\mu)^2$ and D_L as functions of separation time between two populations of individuals with 40 loci each (averaged over all 40 loci). [Upper curve $(\delta\mu)^2$, lower curve D_L.] (B) Same as A but with two populations of individuals with 80 loci each (averaged over all 80 loci).

divergence, LM, resulting in a log domain error. As noted above, we expect that for given t, the chance of satisfying the inequality $LM - \Sigma(\delta\mu)_i^2 < 0$ (thus causing a log domain error) will decrease with the number of loci, thereby extending the time over which D_L will remain linear with t. The simulation with 80 loci confirms this expectation. In this case D_L is approximately linear until generation 10,000 or so. As would be expected, the linearity of $(\delta\mu)^2$ is not affected by the number of loci; therefore D_L, unlike $(\delta\mu)^2$, has the interesting property that its useful temporal range can be extended by increasing the number of loci studied.

To compare the overall reliability of the two distances, we use the mean squared error (MSE). Suppose that the values of either $(\delta\mu)^2$ or D_L are observations on the following process: $\Delta(t) = t + b + \epsilon$, where $\Delta(t)$ is the estimated time using either $(\delta\mu)^2$ or D_L as the time estimator, t is the true time of separation, b is a constant, or systematic, error, often called the bias, and ϵ is the random component of the error. Here ϵ is a random variable with $E[\epsilon] = 0$ and Var $(\epsilon) = \sigma^2$. An overall measure of the size of the error is the MSE, which is defined as

$$\text{MSE} = E[(\Delta(t) - t)^2].$$

MSE can be decomposed into contributions from the bias and the variance as follows:

$$\text{MSE} = b^2 + \sigma^2.$$

The values of $(\delta\mu)^2$ and D_L are computed as the average over all loci, either 40 or 80, and over all 500 replicates of the simulation. The values of MSE and σ^2 are measured over the 500 replicates. Figure 2A demonstrates that, due to the asymptotic behavior of $(\delta\mu)^2$, its MSE very quickly becomes substantially larger than that of D_L. Figure 2B shows that this difference increases with the number of loci, as expected.

Another way to assess the expected performance of these distances in phylogeny reconstruction uses the accuracy index introduced by TAJIMA and TAKEZAKI (1994). Very early in the evolution $(\delta\mu)^2$ has a higher accuracy, but for the majority of times the accuracies of the two distances are nearly identical (data not shown). For distantly related groups we suspect that D_L

may perform better, especially with a large number of loci. Since the slope of $(\delta\mu)^2$ is not sensitive to the number of loci, and the point at which the linearity of D_L is lost increases with the number of loci, it seems clear that for a sufficiently large number of loci and sufficiently large t, D_L will have a higher accuracy than $(\delta\mu)^2$. These theoretical considerations suggest that D_L has potential as a distance for microsatellite loci. The major difficulty in its implementation is finding a large number of microsatellite loci with sufficiently similar mutation rates and range constraints. Fortunately, it would appear that D_L is not highly sensitive to moderate variation in β and R.

To assess the sensitivity of D_L to rate variation, we compared simulations with constant and variable ranges and mutation rates. We considered 30 loci in which the mutation rate was either fixed at 0.045 or chosen randomly from the interval 0.015–0.075 for an average of 0.045. Similarly, the range R was either fixed at 60 or chosen randomly from an interval of 20–100 for an average of 60. Figure 3 presents the results from the simulation with constant R and β, while Figure 4 show the simulations with variable R and β. The most striking aspect of these results is that even though the transformation is analytically incorrect when R and β vary, applying this transformation when either R or β vary over a factor of 5 (Figure 4) results in behavior almost identical to that when the transformation is applied to a set of loci with fixed values that are the averages of the varying values (Figure 3). That is, under variation in either R or β, D_L remains much more linear than $(\delta\mu)^2$ and in fact is nearly as linear as it is in the case of no variation among loci. Furthermore, the MSE of the transformed distances remains well below that of $(\delta\mu)^2$ for the bulk of the evolution, and is even generally below that of D_L applied to the case of no variation (Figure 4B). These results suggest that D_L, despite being an inappropriate transformation analytically, nevertheless behaves very well in the case of variation in β and R.

Careful inspection of Figure 4B also reveals that the MSE of D_L in the case of fivefold variation in R is somewhat lower than in the case of fivefold variation in β. This suggests that D_L is less sensitive to variation in R than variation in β, as might be predicted based on the expectation shown in (31). Notice that the only place where variation in R will affect linearity is in the term $\cos \pi/R$ inside the exponential. This is the only important term because the leading $LM(R)$ averages, and the term against the exponential will only contribute an additive term upon log transformation. Because $\cos \pi/R$ in the exponential is asymptotic in R, as long as all loci have reasonably high values of R (say above 25 or so), variation in R is expected to have a relatively small impact. Variation in β is expected to be more important because it is linear inside the exponential term. Thus, in applying D_L it would seem that clumping loci into sets with similar mutation rates will be more critical than clumping them on the basis of R.

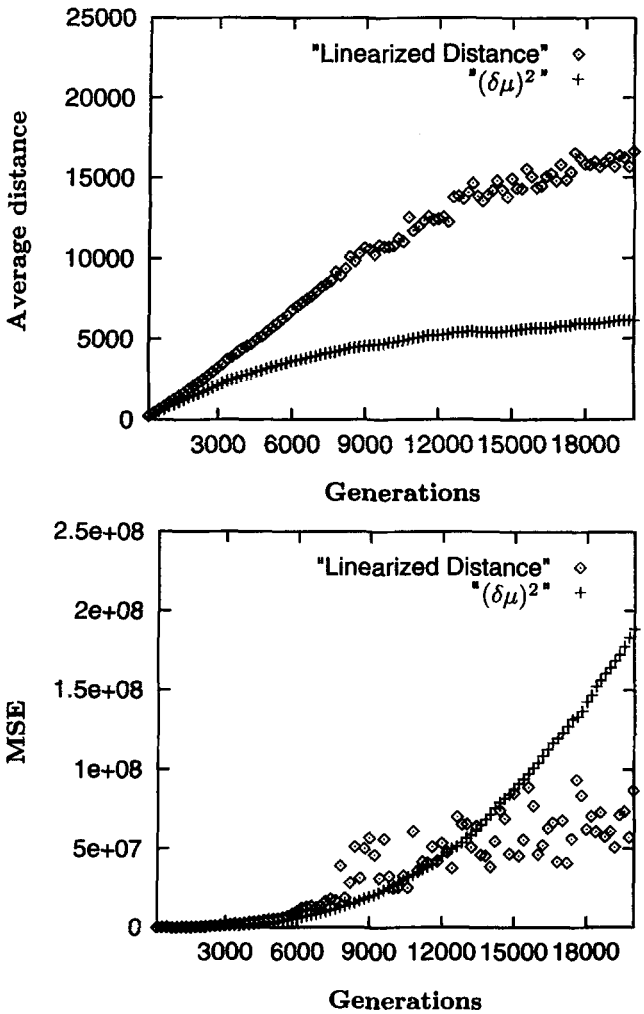

FIGURE 3.—Constant parameter simulations. Conditions are the same as above except that the mutation rate is 0.045, R is 60, and the population size is 60. Top: the average distances among 150 replications; bottom: the MSE for each of the two distances.

DISCUSSION

The tremendous variability of microsatellite loci has established them as the preferred markers for most intraspecific applications. Their use in interspecific phylogeny reconstruction, however, has been much less successful. One reason for this seems to be that at least some microsatellites degrade quickly, causing their mutation rates and other parameters to differ greatly from taxon to taxon (ELLEGREN *et al.* 1995; RUBINSZTEIN *et al.* 1995; GOLDSTEIN *et al.* unpublished data). As a result, distances at these loci do not appear well correlated with time for more divergent taxa. Another difficulty is that even when microsatellites persist over sufficiently long intervals, they nevertheless fail to reflect separation times among species due to constraints on maximal allele size (GARZA *et al.* 1995; GOLDSTEIN *et al.* 1995a). Such constraints, which

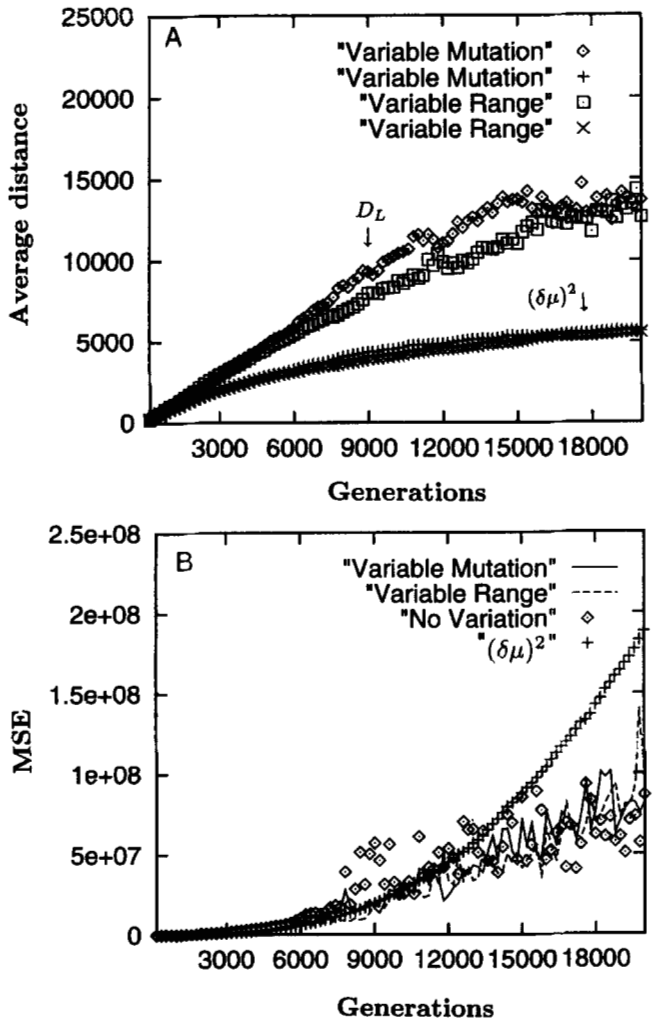

FIGURE 4.—Variable parameter simulations. Conditions are the same as in Figure 3 except that either the mutation rate is randomly drawn from the interval (0.015, 0.075) (average of 0.045 as used above) or the range constraint is randomly drawn from the interval (20, 100) (average of 60 as for Figure 3). (A) The average distances among 150 replications. The two upper curves are the transformed distances (applied in the case of variable R or variable β) and the two lower curves are $(\delta\mu)^2$. (B) The MSE for the linearized distance applied to all three cases (no variation, variation in R, and variation in β). The MSE for $(\delta\mu)^2$ is the same for all cases, so only one curve is shown.

have been well established empirically (BECKMANN and WEBER 1992; BOWCOCK *et al.* 1994), mean that microsatellites will eventually lose their phylogenetic information (GOLDSTEIN *et al.* 1995a,b). Here we have outlined an analytic method for characterizing the effect of a specific constraint on maximal allele size. Based on this analytic characterization, we have developed new distances that recover the linear relation with time even when ranges are constrained.

Our analysis incorporates what seems to us the simplest mutation process with fixed range constraints. For all i allelic states for $1 < i < R$ symmetric stepwise mutations occur at a rate β in each direction. For the boundary states 1 and R, mutation only increases or decreases size, respectively, again at a rate of β. In reality, it is clear that the mutation process has a much more complicated dependence on allele size (RICHARDS and SUTHERLAND 1994; GOLDSTEIN and CLARK 1995). Unfortunately, the exact details of the mutation process are not sufficiently well understood to allow an accurate representation. Furthermore, they are likely to vary from one type of microsatellite to another. As such details become available it will be worthwhile to develop more complicated models. In theory, the analytical approach employed here allows arbitrary assumptions about the mutation process since an explicit transition matrix is used. In practice, however, it may be that all but the simplest matrices will resist direct analysis. It seems reasonable to assume, however, that the primary conclusions obtained here reflect the existence of range constraints and are likely to apply when other assumptions are made about the exact details of the process imposing the limits on size. This conjecture is supported by the qualitative similarity of our results to those of GARZA *et al.* (1995) who used different analytic techniques and assumed a very different mutation process.

One important general point to emerge from this study is that it is probably not possible to develop a formally accurate analytic distance for microsatellite loci in the general case in which mutation rates and range constraints vary across loci. This follows from the form of the distance under range constraints, which necessitates multiplying distances across loci before logarithms are taken. Because of the high variance of the underlying distance, this process invariably results in log domain errors and is therefore essentially useless. On the other hand, we have shown that if a large set of loci with sufficiently similar mutation rates and range constraints can be identified, a simple analytic correction can provide a distance that is much less biased than existing distances. An interesting and unusual property of the new distance is that its linearity increases with the number of loci. Most importantly, we have used computer simulations to show that even though it is formally insufficient to recover linearity in the case of variation among loci, the simple correction is not strongly sensitive to such variation. If loci can be clustered at least to within a factor of 5 or so, D_L can be expected to substantially improve linearity.

Our analysis also provides a framework for comparing the expected behaviors of different distances as functions of the particular loci and phylogenetic problem under study. For example, for studies involving a relatively large number of taxa, it is possible to obtain estimates of the parameters of each locus under study (POLLOCK *et al.* 1996). Once the parameters have been estimated, observed distances can be compared with (24) to assess whether the separation time is beyond the useful range of the locus. Furthermore, it is possible

to use the analytic expectations developed here to compare the behavior of stepwise and nonstepwise distances. In particular, it is important to delimit the boundaries of the window of time within which stepwise distances outperform nonstepwise distances (GOLDSTEIN *et al.* 1995a,b). These considerations suggest that variation at certain microsatellite loci might not be useful for some phylogenetic problems. Just as one would not use sequence variation in histones or other slowly evolving proteins to assess relationships among closely related species, one must choose microsatellite appropriate to the phylogenetic problem under study. For example, in studying relationships among more divergent taxa, it is necessary to choose microsatellites that (1) have not degraded in either taxon, (2) have sufficiently large ranges, and (3) have a sufficiently small mutation rate. It is well known that these parameters vary among types of microsatellite loci (WEBER and WONG 1993). Therefore, it is little wonder that the application of microsatellites chosen strictly on the basis of polymorphism in a focal species has been of relatively little use in recovering interspecific relationships. The challenge now is to develop the machinery necessary to estimate the parameters of microsatellite loci. Once these parameters are estimated, models like the one described here can be used to partition microsatellites into categories appropriate for different kinds of phylogenetic problems. Finally, the statistical properties of the distance to be used must be considered to determine the minimum number of loci required for a specified degree of accuracy. In general, we expect that the requisite number will be somewhat larger than is customary at present. While the exact numbers will depend on the phylogenetic problem at hand and on the exact parameter values estimated, we doubt that fewer than ~15–20 microsatellites will ever provide very reliable estimates of interspecific relationships. Furthermore, we expect that some phylogenetic problems will require characterization of a great many more loci. Fortunately, eukaryotic genomes seem to harbor sufficient numbers of microsatellites that reliable information should be obtainable if one is willing to invest the effort.

The authors thank Drs. NAUTA and WEISSING for providing an advance copy of their manuscript. This research supported in part by National Institutes of Health (NIH) grant GM-28428 to M.W.F., and an NIH National Service award to D.B.G.

LITERATURE CITED

BARNETT, S., 1990 *Matrices: Methods and Applications.* Oxford University Press, Oxford.

BOWCOCK, A. M., A. RUIZ LINARES, J. TOMFOHRDE, E. MINCH, J. R. KIDD *et al.*, 1994 High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368:** 445–457.

CHAKRABORTY, R., and M. NEI, 1977 Bottleneck effects on average heterozygosity and genetic distance with the stepwise mutation model. Evolution **31:** 347–356.

DEKA, R., L. JIN, M. D. SHRIVER, L. M. YU, S. DECROO *et al.*, 1995 Population genetics of dinucleotide $(dC\text{-}dA)_n(dG\text{-}dT)_n$ polymorphisms in world populations. Am. J. Hum. Genet. **56:** 461–474.

ESTOUP, A., L. GARNERY, M., SOLIGNAC and J. CORNUET, 1995 Microsatellite variation in honey bee (*Apis Mellifera L.*) populations: hierarchical genetic structure and test of the infinite allele and stepwise mutation models. Genetics **140:** 679–695.

GARZA, J. C., M. SLATKIN and N. B. FREIMER, 1995 Microsatellite allele frequencies in humans and chimpanzees, with implications for constraints on allele size. Mol. Biol. Evol. **12:** 594–603.

GOLDSTEIN, D. B., and A. G. CLARK, 1995 Microsatellite variation in North American populations of *Drosophila melanogaster.* Nucleic Acids Res. **23:** 3882–3886.

GOLDSTEIN, D. B., and D. D. POLLOCK, 1994 Least squares estimation of molecular distance—noise abatement in phylogenetic reconstruction. Theor. Popul. Biol. **45:** 219–226.

GOLDSTEIN, D. B., A. RUIZ LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995a An evaluation of genetic distances for use with microsatellite loci. Genetics **139:** 463–471.

GOLDSTEIN, D. B., A. RUIZ LINARES, L. L. CAVALLI-SFORZA and M. W. FELDMAN, 1995b Genetic absolute dating based on microsatellites and modern human origins. Proc. Natl. Acad. Sci. USA **92:** 6723–6727.

GOLDSTEIN, D. B., L. A. ZHIVOTOVSKY, K. NAYAR, A. RUIZ LINARES, L. L. CAVALLI-SFORZA *et al.*, 1995c Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y chromosomes. Mol. Biol. Evol. (in press).

HENDERSON, S., and T. PETES, 1992 Instability of simple sequence DNA in *Sacharomyces cerevisiae.* Mol. Cell Biol. **12:** 2749–2757.

LAGERCRANTZ, U., H. ELLEGREN and L. ANDERSON, 1993 The abundance of various polymorphic microsatellite motifs differs between plants and vertebrates. Nucleic Acids Res. **21(5):** 1111–1115.

MACHUGH, D. E., R. T. LOFTUS, D. G. BRADLY, P. M. SHARP and P. CUNNINGHAM, 1994 Microsatellite DNA variation within and among European cattle breeds. Proc. Roy. Soc. Lond. B **256:** 25–31.

MORAN, P., 1975 Wandering distributions and the electorphoretic profile. Theor. Pop. Biol. **8:** 318–330.

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

NEI, M., 1987 *Molecular Evolutionary Genetics.* Columbia University Press, New York.

NEI, M., and R. CHAKRABORTY, 1973 Genetic distance and electrophoretic identity of proteins between taxa. J. Mol. Evol. **2:** 323–328.

OHTA, T., and K. KIMURA, 1973 A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22:** 201–204.

PEPIN, L., Y. AMIGUES, A. LEINGLE, J. L. BERTHIER, A. BENSAID *et al.*, 1995 Sequence conservation of microsatellites between *Bos taurus* (cattle) and *Capra hircus* (goat) and related species. Heredity **74:** 53–61.

POLLOCK, D. D., A. BERGMAN, M. W. FELDMAN and D. B. GOLDSTEIN, 1996 An improved distance estimator for microsatellites with range constraints. Manuscript.

RICE, J. A., 1995 *Mathematical Statistics and Data Analysis,* Ed. 2. Duxbury Press, Belmont, CA.

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

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

TAJIMA, F., and N. TAKEZAKI, 1994 Estimation of evolutionary distance for reconstructing molecular phylogenetic trees. Mol. Biol. Evol. **11:** 278–286.

TAUTZ, D., 1993 Note on the definition and nomenclature of tandemly repetitive DNA sequences, pp. 21–28 in *DNA Fingerprinting: State of the Science,* edited by S. D. J. PENA, J. T. EPLEN and A. J. JEFFREYS. Birkhauser Verlag, Basel.

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

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

Communicating editor: M. SLATKIN