Genetic Distances and Reconstruction of Phylogenetic Trees From Microsatellite DNA

Naoko Takezaki and Masatoshi Nei

Institute of Molecular Evolutionary Genetics and Department of Biology, The Pennsylvania State University, University Park, Pennsylvania 16802

Manuscript received February 29, 1996
Accepted for publication June 6, 1996

ABSTRACT

Recently many investigators have used microsatellite DNA loci for studying the evolutionary relationships of closely related populations or species, and some authors proposed new genetic distance measures for this purpose. However, the efficiencies of these distance measures in obtaining the correct tree topology remains unclear. We therefore investigated the probability of obtaining the correct topology (P_C) for these new distances as well as traditional distance measures by using computer simulation. We used both the infinite-allele model (IAM) and the stepwise mutation model (SMM), which seem to be appropriate for classical markers and microsatellite loci, respectively. The results show that in both the IAM and SMM CAVALLI-SFORZA and EDWARDS' chord distance (D_C) and NEI *et al.*'s D_A distance generally show higher P_C values than other distance measures, whether the bottleneck effect exists or not. For estimating evolutionary times, however, NEI's standard distance and GOLDSTEIN *et al.*'s $(\delta\mu)^2$ are more appropriate than other distances. Microsatellite DNA seems to be very useful for clarifying the evolutionary relationships of closely related populations.

ALLELE frequency data are useful for studying evolutionary relationships of closely related species or populations. In this case it is customary to use some genetic distance measures for constructing trees. Using computer simulation, NEI *et al.* (1983) examined the relative efficiencies of different distance measures for obtaining the correct tree topology with the assumption that new mutant alleles are always different from the existing ones in the population [infinite-allele model (IAM); KIMURA and CROW 1964]. This model seems to apply approximately to classical genetic markers such as protein and blood group polymorphisms (NEI 1987). In recent years, however, microsatellite DNA loci are often used for phylogenetic analysis (*e.g.*, BOWCOCK *et al.* 1994; ROY *et al.* 1994; DEKA *et al.* 1995), and these loci seem to be subject to a mutational pattern that roughly follows OHTA and KIMURA's (1973) stepwise mutation model (SMM). For this reason, GOLDSTEIN *et al.* (1995a) and SHRIVER *et al.* (1995) developed new genetic distance measures.

However, the efficiencies of these distance measures in phylogenetic reconstruction compared with those of traditional distance measures are unclear when they are applied to microsatellite DNA loci. Furthermore, NEI *et al.*'s (1983) study is based on a small number of replications, so that a more careful study is necessary even for the IAM. The purpose of this paper is to study the efficiencies of genetic distance measures in phylogenetic reconstruction by using computer simulation. In this paper we will consider only two methods of phylogenetic reconstruction: the neighbor-joining (NJ) method (SAITOU and NEI 1987) and the unweighted pair group method with arithmetic mean (UPGMA; SNEATH and SOKAL 1973). UPGMA seems to be useful for allele frequency data when the evolutionary rate is nearly the same for all populations (NEI 1987), whereas the NJ method is known to be applicable for a variety of situations (NEI 1991). UPGMA usually produces a rooted tree, but in this paper we removed the root to make a fair comparison with the NJ method that produces only unrooted trees. In this paper we are primarily interested in the accuracy of the phylogenetic tree topology obtained, but the time dependency and sampling error of each distance measure will also be considered. Preliminary results of some parts of this study have been published by NEI and TAKEZAKI (1994).

Corresponding author: Masatoshi Nei, Institute of Molecular Evolutionary Genetics, The Pennsylvania State University, 328 Mueller Laboratory, University Park, PA 16802.
E-mail: nxm2@psu.edu

METHODS OF COMPUTER SIMULATION

Genetic distance measures: Various genetic distance measures used for gene frequency data have been described by NEI (1987). Here, we present only the definitions and brief explanations of the distance measures examined in this study.

NEI's (1972) standard genetic distance (D_S):

$$D_S = -\ln [J_{XY}/\sqrt{J_X J_Y}], \quad (1)$$

where $J_X = \sum_j^r \sum_i^{m_j} x_{ij}^2/r$ and $J_Y = \sum_j^r \sum_i^{m_j} y_{ij}^2/r$ are the average homozygosities over loci in populations X and Y, respectively, and $J_{XY} = \sum_j^r \sum_i^{m_j} x_{ij}y_{ij}/r$. Here, x_{ij} and y_{ij}

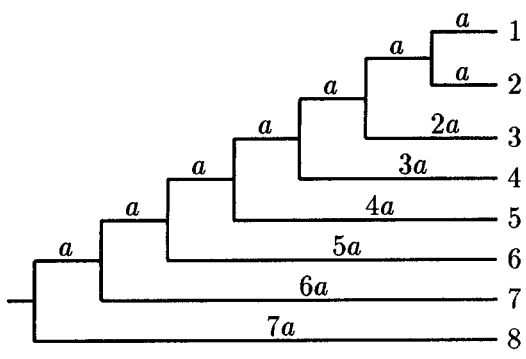

FIGURE 1.—Model tree used for computer simulation. a is a branch length in unit of expected number of mutations accumulated per locus (vt). a was 0.1 and 0.04 for the IAM, and 0.4 and 0.04 for the SMM.

are the frequencies of the ith allele at the jth locus in populations X and Y, respectively, m_j is the number of alleles at the jth locus, and r is the number of loci examined. Under the IAM, D_S is expected to increase linearly with time, if the mutation-drift balance is maintained throughout the evolutionary process. That is, $E(D_S) = 2vt$ where v is a mutation rate per locus per generation and t is the time measured in generations after divergence of the two populations.

NEI's (1973) minimum genetic distance (D_m):

$$D_m = (J_X + J_Y)/2 - J_{XY}. \tag{2}$$

The expectation of D_m is known to be $E(D_m) = J(1 - e^{-2vt})$, where J is the expected homozygosity of the two populations.

LATTER's (1972) F_{ST} ($=\phi^*$) distance:

$$F_{ST} = \frac{(J_X + J_Y)/2 - J_{XY}}{1 - J_{XY}}. \tag{3}$$

REYNOLDS *et al.*'s (1983) θ_W becomes essentially the same as this F_{ST} when sample size is large. The measure $D_L = -\ln(1 - F_{ST})$ has also been proposed (LATTER 1972; REYNOLDS *et al.* 1983), but our study has shown that the efficiency of D_L in obtaining the correct tree topology is almost the same as or slightly lower than that of F_{ST} (NEI and TAKEZAKI 1994; unpublished results). Therefore, the results for D_L will not be considered in this paper.

ROGERS' (1972) distance:

$$D_R = \frac{1}{r}\sum_j^r \sqrt{\frac{\sum_i^{m_j}(x_{ij} - y_{ij})^2}{2}}. \tag{4}$$

PREVOSTI *et al.*'s (1975) distance has a statistical property similar to that of D_R and is defined as

$$C_p = \sum_j^r \sum_i^{m_j} |x_{ij} - y_{ij}|/(2r). \tag{5}$$

CAVALLI-SFORZA and EDWARDS' (1967) chord distance:

$$D_C = (2/\pi r)\sum_j^r \sqrt{2\left(1 - \sum_i^{m_j}\sqrt{x_{ij}y_{ij}}\right)}. \tag{6}$$

If we represent two populations on the surface of a multidimensional hypersphere using allele frequencies at the jth locus, D_C for the locus gives the chord distance between the two populations, where the angle (θ_j) for the two populations is given by $\cos\theta_j = \Sigma_i^{m_j}\sqrt{x_{ij}y_{ij}}$. By contrast, BHATTACHARYYA (1946) and NEI (1987) suggested that the distance between the two populations be measured by $\theta^2 = \Sigma_j^r(\arccos\theta_j)^2/r$. However, the performance of this distance measure in obtaining the correct topology is essentially the same as that of D_A given below, so that we shall not consider this measure in this paper (see NEI and TAKEZAKI 1994 for the results for the IAM).

NEI *et al.*'s (1983) D_A distance:

$$D_A = 1 - \frac{1}{r}\sum_j^r \sum_i^{m_j}\sqrt{x_{ij}y_{ij}}. \tag{7}$$

NEI *et al.*'s (1983) computer simulation with the IAM has shown that this distance is more efficient than D_S, D_m, D_R, and CAVALLI-SFORZA's (1969) f_θ in obtaining the correct topology. SANGHVI's (1953) X^2 distance is closely related to θ^2 and is given by

$$X^2 = \sum_j^r \sum_i^{m_j} 2\frac{(x_{ij} - y_{ij})^2}{(x_{ij} + y_{ij})}\Big/ r. \tag{8}$$

This is approximately given by $4\theta^2$ when θ_j's are small (NEI 1987).

In recent years many authors have started to use microsatellite DNAs that are repeats of generally two to five nucleotides. There is a high degree of polymorphism with respect to the number of repeats at a microsatellite locus. An allele at this locus is usually represented by the number of tandem repeats, and this number may increase or decrease by mutation roughly following the SMM, as mentioned earlier. In the SMM an allele in state i (an allele with i repeats) is assumed to mutate to an allele either in state $i + 1$ or $i - 1$ with an equal probability. With this assumption, GOLDSTEIN *et al.* (1995a) proposed that the following distance measure be used for microsatellite loci.

$$(\delta\mu)^2 = \sum_j^r (\mu_{X_j} - \mu_{Y_j})^2/r, \tag{9}$$

where $\mu_{X_j}(=\Sigma_i ix_{ij})$ and $\mu_{Y_j} = (\Sigma_i iy_{ij})$ are average allelic states at the jth locus, and x_{ij} and y_{ij} are the frequencies of the allele in state i at the jth locus in populations X and Y, respectively.

A distance measure closely related to $(\delta\mu)^2$ is the average square distance (ASD) (GOLDSTEIN *et al.* 1995b; SLATKIN 1995), defined as $ASD = \Sigma_k^r \Sigma_{i,j}(i - j)^2 x_{ik}y_{jk}/r$. Under the assumption of mutation-drift balance, the expected values of $(\delta\mu)^2$ and ASD are given by $E[(\delta\mu)^2] = 2vt$ and $E(ASD) = 2vt + 2(2N - 1)v$, respectively,

TABLE 1

Percentage of replications in which the correct topology was obtained for the infinite-allele model

	NJ							Unrooted UPGMA						
No. loci	D_R	D_m	D_S	F_{ST}	X^2	D_C	D_A	D_R	D_m	D_S	F_{ST}	X^2	D_C	D_A
						$a = 0.1$; $H = 0.5$								
10	8	7	4	6	13	17	17	10	7	9	6	14	17	19
30	33	36	25	38	57	58	62	42	42	43	27	57	59	63
50	55	60	44	65	73	77	80	58	61	62	56	81	80	83
100	82	85	73	87	92	94	94	88	88	87	80	96	95	98
						$a = 0.1$; $H = 0.16$								
10	5	5	3	7	7	6	7	2	4	3	3	4	3	4
30	24	29	17	32	32	31	36	21	24	24	15	27	24	30
50	45	48	34	51	56	50	59	39	42	42	21	43	41	45
100	71	74	60	78	77	74	80	67	68	68	46	70	69	73
						$a = 0.004$; $H = 0.5$								
10	10	5	5	6	22	31	25	9	8	7	7	33	33	31
30	35	24	22	24	63	71	69	46	40	40	39	74	80	80
50	49	39	39	41	84	89	88	65	63	60	64	93	95	95
100	81	72	72	72	98	100	98	90	87	87	86	99	100	100
						$a = 0.004$; $H = 0.16$								
10	1	2	2	2	2	3	3	2	1	1	2	1	3	2
30	5	4	4	5	10	18	15	5	6	5	4	14	13	13
50	11	7	7	8	22	31	28	14	11	11	11	26	26	31
100	32	22	21	27	44	58	54	41	33	32	28	53	57	62

$a = 0.1$. $N = 100$ and $v = 0.0025$ for $H = 0.5$. $N = 100$ and $v = 0.0005$ for $H = 0.16$. $a = 0.004$. $N = 250$ and $v = 0.001$ for $H = 0.5$. $N = 100$ and $v = 0.0005$ for $H = 0.16$.

where v is the mutation rate per locus per generation and N is the effective population size. Therefore, both measures are linearly related to evolutionary time t. However, *ASD* has a larger variance than $(\delta\mu)^2$, so that we consider only $(\delta\mu)^2$ in this paper. Another related distance measure is SHRIVER *et al.*'s (1995), which is given by

$$D_{SW} = W_{XY} - (W_X + W_Y)/2, \qquad (10)$$

where $W_X = \Sigma_k^r \Sigma_{i\neq j} |i - j| x_{ik}x_{jk}/r$, $W_Y = \Sigma_k^r \Sigma_{i\neq j} |i - j| y_{ik}y_{jk}/r$, $W_{XY} = \Sigma_k^r \Sigma_{i\neq j} |i - j| x_{ik}y_{jk}/r$.

Simulation procedures: The method of computer simulation was essentially the same as that of NEI *et al.* (1983). Figure 1 shows the model tree used in the simulation. The allele frequency data for monoecious diploid populations were generated by introducing mutations and sampling alleles at random in each generation. In the case of the IAM, it was assumed that whenever a mutation occurs, it creates a new allele. In the SMM, a mutation was assumed to create a new allele as described earlier. When a population was split into two descendant populations, the initial allele frequencies of the two populations were assumed to be identical with those of the ancestral one, and the reproductive isolation occurred immediately. In the study of the effects of population bottlenecks (population size change), we assumed that the size of populations 2, 5, and 8 in Figure 1 was reduced to half the original size as soon as population splitting occurred. This scheme of simulation is certainly artificial. But it will give some idea about the bottleneck effect on reconstruction of phylogenetic trees. Allele frequency data were generated for 10, 30, 50, . . . , independent loci, and NJ and UPGMA trees were constructed by using various genetic distance measures. After 200 replications, the percentage (P_C) of replications in which the correct topology was obtained was calculated.

The expected number of gene substitutions per locus along the shortest branch ($a = vt$) of the model tree was 0.1 or 0.004 for the IAM, as was in the case of NEI *et al.*'s (1983) simulation. The value of $a = 0.1$ is intended to represent trees for different species, whereas $a = 0.004$ is appropriate for trees constructed for populations within species (see NEI 1987, p. 242). In the case of the SMM, a was set to 0.4 and 0.04. When $a = 0.4$, the largest expected distance (5.6) between populations is similar to the observed distance between African and non-African human populations for 30 microsatellite loci (GOLDSTEIN *et al.* 1995a). The value of $a = 0.04$ is intended to represent the divergence level of very

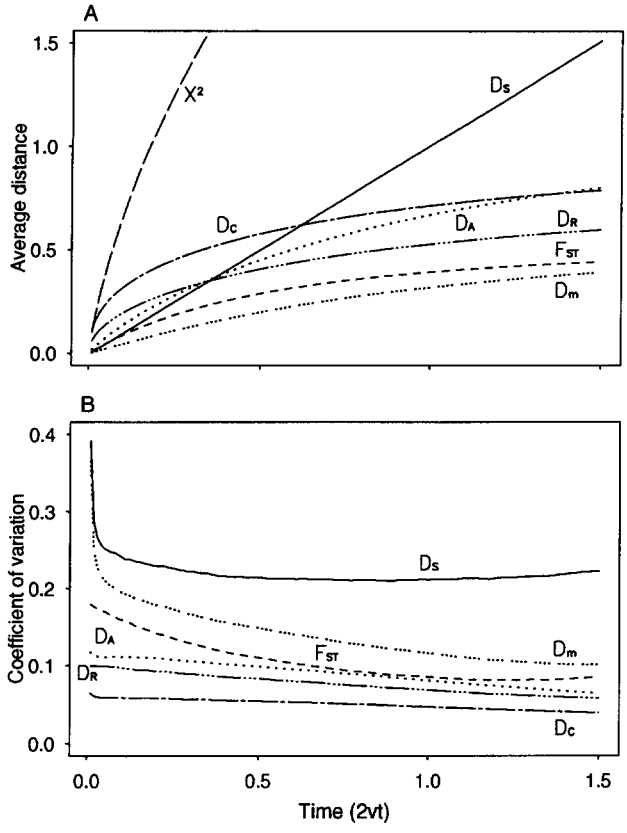

FIGURE 2.—Relationships of the mean distance and coefficient of variation with evolutionary time ($2vt$) for the case of the IAM with $H = 0.5$. (A) Mean distance. (B) Coefficient of variation (CV). The means and CV's for each distance measure were computed for 100 loci and 30 loci, respectively, with 1000 replications.

closely related populations. The divergence level between species can be much greater than that expected for the case of $a = 0.4$. However, the conclusion from the case of $a = 0.4$ essentially holds for trees of such highly divergent populations, as will be shown later.

The expected heterozygosity (H) was assumed to be 0.5 or 0.16 for the IAM and 0.8 or 0.5 for the SMM. For the IAM, $H = 0.5$ represents a high level of heterozygosity and $H = 0.16$ a low level of heterozygosity for classical genetic markers (NEI 1987, p. 197). In the case of microsatellite loci, the average heterozygosity seems to be generally between 0.5 and 0.8 (BOWCOCK *et al.* 1994; GOTTELLI *et al.* 1994; ROY *et al.* 1994; DEKA *et al.* 1995; FORBES *et al.* 1995; JORDE *et al.* 1995; PAETKAU *et al.* 1995). Theoretically, H is a function of an effective population size (N) and a mutation rate (v) per generation. That is, $H = 4Nv/(1 + 4Nv)$ for the IAM and $H = 1 - 1/\sqrt{1 + 8Nv}$ for the SMM. Various values of N and v were used to generate the H values mentioned above, but to save computer time we used N, which was ≤ 300. NEI *et al.*'s (1983) results and our preliminary study have shown that the P_C value primarily depends on vt and Nv rather than on v, t, and N, separately, as expected from LI and NEI's (1975) theoretical study of the means and variances of D_m and D_S. This was the case even in the presence of bottleneck effects.

RESULTS

Infinite-allele model: *Constant population size:* Table 1 shows the P_C values for the case of constant population size for the IAM. The P_C values for distance measure C_p are not shown here because they are similar to those for D_R (see NEI and TAKEZAKI 1994). This table also does not include P_C's for $(\delta\mu)^2$ and D_{SW}, because these distances are not computable in this case.

The P_C values for all the traditional distance measures for 10 loci are quite low whether the levels of divergence and heterozygosity are high or low. In many cases, P_C values are $<10\%$. This indicates that phylogenetic trees constructed with a small number of loci are not reliable. As the number of loci increases, however, P_C gradually increases. Therefore, it is important to use a large number of loci for constructing a population tree (NEI *et al.* 1983). Irrespective of the divergence level ($a = 0.1$ or 0.004), P_C is higher for a high heterozygosity level ($H = 0.5$) than for a low heterozygosity level ($H = 0.16$). When H is high, P_C is higher for a low divergence level ($a = 0.004$) than for a high divergence level ($a = 0.1$). When H is low, however, P_C is higher for a high divergence level than for a low divergence level.

Since the population size remained the same throughout the evolutionary process in this case, the expected evolutionary rate was constant for all different lineages. In such a case UPGMA is known to be efficient for obtaining a correct tree (NEI 1987, 1991). Table 1 shows that P_C's for UPGMA are slightly higher than those for NJ except for the case of $a = 0.1$; $H = 0.16$ and for F_{ST} for the case of $a = 0.1$; $H = 0.5$. UPGMA is based on the assumption of the molecular clock (NEI 1975, p. 200). Therefore, a distance measure that is linearly related to evolutionary time is expected to perform well in UPGMA. Indeed, the linear distance D_S shows a slightly higher P_C value in UPGMA than in NJ for all cases, but there are other distance measures that give a higher P_C than D_S when UPGMA is used. This has occurred because D_S has a larger variance than some other distance measures.

There are considerable differences in P_C among different distance measures. The P_C values for D_A, D_C, and X^2 are generally higher than those for the others. Among these three distance measures, the P_C values for X^2 are slightly lower than those for the first two. Although D_A and D_C show similar P_C values, D_A performs slightly better than D_C for the case of NJ with $a = 0.1$, whereas D_C's performance is slightly better than D_A when $a = 0.004$. Among the remaining distance measures, F_{ST} tends to show the highest P_C when NJ is used but not when UPGMA is used. Previously, we mentioned that the expected value of D_S increases linearly with time, but when NJ is used, it has a low P_C value.

TABLE 2

Percentage of replications in which the correct topology was obtained for the infinite-allele model in the case of the bottleneck

	NJ							Unrooted UPGMA						
No. loci	D_R	D_m	D_S	F_{ST}	X^2	D_C	D_A	D_R	D_m	D_S	F_{ST}	X^2	D_C	D_A
						$a = 0.1$; $H = 0.5$								
10	5	5	3	6	14	15	15	4	4	6	1	15	15	13
30	25	25	16	32	48	46	50	11	6	27	0	42	46	42
50	39	44	31	52	65	66	69	13	6	39	0	57	67	61
100	68	75	57	76	87	87	91	11	3	63	0	78	84	78
						$a = 0.1$; $H = 0.16$								
10	5	6	3	5	6	6	8	3	3	3	1	4	4	5
30	24	26	17	25	31	28	33	20	19	22	2	26	23	28
50	45	50	36	40	53	49	56	35	34	40	2	44	40	45
100	71	77	63	72	80	75	81	64	58	67	0	73	69	77
						$a = 0.004$; $H = 0.5$								
10	2	2	1	2	15	26	22	2	2	3	2	13	16	13
30	16	10	12	10	41	56	52	5	6	8	5	22	27	23
50	32	23	21	26	59	66	68	4	8	9	5	23	34	26
100	59	51	48	51	85	86	91	4	3	5	1	15	30	22
						$a = 0.004$; $H = 0.16$								
10	2	1	2	1	4	2	3	2	2	2	1	1	3	3
30	4	3	3	4	7	7	8	7	4	3	1	9	12	9
50	10	5	4	6	14	18	16	10	5	5	2	9	24	14
100	20	15	15	20	30	35	37	12	7	8	4	19	39	22

$N = 250$ and $v = 0.001$ for $H = 0.5$. $N = 250$ and $v = 0.0002$ for $H = 0.16$. The size of populations 2, 5, and 8 are reduced to a half.

When H is low and the divergence level is high, P_C is slightly higher for F_{ST} than for D_C.

Means and sampling errors of distance measures: The efficiency of constructing phylogenetic trees by means of genetic distances depends on the linear relationship with time and the sampling error of the distance measures used (NEI *et al.* 1983; GOLDSTEIN and POLLOCK 1994; TAJIMA and TAKEZAKI 1994). We have therefore examined these properties. Figure 2A shows the relationships of the mean distance and evolutionary time ($2vt$) for various distance measures for the case of $H = 0.5$. The results for $H = 0.16$ are not shown here, because the relationships are essentially the same as those for $H = 0.5$. The mean D_S increases linearly with time, as expected. However, the average values of all other distance measures reach a certain level of plateau, though when the divergence level is low, they increase approximately linearly with time. The average D_m increases fairly linearly with time even for a higher divergence level. By contrast, the rate of increase with time quickly decreases in F_{ST} and D_C compared with other distances. Generally, D_S, D_m, X^2, and D_A are better than F_{ST}, D_R, and D_C with respect to the linear relationship with time.

To measure the extent of sampling error, we computed the coefficient of variation (CV) for each distance measure. Figure 2B shows the relationships between the CV and the evolutionary time ($2vt$) for $H = 0.5$ for 30 loci. The relationships for $H = 0.16$ were similar to those for $H = 0.5$, though the CV's were generally higher in this case. The CV's for X^2 are not shown in Figure 2B, because they were virtually identical with those for D_A. As expected, the CV's are large for low divergence levels in all distance measures but decrease as the divergence level increases. For a very high divergence level, however, the CV's for D_S and F_{ST} tend to increase.

Figure 2B shows that D_S has the largest CV's among all distance measures examined. Apparently, because of this large CV, the P_C's for D_S are low even though the expected value of D_S has a perfect linear relationship with time. The CV's for D_m were the second largest following those for D_S, whereas D_C had the smallest CV. F_{ST} has a relatively high CV for $H = 0.5$, but in the case of $H = 0.16$, the CV for F_{ST} rapidly decreases as the divergence level increases. These low CV's seem to be responsible for a relatively high P_C for F_{ST} for the case of $a = 0.1$ and $H = 0.16$.

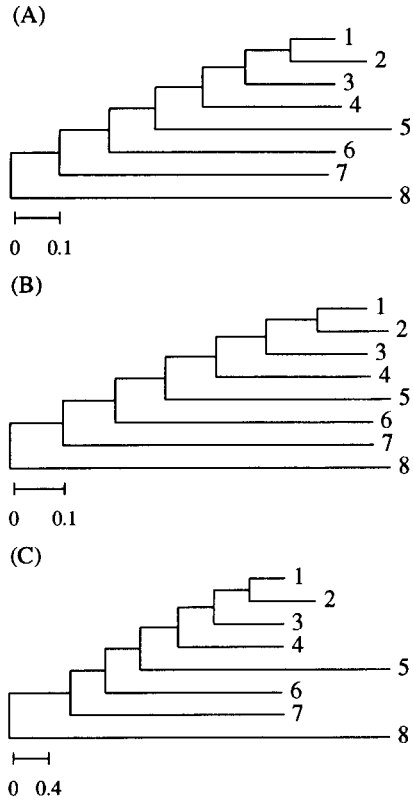

FIGURE 3.—NJ trees constructed by using average distances affected by the bottleneck effect. (A) D_S for $H = 0.5$; $a = 0.1$ with the IAM. (B) D_S for $H = 0.16$; $a = 0.1$ with the IAM. (C) $(\delta\mu)^2$ for $H = 0.8$; $a = 0.4$ with the SMM. The branch lengths for populations 2, 5, and 8 are elongated because of the bottleneck effect. When $a = 0.1$, the extent of elongation is higher for $H = 0.5$ than for $H = 0.16$. The average values of D_S and $(\delta\mu)^2$ were computed for 100 loci.

Bottleneck effect: Table 2 shows the P_C values for the case where populations undergo a bottleneck. In this case the P_C's for NJ and UPGMA both become smaller than those for the case of a constant population size (Table 1). Further, the P_C's for UPGMA are generally lower than those for NJ in this case. When a bottleneck occurs, the distance values are known to increase rapidly (CHAKRABORTY and NEI 1977). Figure 3, A and B show NJ trees constructed by using the average D_S for $H = 0.5$ and 0.16 when a divergence level is high ($a = 0.1$). In these trees, the branch lengths for the populations that went through a bottleneck are elongated. Thus, the evolutionary rate varies considerably with population, even if the mutation rate remains the same. In this case UPGMA is not efficient in obtaining the correct topology, because the assumption of rate constancy is violated. Thus, P_C values of D_R, D_m, and F_{ST} are nearly zero irrespective of the number of loci used.

The extent of decrease of P_C due to the bottleneck effect clearly depends on the bottleneck effect on distance measures. Thus, when the divergence level is high, the bottleneck effect on distance values is larger for $H = 0.5$ than for $H = 0.16$ (Figure 3, A and B), and the P_C values decline more drastically in the former case than in the latter. By contrast, when the divergence level was low, distance values increased considerably owing to the bottleneck effect for both heterozygosity levels (data not shown), and P_C's declined substantially.

Although every distance measure is affected by the bottleneck effect, D_A, D_C, and X^2 generally show a higher P_C than the other measures, and P_C's for D_A and D_C are higher than those for X^2 except for the case of $a = 0.1$; $H = 0.16$. The difference in P_C among other distance measures shows a pattern similar to that for the constant population size case when NJ is used. When UPGMA is used, however, F_{ST} does not perform well, giving an incorrect tree in most cases, but the P_C's for D_S are relatively high.

Stepwise mutation model: *Constant population size:* The P_C values for the case of the SMM are presented in Table 3. To save space, the P_C's for D_m and X^2 are not presented because their efficiencies relative to the traditional distance measures were essentially the same as those for the case of the IAM. The P_C's for UPGMA are also not shown here for the same reason. As in the case of the IAM, UPGMA performed better than NJ for the case of constant population size, but not for the case where the bottleneck effect was considered.

Table 3 shows that in the case of $a = 0.4$, P_C's are very low for any distance measure unless 300 or more loci are studied (except in the case of constant population size with $H = 0.8$). They are lower than those for the case of the IAM with $H = 0.5$ when the same number of loci are used (Table 1). This indicates that when the divergence level is high, microsatellite loci may be no better than classical markers unless the average heterozygosity for the latter markers are very low.

In the case of $a = 0.04$, however, the P_C values for traditional distance measures, particularly for D_A and D_C, are reasonably high if 30 or more loci are used. These values are comparable to those for the case of the IAM with $H = 0.5$ but are higher than those for the case of the IAM with $H = 0.16$ (Table 1). Since the heterozygosity level of classical markers is usually low, microsatellite loci seem more useful than classical markers for a study of closely related populations. Curiously, however, D_{SW} and $(\delta\mu)^2$ show lower P_C values than traditional distance measures, except for $a = 0.4$ and $H = 0.5$. This indicates that D_{SW} and $(\delta\mu)^2$ are generally less useful than traditional distance measures for phylogenetic inference even for microsatellite loci, for which the distance measures were specifically designed. Particularly when $a = 0.04$, D_{SW} shows a very poor performance for obtaining the correct tree even when 300 or more loci are studied.

The relative P_C values for traditional distance measures are essentially the same as those for the IAM, and D_C and D_A show a higher P_C than the other distance measures irrespective of the values of a and H.

Means and sampling errors of distance measures: Figure

TABLE 3

Percentage of replications in which the correct topology was obtained for the stepwise mutation model with the NJ method

	$a = 0.4$							$a = 0.04$						
No. loci	D_R	D_S	F_{ST}	D_C	D_A	D_{SW}	$(\delta\mu)^2$	D_R	D_S	F_{ST}	D_C	D_A	D_{SW}	$(\delta\mu)^2$
						$H = 0.8$; constant population size								
10	4	4	3	5	5	0	1	16	11	12	30	25	0	1
30	17	19	21	26	25	5	4	52	44	46	69	67	1	7
100	53	51	56	67	66	14	23	91	89	89	98	97	2	26
300	84	83	86	94	94	55	58	100	100	100	100	100	7	62
						$H = 0.5$; constant population size								
10	0	1	0	1	3	1	1	3	3	3	6	5	1	1
30	1	4	2	8	7	7	8	20	14	17	27	26	1	4
100	18	18	23	26	29	37	29	57	53	60	69	70	7	17
300	43	47	51	57	67	76	72	93	90	94	97	96	30	58
						$H = 0.8$; bottleneck								
10	3	2	2	4	4	0	1	11	8	6	17	18	0	0
30	11	8	9	19	15	1	2	34	29	29	45	46	0	3
100	28	35	39	41	51	3	17	66	74	75	80	85	3	17
300	47	71	74	61	78	0	55	92	98	97	97	100	2	44
						$H = 0.5$; bottleneck								
10	1	0	1	1	1	3	1	3	2	3	4	5	0	1
30	3	1	3	4	6	3	5	11	10	16	15	19	1	2
100	12	14	18	22	24	17	28	38	39	45	47	54	0	12
300	37	41	43	51	61	32	78	74	77	82	74	84	5	43

$a = 0.4$. $N = 150$, and $v = 0.02$ for $H = 0.8$. $N = 94$, and $v = 0.004$ for $H = 0.5$. $a = 0.04$. $N = 300$ and $v = 0.01$ for $H = 0.8$. $N = 188$ and $v = 0.002$ for $H = 0.5$. In the case of bottleneck, N and v were the same as the above except for $N = 188$ and $v = 0.002$ for $a = 0.4$ and $H = 0.5$. N was reduced to a half after separation from the ancestral populations in populations 2, 5, 8.

4A shows the relationships between the mean distance and evolutionary time ($2vt$) for various distance measures for the case of $H = 0.8$. The results for the case of $H = 0.5$ are not shown, because they are similar to those for the case of $H = 0.8$. As expected, $(\delta\mu)^2$ has a linear relationship with time. D_{SW} also increases almost linearly with time. All the traditional distance measures are nonlinearly related with time, though the relationships are initially linear. The initial linear relationship is slightly better for D_S and D_A than D_C.

Figure 4B shows the relationships between the CV's and evolutionary time ($2vt$) for $H = 0.8$. The CV's for F_{ST} and D_R are not shown, because they are close to those for D_A and D_C, respectively. The CV's for $(\delta\mu)^2$ and D_{SW} are clearly much greater than those of the traditional distance measures. These high CV's explain the poor performance of $(\delta\mu)^2$ and D_{SW} in phylogenetic reconstruction, though the mean values are linearly related with time (see Figure 4A).

ZHIVOLOVSKY and FELDMAN (1995) analytically showed that CV for $(\delta\mu)^2$ is approximately constant if a sufficiently high divergence level is considered. In the range of divergence shown in Figure 4B ($2vt \leq 6$), however, it is still decreasing for $H = 0.8$. For $H = 0.5$ CV for $(\delta\mu)^2$ reached a plateau more quickly (data not shown). In contrast to the case of $(\delta\mu)^2$, D_{SW} has a very high CV, but the CV drops quickly for high divergence. Therefore, D_{SW} has a larger CV than $(\delta\mu)^2$ for a low divergence level but a smaller CV for a high divergence level.

The CV's for traditional distance measures are nearly constant over time except in the initial stage. Particularly, D_S has a large CV when $2vt$ is small, but the CV quickly declines to reach an apparent constant value. Here, too, D_C and D_A have a smaller CV than D_S. The relationships between CV's and $2vt$ for $H = 0.5$ were more or less the same as those for $H = 0.8$.

Bottleneck effect: The effect of population size change (bottleneck effect) on P_C's for the NJ trees are shown in the lower half of Table 3. In the presence of the bottleneck effect, P_C's are generally smaller than those for the constant population size case (upper half of Table 3), as in the case of the IAM. The distance values increase rapidly in the presence of the bottleneck effect (Figure 3C), and the extent of decrease of P_C's is positively correlated to the extent of distance increase.

The relative values of P_C's for different distance mea-

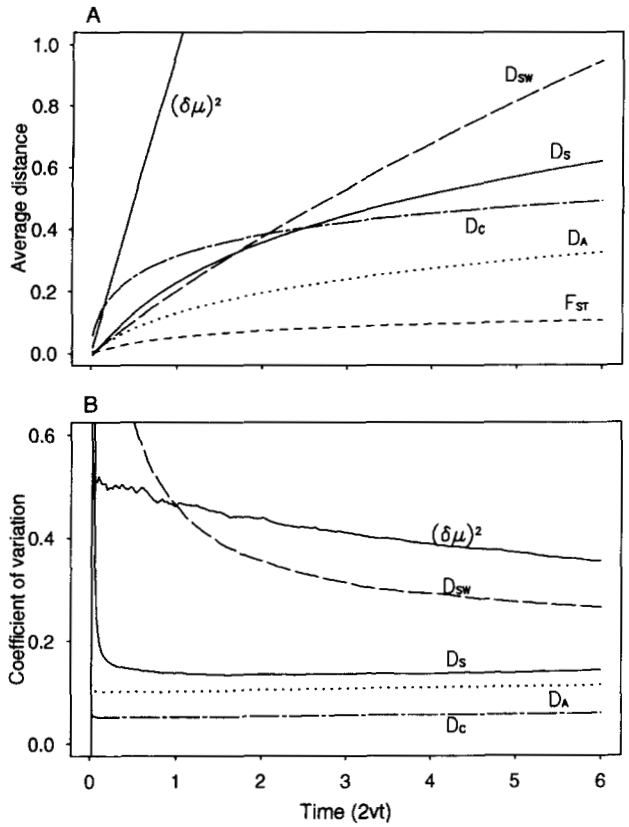

FIGURE 4.—Relationships of the mean distance and the coefficient of variation with evolutionary time ($2vt$) for the case of the SMM with $H = 0.8$. (A) Mean distance. (B) Coefficient of variation (CV). The means and CV's for each distance measure were computed for 100 loci and 30 loci, respectively, with 10,000 replications.

sures are generally the same as those for the constant population size. The P_C values for the traditional distance measures are again higher than those of D_{SW} and $(\delta\mu)^2$ for $a = 0.04$; $H = 0.8$ and 0.5 and $a = 0.4$; $H = 0.8$. Since the efficiencies of D_A and D_C are higher than other traditional distance measures, D_A and D_C performed best in these cases. By contrast, in the case of $a = 0.4$ and $H = 0.5$, $(\delta\mu)^2$ tends to show a higher P_C value than all other distance measures. In the case of constant population size, the P_C's for D_{SW} and $(\delta\mu)^2$ are similar to each other when $a = 0.4$, but in the presence of the bottleneck effect, the P_C's for D_{SW} drastically decrease. Although the CV for D_{SW} is smaller than that for $(\delta\mu)^2$ when the divergence level is high (Figure 4B), the P_C for D_{SW} is severely affected by the bottleneck effect. In practice, population size would change frequently in the evolutionary process, so $(\delta\mu)^2$ seems superior to D_{SW} in actual data analysis.

Sample size: So far we assumed that the P_C values are computed by sampling all individuals in the populations. In practice, however, the number of individuals sampled are usually limited. Therefore, we examined the effect of sample size on P_C values. In Table 4 we present only the P_C values for D_A, since other distance measures showed similar results. The number of individuals sampled varied from 10 to 50. The P_C values for the entire populations sampled are also shown in the table.

When the divergence level is high, the sample size (s) does not make much difference for the P_C values in both the IAM and SMM as long as $s \geq 20$. In the case of the IAM, this is true whether the heterozygosity level is high ($H = 0.5$) or low ($H = 0.16$). Similarly, in the case of the SMM, the sample size effect is very small when the heterozygosity level is low ($H = 0.5$). When the heterozygosity level is high ($H = 0.8$), however, P_C is lower for $s = 10 \sim 20$ than for $s \geq 30$. This is in agreement with SHRIVER *et al.*'s (1993) result. By contrast, when the divergence level is low, P_C increases with increasing sample size up to $s = 50$ if the heterozygosity is high ($H = 0.5$ for IAM and $H = 0.8$ for SMM). When average heterozygosity is low, however, the sample size effect is not substantial. Therefore, in general, it is more important to examine a large number of loci rather than a large number of genes per locus to have a higher P_C value. However, if the average heterozygosity is high and divergence level is low, a large number of individuals ($s \simeq 50$) should be also examined.

DISCUSSION

Our computer simulation has shown that D_A and D_C are most efficient in obtaining the correct tree topology in many different conditions examined. This is so despite the fact that D_A and D_C are based on geometric distances of populations represented on a multidimensional hypersphere and are independent of the mutational models (NEI 1987). The expected values of D_S and $(\delta\mu)^2$ increase linearly with time under the IAM and the SMM, respectively, as mentioned earlier. However, the efficiencies of D_S and $(\delta\mu)^2$ for obtaining the correct tree are generally low. This has happened because D_S and $(\delta\mu)^2$ have a large sampling error. Therefore, the extent of sampling error is an important factor for determining the efficiency of a distance measure in phylogenetic reconstruction. However, the coefficient of variation is not the sole determinant of the efficiency, and the linear relationship with time also plays some role, as is clear from the comparison of CV's between D_C and D_A. This conclusion is similar to NEI *et al.*'s (1983) and suggests that different distance measures should be used for reconstructing a topology and estimating branch lengths. In general, D_A and D_C are superior to other distance measures in topology construction and D_S and $(\delta\mu)^2$ are better than others in branch-length estimation, depending on the model used. Therefore, it seems to be preferable to use D_A and D_C for constructing a topology and then use D_S for estimating branch lengths for electrophoretic or immunological data and $(\delta\mu)^2$ for microsatellite loci (NEI 1995). A statistical method for conducting this type of statistical

TABLE 4

Percentage of replications in which the correct topology was obtained by the NJ method for different sample sizes

	No. loci															
	Infinite allele model								Stepwise mutation model							
	a = 0.1				a = 0.004				a = 0.4				a = 0.04			
Sample size (s)	10	30	50	100	10	30	50	100	10	30	50	100	10	30	50	100
	H = 0.5								H = 0.8							
10	13	59	75	94	1	5	12	46	2	8	24	45	1	4	9	32
20	11	60	80	94	5	24	50	81	4	16	30	55	2	16	34	74
30	15	61	81	95	9	39	65	94	4	17	38	58	7	26	46	82
50	15	63	80	94	13	54	76	94	4	19	41	64	9	35	64	90
All	17	62	80	94	25	69	88	98	4	23	43	66	23	57	73	97
	H = 0.16								H = 0.5							
10	8	33	57	80	1	5	13	36	1	7	12	29	2	7	32	58
20	6	34	57	80	2	8	17	48	2	6	13	31	2	16	41	67
30	7	35	58	79	2	9	24	51	2	7	16	29	3	20	41	68
50	6	36	58	80	2	12	24	54	3	7	16	31	3	19	46	72
All	7	36	59	80	3	15	28	54	3	7	16	29	3	19	47	71

s is the number of individuals sampled. D_A distance was used. Infinite-allele model. In the case of a = 0.1 N = 250 for H = 0.5 and for H = 0.16. In the case of α = 0.004, N = 250 for H = 0.5, and N = 100 for H = 0.16. Stepwise mutation model. In the case of a = 0.4, N = 150 for H = 0.8, and N = 94 for H = 0.5. In the case of a = 0.04, N = 300 for H = 0.8 and N = 188 for H = 0.5.

analysis has already been developed (TAKEZAKI *et al.* 1995). It should also be noted that the linear relationship with time is better for D_A than for D_C. Therefore, a tree constructed with D_A may be sufficient for many purposes unless the evolutionary time considered is very long.

In the present paper we are primarily concerned with phylogenetic reconstruction of intraspecific populations for microsatellite loci. However, when a phylogenetic relationship of different species is to be studied, $(\delta\mu)^2$ can be quite high. For example, the $(\delta\mu)^2$ between humans and chimpanzees for 25 microsatellite loci is ~25 (data from BOWCOCK *et al.* 1994), whereas the $(\delta\mu)^2$ between domestic sheep and Rocky Mountain bighorn sheep for eight microsatellite loci (data from FORBES *et al.* 1995) is ~39. In the latter case there is one locus that shows an extremely high degree of variation, and if we exclude this locus, $(\delta\mu)^2$ becomes ~14.

To obtain some idea about the utility of microsatellite loci for a study of genetic differentiation of different species, we conducted an additional computer simulation with a = 1 and 2, for which the largest expected number of gene substitutions ($2vt$) were 14 and 28, respectively, without the bottleneck effect. Table 5 shows the results of this simulation. The P_C values for traditional distance measures are now considerably smaller than those for the cases of a = 0.04 and 0.4. By contrast, $(\delta\mu)^2$ and D_{SW} tend to show slightly higher values than those in Table 3. Indeed, when H = 0.5, they show higher P_C values than traditional distance measures. However, the P_C's for them are still quite low unless 300 or more loci are studied.

In the present study we assumed that microsatellite loci are subject to the mutational pattern of the SMM. In practice, this assumption does not necessarily hold, and there is evidence that mutation sometimes produces nucleotide repeat patterns that require two or more step changes (KWIATKOWSKI *et al.* 1992; WEBER and WONG 1993). For this reason, VALDES *et al.* (1993), DI RIENZO *et al.* (1994), and GARZA *et al.* (1995) proposed mathematical models that allow changes of more than one step with some probability. Our computer simulation has shown that these models give essentially the same relative P_C values as those for the SMM used here (data not shown) and that details of mathematical models of the evolutionary change of microsatellite loci are unimportant for phylogenetic reconstruction.

However, the real problem with microsatellite loci is that the mutational pattern is often irregular and that there seems to be an upper limit of the number of repeats (FORBES *et al.* 1995; GARZA *et al.* 1995; GOLDSTEIN *et al.* 1995b). Furthermore, some microsatellite loci are highly polymorphic in some populations or species but monomorphic in others. For example, four out of the 25 polymorphic loci in humans are apparently monomorphic in chimpanzees (data from BOWCOCK *et al.* 1994). Even within humans some highly polymorphic loci in Europeans are monomorphic in Amerindians. This raises a question about the long-term stability of microsatellite loci. At the present time, we

TABLE 5

Percentage of replications in which the correct topology was obtained for highly divergent populations in the case of the stepwise mutation model

	$a = 2$							$a = 1$						
No. loci	D_R	D_S	F_{ST}	D_C	D_A	D_{SW}	$(\delta\mu)^2$	D_R	D_S	F_{ST}	D_C	D_A	D_{SW}	$(\delta\mu)^2$
						$H = 0.8$								
10	1	1	0	0	0	3	1	2	1	2	4	4	2	5
30	4	4	6	8	9	8	5	8	7	9	12	11	7	4
100	16	17	22	38	40	38	26	30	31	32	47	49	30	22
300	55	59	63	76	78	81	78	64	66	68	82	84	77	64
						$H = 0.5$								
10	2	0	0	0	0	1	3	1	1	0	0	0	3	2
30	0	1	1	2	2	9	8	1	1	1	1	1	4	6
100	5	4	4	8	10	40	38	9	6	8	13	18	41	42
300	28	25	29	40	47	88	87	32	30	35	50	56	81	83

The NJ method was used. $N = 75$, and $v = 0.04$ for $H = 0.8$. $N = 47$, and $v = 0.008$ for $H = 0.5$. The size of all populations remained constant.

do not know the reason for the difference in the extent of polymorphism between different populations. Therefore, some caution is necessary in the extrapolation of our results to the study of distantly related populations.

Another disturbing factor in the use of microsatellite loci is that the mutation rate seems to vary considerably from locus to locus and this will increase the variance of distance values. Therefore, the coefficients of variation given in Figure 4B should be considered to be minimum, and the actual P_C values are expected to be smaller than those given in Tables 3–5. The bottleneck effect also would contribute to reducing the P_C values, as we have seen before.

Despite these problems with microsatellite loci, they are clearly useful for studying the genetic relationships of closely related populations. With classical genetic markers, it is often difficult to study this problem, because there are not many polymorphic loci that are used for this purpose. In the case of microsatellite loci, it is relatively easy to examine many polymorphic loci. Therefore, at least for studying closely related populations microsatellite loci are expected to be very useful.

This study was supported by research grants from the National Institutes of Health and the National Science Foundation to M.N.

LITERATURE CITED

BHATTACHARYYA, A., 1946 On a measure of divergence between two multinomial populations. Sankhya **7:** 401–407.

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:** 455–457.

CAVALLI-SFORZA, L. L., 1969 Human diversity. Proc. 12th Intl. Cong. Genet., Tokyo **3:** 405–416.

CAVALLI-SFORZA, L. L., and A. W. F. EDWARDS, 1967 Phylogenetic analysis: models and estimation procedures. Amer. J. Hum. Genet. **19:** 233–257.

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 $(dC\text{-}dA)_n \cdot (dG\text{-}dT)_n$ polymorphisms in world populations. Amer. J. Hum. Genet. **56:** 461–474.

DI RIENZO, A., A. C. PETERSON, J. C. GARZA, A. M. VALDES, M. SLATKIN *et al.*, 1994 Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91:** 3166–3170.

FORBES, S. H., J. T. HOGG, F. C. BUCHANAN, A. M. CRAWFORD and F. W. ALLENDORF, 1995 Microsatellite evolution in congeneric mammals: domestic and bighorn sheep. Mol. Biol. Evol. **12:** 1106–1113.

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 D. D. POLLOCK, 1994 Least squares estimation of molecular distance-noise abatement in phylogenetic reconstruction. Theor. Pop. Biol. **45:** 219–226.

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

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

GOTTELLI, D., C. SILLERO-ZUBIRI, G. D. APPLEBAUM, M. S. ROY, D. J. GIRMAN *et al.*, 1994 Molecular genetics of the most endangered canid: the Ethiopian wolf *Canis simensis*. Mol. Ecol. **3:** 301–312.

JORDE, L. B., M. J. BAMSHAD, W. S. WATKINS, R. ZENGER, A. E. FRALEY *et al.*, 1995 Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. Am. J. Hum. Genet. **57:** 523–538.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KWIATKOWSKI, D. J., E. P. HENSKE, K. WEIMER, L. OZELIUS, J. F. GUSELLA *et al.*, 1992 Construction of a GT polymorphism map in human 9q. Genomics **12:** 229–240.

LATTER, B. D. H., 1972 Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. Genetics **70:** 475–490.

LI, W.-H., and M. NEI, 1975 Drift variances of heterozygosity and genetic distance in transient states. Genet. Res. Camb. **25:** 220–248.

NEI, M., 1972 Genetic distance between populations. Amer. Nat. **106:** 283–291.

NEI, M., 1973 The theory and estimation of genetic distance, pp.

45–54 in *Genetic Structure of Populations*, edited by N. E. MORTON. University Press of Hawaii, Honolulu.

NEI, M., 1975 *Molecular Population Genetics and Evolution*. North-Holland, Amsterdam.

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

NEI, M., 1991 Relative efficiencies of different tree-making methods for molecular data, pp. 90–128 in *Phylogenetic Analysis of DNA Sequence*, edited by M. M. MIYAMOTO and J. CRACRAFT. Oxford University Press, New York.

NEI, M., 1995 Genetic support for the out-of-Africa theory of human evolution. Proc. Natl. Acad. Sci. USA **92:** 6653–7136.

NEI, M., and N. TAKEZAKI, 1994 Estimation of genetic distances and phylogenetic trees from DNA analysis. Proc. 5th World Cong. Genet. Appl. Livestock Prod. **21:** 405–412.

NEI, M., F. TAJIMA and Y. TATENO, 1983 Accuracy of estimated phylogenetic trees from molecular data. J. Mol. Evol. **19:** 153–170.

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

PAETKAU, D., W. CALVERT, I. STIRLING and C. STROBECK, 1995 Microsatellite analysis of population structure in Canadian polar bears. Mol. Ecol. **4:** 347–354.

PREVOSTI, A., J. OCANA and G. ALONZO, 1975 Distances between populations for *Drosophila subobscura* based on chromosome arrangement frequencies. Theor. Appl. Genet. **45:** 231–241.

REYNOLDS, J., B. S. WEIR and C. C. COCKERHAM, 1983 Estimation of the coancestry coefficient: basis for a short-term genetic distance. Genetics **105:** 767–779.

ROGERS, J. S., 1972 Measures of genetic similarity and genetic distance, pp. 145–153 in *Studies in Genetics* VII. University of Texas Publication 7213, Austin, TX.

ROY, M. S., E. GEFFEN, D. SMITH, E. A. OSTRANDER and R. K. WAYNE, 1994 Patterns of differentiation and hybridization in North American wolflike canids, revealed by analysis of microsatellite loci. Mol. Biol. Evol. **11:** 553–570.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

SANGHVI, L. D., 1953 Comparison of genetical and morphological methods for a study of biological differences. Amer. J. Phys. Anthropol. **11:** 385–404.

SHRIVER, M. D., L. JIN, R. CHAKRABORTY and E. BOERWINKLE, 1993 VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134:** 983–993.

SHRIVER, M., L. JIN, E. BOERWINKLE, R. DEKA, R. E. FERRELL *et al.*, 1995 A novel measure of genetic distance for highly polymorphic tandem repeat loci. Mol. Biol. Evol. **12:** 914–920.

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

SNEATH, P. H. A., and R. R. SOKAL, 1973 *Numerical Taxonomy*. W. H. Freeman, San Francisco.

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

TAKEZAKI, N., A. RZHETSKEY and M. NEI, 1995 Phylogenetic test of the molecular clock and linearized trees. Mol. Biol. Evol. **12:** 823–833.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics **133:** 737–749.

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: N. TAKAHATA