*Amer J Hum Genet* 25:247–261, 1973

# The Estimation of Genetic Divergence between Populations Based on Gene Frequency Data

B. D. H. LATTER[1]

Population genetics is concerned with the origin and maintenance of genetic differences *within* populations and with the process of genetic differentiation *between* populations. It is now possible to identify an extensive array of genotypes in natural populations by the use of immunological and electrophoretic techniques [1]; a quantitative description can therefore be given of the extent of genetic divergence between any two populations due to selection, mutation, migration, and drift. There is, however, no general agreement as to the most useful of the many "distance" measures which have been proposed for this purpose [2–7].

There are basically four kinds of distance parameters currently being used for the analysis of differences in gene frequency between populations. (1) Sanghvi [8] proposed an index based on the $\chi^2$ statistic, designated $G_S^2$ by Balakrishnan and Sanghvi [2], which has obvious advantages for tests of the statistical significance of differences in gene frequency and which also can be given a simple interpretation in terms of $F$ statistics [9]. (2) An angular transformation of gene frequencies has been used by Cavalli-Sforza and Edwards [10] to derive a measure of distance, $2\theta/\pi$, for which the drift variance is expected to be independent of mean gene frequencies which lie in the interval .05–.95. A related statistic, $f_\theta$, which is more closely related to the coefficient of kinship, has been used by Cavalli-Sforza et al. [11]. (3) Estimation of the coefficient of kinship, $\phi$, directly from the variance between populations in the untransformed allelic frequencies is advocated by Morton [4]. An appropriate expression for averaging such estimates over alleles has been provided by Yasuda [12]. (4) A measure of genetic distance, $\gamma$, which is unrelated to the coefficient of kinship but which increases as a simple function of time under a regime of mutation, random genetic drift, and centripetal selection, has been proposed by Latter [6]. The parameter is proportional to the rate of gene substitution throughout the evolutionary history of a population and can be interpreted as a rate of *mutational* divergence [7]. A related measure has been used by Selander [5] and by Nei [13].

This paper compares the behavior of a representative set of distance parameters throughout the early stages of divergence of simulated contemporary populations. The model assumes no migration, a constant population size $N$, and continual

Received June 28, 1972, revised Novermber 2, 1972.

[1] Division of Animal Genetics, Commonwealth Scientific and Industrial Research Organization, P.O. Box 90, Epping, New South Wales, Australia, 2121.

mutation at a rate $\mu$ per generation to alleles which are novel to the populations concerned. Loci subject to one of two intensities of natural selection for an optimal level of gene activity have been studied in addition to strictly neutral loci [14]. The use of gene frequency data derived by computer simulation clearly shows up difficulties in the interpretation of some measures of distance and confirms the regular behavior of those based solely on the calculated mean levels of heterozygosity [7].

## FORMULAS

1. $G_S^2$: Balakrishnan and Sanghvi [2] have defined the genetic distance between two populations $P_1$ and $P_2$ as

$$G_S^2 = \sum_i \frac{(p_{i1} - p_{i2})^2}{\bar{p}_i}, \tag{1}$$

where the alleles $A_i$ occur in the two populations with frequencies $p_{i1}$, $p_{i2}$, respectively, and $i = 1, 2, \ldots, n$. Mean allele frequencies are denoted by $\bar{p}_i = \frac{1}{2}(p_{i1} + p_{i2})$.

If each of the two populations assayed is of size $N$, $G_S^2$ as defined by equation (1) is equivalent to $\chi^2/N$ with $n-1$ df [9]. Tests of significance of differences between populations may therefore readily be made for individual loci, and for a group of loci collectively by summing or averaging the measure over loci.

2. $f_\theta$: Cavalli-Sforza [3] has proposed the following measure of distance,

$$f_\theta = 4(1 - \cos\theta)/(n - 1), \tag{2}$$

where

$$\cos\theta = \sum_i (p_{i1}\, p_{i2})^{1/2},$$

and $n$ denotes the number of alleles segregating at the locus concerned. Information from different loci can be combined by the use of mean values of $\cos\theta$ and $n$ in equation (2).

Use of the symbol $f_\theta$ for this parameter emphasizes its approximation to the inbreeding coefficient for small values of $\theta$ [3]. However, the measure has an upper limit of 4 for two-allele polymorphisms and cannot, therefore, be expressed in terms of "gene substitutions" as could the value of $2\theta/\pi$ used in earlier studies by Cavalli-Sforza and Edwards [10].

3. $\phi_Y$: Yasuda [12] has defined a weighted measure of the dispersion in gene frequencies for two populations as follows:

$$\phi_Y = \sum_i \frac{\frac{1}{4}(p_{i1} - p_{i2})^2}{(1 - \bar{p}_i)}. \tag{3}$$

The numerator is equal to

$$\sigma_{p_i}^2 = E(p_i^2) - [E(p_i)]^2, \tag{4}$$

and $\phi_Y$ for a single locus is therefore a mean of $n$ separate values of $\sigma_{p_i}^2/[\bar{p}_i\,(1 -$

$\bar{p}_i)]$, one for each allele, weighted according to the mean frequency of the allele concerned. The measure has an upper limit of unity and may be averaged over loci.

4. $\phi^*$: Latter [7] has suggested the following measure of population differentiation,

$$\phi^* = [\Sigma_i \tfrac{1}{2}(p_{i1} - p_{i2})^2]/(1 - \Sigma_i p_{i1} p_{i2}) \quad (5)$$

$$= 1 - H/H_B, \quad (6)$$

where

$$H = 1 - \tfrac{1}{2} \Sigma_i (p_{i1}^2 + p_{i2}^2)$$

is the mean level of heterozygosity within populations, and

$$H_B = 1 - \Sigma_i (p_{i1} p_{i2})$$

is that predicted in the $F_1$ population, $P_1 \times P_2$. Information from several loci may be combined by the use of mean values of $H$ and $H_B$ in equation (6).

The parameter $\phi^*$ is identical with the coefficient of kinship for a model of population divergence involving genetic drift alone and has a range from zero to unity.

5. $\gamma$: Latter [6] has defined a measure of genetic distance which corresponds to the rate of gene substitution expected in isolated finite populations due to mutation pressure, with or without selection for an optimal level of gene activity. The measure is

$$\gamma = [\Sigma_i (p_{i1} - p_{i2})^2]/[\Sigma_i (p_{i1}^2 + p_{i2}^2)] \quad (7)$$

$$= (H_B - H)/(1 - H), \quad (8)$$

where $H$ and $H_B$ represent mean within-population heterozygosity and between-population heterozygosity, respectively, as before. The value of the parameter represents that fraction of a gene substitution which has occurred at a locus to differentiate the two populations, provided the mean level of heterozygosity over all such loci, $E(H)$, has remained essentially unchanged throughout.

Unlike the measures of distance previously discussed, the estimation of $\gamma$ requires data from a randomly chosen set of loci, both monomorphic and polymorphic. The combined measure over all loci involves the use of mean values of $H$ and $H_B$ in equation (8).

## THE GENETIC MODEL AND SIMULATION PROCEDURES

### *The Model*

We are concerned in this paper with two or more completely isolated populations, produced initially by the splitting of a single panmictic parental population. The effective population size is supposed to be equal to $N$, being the same for all isolates including the parental population. A rapid increase in numbers is therefore envisaged as part of the process of population subdivision: in the computer populations this increase in population size prior to fission is taken to be instantaneous.

The regime in each population involves the following sequence of operations in each generation: (1) natural selection favoring an intermediate optimal level of gene or enzyme activity; (2) random drift in gene frequencies due to finite population size; (3) mutation to new alleles not present in any of the simulated populations; and (4) random mating of surviving individuals. Natural selection is, therefore, supposed to be due to differences in mortality prior to reproductive age and does not involve inherited differences in fertility. Mutation is simulated at the commencement of reproduction, but includes mutational events occurring at all prior stages of development.

The effects of alleles at a given locus are assumed to be additive on a scale which measures differences of adaptive significance among the genotypes: for example, differences in specific catalytic activity, or in the rate of synthesis of the enzyme or protein concerned [1, 15]. A heterozygote therefore has a mean level of "activity" equal to the average of the levels of the two corresponding homozygotes [6, 16]. The allelic effect of a new mutant is $a_i + \delta a_i$, where $a_i$ is that of the parental allele and $\delta a_i$ is a unit random normal deviate.

Natural selection is simulated by assigning a relative selective value to the genotype $A_i A_j$ which is proportional to

$$w^*_{ij} = 1 - \tfrac{1}{2} C^* (d_{ij}^2), \tag{9}$$

where $d_{ij} = \bar{x} + a_i + a_j$ is the deviation of the mean activity of $A_i A_j$ from optimal, and $\bar{x}$ is the weighted mean activity of all genotypes in the population at that time. The coefficient $C^*$ is a measure of the intensity of selection against deviant genotypes [14]. The assumption of a normal distribution of mutant effects about that of the parent allele implies that in a population homozygous for an allele of optimal activity, roughly 5% of new mutants would lead to a reduction in fitness greater than $2C^*$ in heterozygotes and greater than $8C^*$ in homozygotes. In this study, values of $C^* = 0.000$, 0.005, and 0.010 have been used for direct comparison throughout. The behavior of distance measures based on random drift theory can then be determined for neutral loci and compared with that at loci subject to two different intensities of natural selection.

*Simulation Techniques*

The computer techniques have been described in full elsewhere [14]. Natural selection is simulated by an appropriate transformation of the vector of allelic frequencies each generation, to give

$$p'_i = p_i [1 - \tfrac{1}{2} C^* (\bar{x}^2 + 2\bar{x} a_i + a_i^2 + \tfrac{1}{2} \sigma_g^2)] / \bar{w}^*, \tag{10}$$

where $\bar{x}$ is the population mean on the scale of gene or enzyme activity, $\sigma_g^2$ is the corresponding genotypic variance, $\bar{w}^* = 1 - \tfrac{1}{2} C^* (\bar{x}^2 + \sigma_g^2)$, and the $a_i$ are coded each generation so that $\Sigma p_i a_i = 0$.

Drift in gene frequency due to genetic sampling is simulated each generation by drawing a random sample from a multinomial distribution (parameters $2N$; $p_i$, $i = 1, 2, \ldots n$). The number of mutants is determined each generation as a random

Poisson variate with mean $2N\mu$, and mutational events are allocated at random to the existing alleles, weighted by the current values of $p_i'$.

*Choice of Parameter Values and Initial Populations*

The parental populations were sampled at 5,000-generation intervals from the long-term computer populations depicted in figures 2–4 of Latter [6], which had reached equilibrium under a regime with a population size of $N = 500$, a mutation rate given by $N\mu = 0.05$, and intensities of selection for optimal activity corresponding to $NC^* = 0.0$, 2.5, and 5.0, respectively. A value of $N\mu = 0.05$ can be taken to apply to a range of population size–mutation rate combinations from $N = 5{,}000$, $\mu = 10^{-5}$ to $N = 500$, $\mu = 10^{-4}$, and leads to mean levels of heterozygosity at equilibrium of approximately $H = 0.16$, 0.15, and 0.13, respectively, for the chosen values of $NC^* = 0.0$, 2.5, and 5.0 [6]. These levels of heterozygosity are of the same order as those observed in electrophoretic surveys of randomly selected loci in populations of man, mice, and *Drosophila* [16–19].

A total of 40 parental populations was chosen for each selection intensity. Five replicates of each parental population were continued for 2,500 generations with the same value of $N$, $\mu$, and $C^*$ as before. Pairwise measures of genetic distance between the isolates were evaluated at regular intervals during this 2,500-generation period, and averages taken over the 10 possible comparisons of gene frequencies in the five isolates.

## RESULTS

It is necessary at the outset to describe in some detail the genetic variability expected over an evolutionary time span with this genetic model. The allelic variation observed in a population in equilibrium under a regime of drift, mutation, and centripetal selection may be (1) neutral, if either $C^* = 0.0$ or if the alleles segregating differ so little in activity that selection cannot effectively discriminate among them; (2) subject to directional selection due to the average selective advantage of one allele over another; or (3) subject to balancing selection because of the selective advantage of a heterozygote. The latter may be closer to the optimum in mean activity than either of the two corresponding homozygotes, one of which is above optimal and the other below optimal. An earlier study has shown this phenomenon to be an important feature of the model under discussion [6].

If an allelic difference is neutral, the gene frequencies concerned are expected to decrease slowly due to mutational pressure alone, though genetic drift will be a far more important process in small populations. In the long term, any allele will eventually be replaced by a mutant derivative, an event which is scored as one gene substitution in the evolutionary history of the population. With centripetal selection this process is retarded due to selection against mutant alleles with above- or below-optimal effect, and by selection for the heterozygote in balanced polymorphisms. However, all populations are subject to drift in the population mean, including the case of a population initially homozygous for an allele of optimal activity; thus, mutation combined with genetic sampling will eventually lead to the chance

fixation of suboptimal alleles. This may be followed either by selective replacement of the suboptimal mutant allele by one more closely approximating the optimum or by the establishment of a heterotic polymorphism involving an allele of compensatory effect in the heterozygote [6].

Whatever the merits might be of this particular model of selection for optimal activity, the simulated populations are of considerable general interest in view of the variety of genetic polymorphisms which arise in the populations sampled, namely, neutral polymorphisms, those due to selective replacement of one allele by another, and those maintained by balancing selection. For regimes with $N\mu = 0.05$ and $NC^* = 0.0$, 2.5, and 5.0, the overall probability of genetic polymorphism at a randomly chosen locus is approximately 0.43, 0.39, and 0.31, respectively; for the nonneutral loci, approximately half of the observed polymorphisms become established partly because of the selective advantage of the heterozygote [6].

*Increase in Distance with Time*

Figures 1–6 illustrate changes in the means of the five chosen distance measures over a 5.0 $N$-generation period following population subdivision. Figures 1–4 are based only on the allele frequencies observed at loci which are polymorphic at generation zero, that is, those for which the initial frequency of heterozygotes exceeds 0.10. It is assumed that, in practice, polymorphic loci will be identified by examination of a moderately large number of populations and that all such loci will be used to calculate $\phi_S$, $f_\theta$, $\phi_Y$, and $\phi^*$ for any pair of the populations, even when both happen to be monomorphic for the same allele.

The theoretical curve used for comparison in figures 1–4 is $\phi(t)$, the expected coefficient of kinship for neutral loci after $t$ generations of independent evolution, given by

$$\phi(t) = (1 + 4N\mu)^{-1}\left\{1 - \exp\left[-\frac{t}{2N}(1 + 4N\mu)\right]\right\}, \qquad (11)$$

with a final equilibrium value of $(1 + 4N\mu)^{-1}$ [20].

The statistic $\phi_S$ plotted in figure 1 is related to $G_S^2$ defined by equation (1) and is based on the following expectations. From equation (4) it can be seen that $G_S^2$ is by definition

$$G_S^2 = 4\sum_i \sigma_{p_i}^2/\bar{p}_i = 4\left[\sum_i \frac{\sigma_{p_i}^2}{\bar{p}_i(1-\bar{p}_i)} - \sum_i \frac{\sigma_{p_i}^2}{(1-\bar{p}_i)}\right]. \qquad (12)$$

It is usual [4, 12, 22] to take as the definition of the coefficient of kinship

$$\phi = \frac{\sigma_p^2}{\bar{p}(1-\bar{p})}, \qquad (13)$$

leading to an expectation of $4(n-1)\phi$ for the observed value of $G_S^2$ [9]. We have therefore defined $\phi_S$ to be

$$\phi_S = \tfrac{1}{4}G_S^2/(n-1) \qquad (14)$$

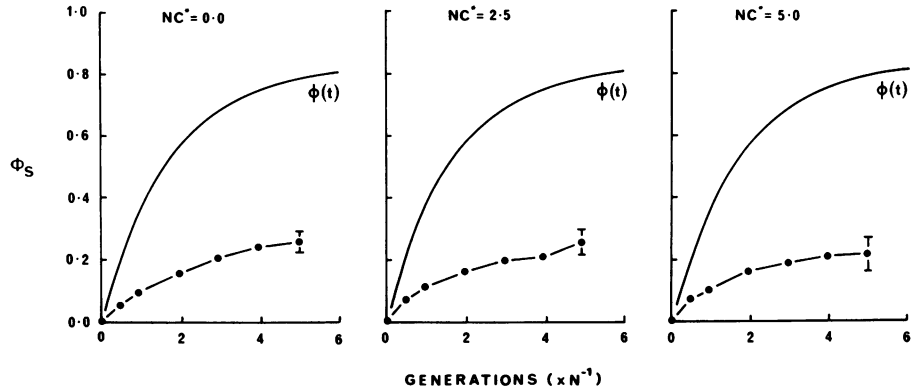


FIG. 1.—Observed values of $\phi_S = \frac{1}{4}\Sigma G_S^2/\Sigma(n-1)$, measuring the genetic distance between two completely isolated contemporary populations based on a random sample of polymorphic loci. The statistic $G_S^2$ is defined by equation (1), and $n$ denotes the number of alleles contributing to the value of $G_S^2$ with a frequency $\geqslant .01$ in one of the two populations concerned. The theoretical curve is the coefficient of kinship, $\phi(t)$, defined by equation (11). The parameter $C^*$ specifies the intensity of natural selection for an intermediate optimal level of gene or enzyme activity. The sampling errors shown are $\pm$ SE of the mean.

with an expected range from zero to unity. Note that equation (13) refers to the mean coefficient of kinship within populations and corresponds to Wright's statistic $F_{ST}$.

It is clear from figure 1 that the value of $\phi$ given by equation (13) is a gross underestimate, if based on gene frequencies observed in only two, or a small number of populations. The bias stems from the definition of $\sigma_p^2$ given by equation (4) and for small values of $\phi$ is equal to $(k-1)/k$, where $k$ is the number of populations contributing to estimates of the variance in gene frequency. The initial rate of change of $\phi_S$ is therefore equal to $t/4N$, whereas $\phi(t)$ given by equation (11) is equal to $t/2N$ for small values of $\phi$. The actual rate of change $\phi_S$ can be seen from figure 1 to fall off very rapidly from the expected value of $\phi_S = t/4N$ as $t/N$ increases, and to be little affected by natural selection for an intermediate optimum at the intensities simulated.

Figure 2 shows the corresponding changes in the mean value of $f_\theta$ defined by equation (2). All alleles, irrespective of their frequency, have been used in the calculation of the mean value of cos $\theta$, but it has been found empirically that the most satisfactory fit to the theoretical value of $\phi(t)$ is obtained if the mean number of alleles per locus, $n$, is taken to refer only to those with a frequency $\geqq 0.01$ in at least one of the two populations being compared. The agreement between observation and expectation can be seen from figure 2 to be excellent for $t/N \leqq 3$, and is not affected appreciably by natural selection in the two regimes with $NC^* = 2.5$ and 5.0. At values of $t/N > 4$, the observed values of $f_\theta$ appreciably exceed those of $\phi(t)$ because $f_\theta$ does not have an upper limit of unity and may take values as high as 4 for two-allele polymorphisms. The statistic clearly is not designed for populations as divergent as these [3].

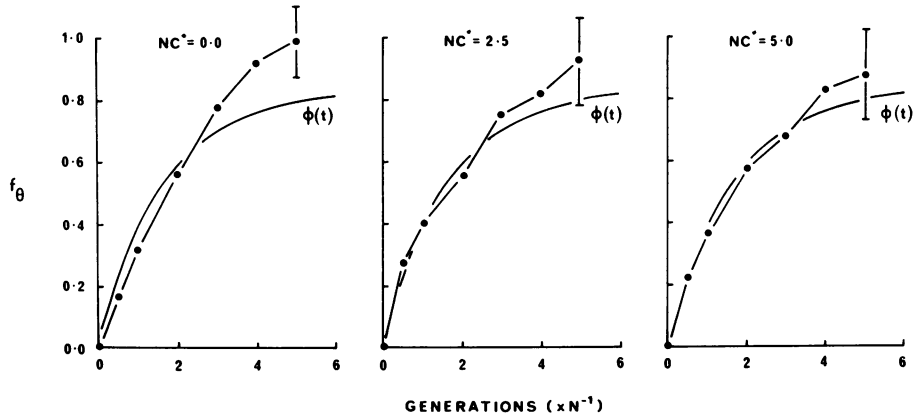


FIG. 2.—Values of Cavalli-Sforza's statistic, $f_\theta = 4\Sigma(1 - \cos\theta)/\Sigma(n - 1)$, as a function of the time elapsed since population fission, based on a random sample of polymorphic loci. Cos $\theta$ is defined as $\Sigma(p_{i1}p_{i2})^{\frac{1}{2}}$, and $n$ denotes the number of alleles contributing to the value of $f_\theta$, with a frequency $\geqslant .01$ in at least one of the two populations concerned.

The behavior of Yasuda's measure of distance, $\phi_Y$, is illustrated in figure 3. As with the observed values of $\phi_S$ in figure 1, the statistic $\phi_Y$ seriously underestimates the coefficient of kinship appropriate to the comparison of two populations because it is based on the definition of variance in gene frequency given by equation (4). The bias is the same for each of the two measures, leading to an initial rate of change of $\phi_Y = t/4N$. However, Yasuda's measure falls off much less rapidly than Sanghvi's, since it gives greatest weight to the common alleles at multiallelic loci. It may be noted here that the measure $\theta$ defined by Morton, Yee, and Harris [21] is related to Yasuda's $\phi_Y$ and is subject to the same bias. In terms of $\phi^*$, $\theta$ is given by

$$\theta = \tfrac{1}{2}\phi^*/(1 - \tfrac{1}{2}\phi^*). \qquad (15)$$

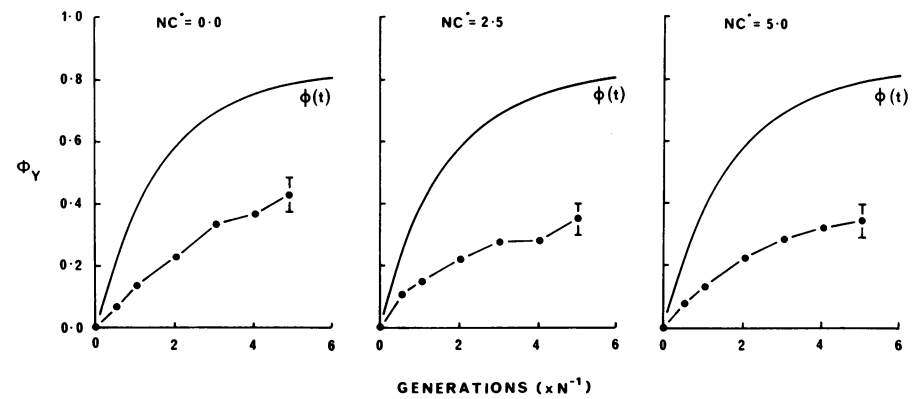


FIG. 3.—Values of Yasuda's statistic $\phi_Y$ obtained by averaging equation (3) over a random sample of polymorphic loci, by comparison with the expected coefficient of kinship given by equation (11).

The observed values of $\phi^* = 1 - H/H_B$ are plotted in figures 4 and 5. The means of figure 4 are based on polymorphic loci alone, as were the corresponding values of $\phi_S$, $f_\theta$ and $\phi_Y$ in figures 1–3. The statistic $\phi^*$ can be seen to underestimate the true coefficient of kinship $\phi(t)$ at intermediate values of $t/N$, though both have virtually the same initial rate of change $1/2N$ per generation and the same equilibrium value of $(1 + 4N\mu)^{-1}$ [7]. Table 1 compares the observed mean values of

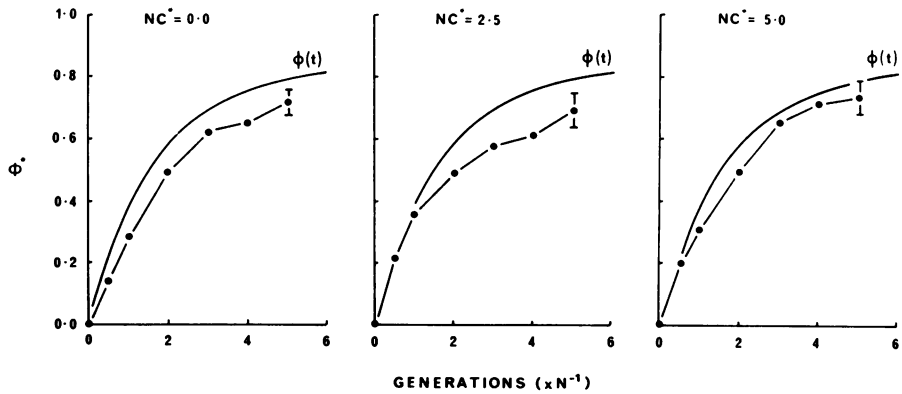


FIG. 4.—Mean values of the statistic $\phi^*$ based on a random sample of polymorphic loci obtained by substituting average values of

$$H = 1 - \tfrac{1}{2} \sum_i (p_{i1}^2 + p_{i2}^2)$$

and

$$H_B = 1 - \sum_i (p_{i1}\, p_{i2})$$

in equation (6). The theoretical curve shows the expected change in the coefficient of kinship given by equation (11) (c.f. table 1).

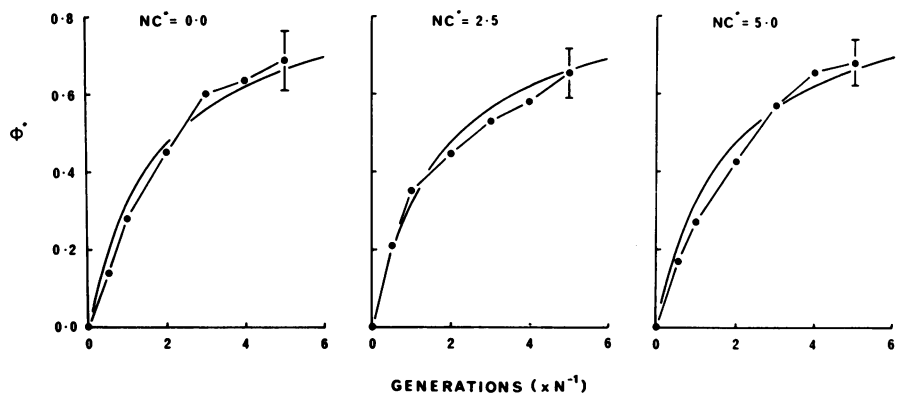


FIG. 5.—Mean values of the statistic $\phi^*$ based on a random sample of all loci, both polymorphic and monomorphic. As in fig. 4, the means are obtained by substituting average values of $H$ and $H_B$ in equation (6). The theoretical curve is the predicted value of $\phi^*$ in the absence of selection. given by equation (16).

TABLE 1

OBSERVED AND PREDICTED VALUES OF $\phi^*$ FOR A RANDOM SAMPLE OF POLYMORPHIC LOCI

| GENERATIONS ($t/N$) | $NC^* = 0.0$; $H_0 = .381$ | | $NC^* = 2.5$; $H_0 = .301$ | | $NC^* = 5.0$; $H_0 = .382$ | |
|---|---|---|---|---|---|---|
| | Obs. | Exp. | Obs. | Exp. | Obs. | Exp. |
| 0.5 | .14 ± .02 | .209 | .22 ± .03 | .206 | .19 ± .04 | .209 |
| 1.0 | .28 ± .04 | .354 | .37 ± .05 | .346 | .31 ± .04 | .354 |
| 2.0 | .49 ± .04 | .531 | .49 ± .05 | .516 | .49 ± .06 | .531 |
| 3.0 | .62 ± .04 | .627 | .58 ± .05 | .608 | .65 ± .07 | .627 |
| 4.0 | .65 ± .04 | .682 | .61 ± .04 | .663 | .71 ± .06 | .682 |
| 5.0 | .72 ± .04 | .716 | .69 ± .04 | .699 | .73 ± .05 | .716 |

NOTE.—The loci included are those for which the initial frequency of heterozygotes ($H_0$) exceeds .10. Predicted values are based on equations (13–15) of Latter [7].

$\phi^*$ with theoretical values predicted for an island model with drift and mutation, but without migration or selection [7]. The agreement with expectation is in general excellent, and there is no detectable effect of natural selection on the rate of change of this parameter (fig. 4).

The means plotted in figure 5 are based on a random sample of *all* loci, both polymorphic and monomorphic, for comparison with the theoretical expectation for populations with an initial mean level of heterozygosity of $H_0 = 4N\mu/(1 + 4N\mu)$, in equilibrium under a regime of drift and mutation alone [7]. The appropriate formula is

$$\phi^*(t) = \frac{1 - \exp(-2t\mu)}{(1 + 4N\mu) - \exp(-2t\mu)}, \tag{16}$$

which for small values of $t/N$ is equal to

$$\phi^*(t) \sim \frac{t/2N}{1 + t/2N} \tag{17}$$

This approximation is excellent for $t \leq N$, and figure 5 shows the simulated populations conforming closely to the theoretical predictions for all three regimes.

All of the measures so far discussed are related to the coefficient of kinship, $\phi$, and have been shown to be insensitive to natural selection favoring an intermediate optimum level of gene or enzyme activity (figs. 1–5). It has nevertheless been shown in a previous study that selection intensities corresponding to $NC^* = 2.5$ and 5.0 lead to reductions of roughly one-third and one-half, respectively, in the rate at which an isolated finite population changes genetically due to drift and mutation [6]. In the absence of natural selection, the expected rate of accumulation of mutational changes in a protein is expected to be $N\mu$ per $N$ generations. This rate of genetic divergence can be measured over short periods by the parameter $\gamma$ defined by equations (7) and (8). For $NC^* = 0.0$, 2.5, and 5.0, the estimated values of $\gamma$ in the earlier study were 0.051 ± .007, 0.033 ± .005, and 0.024 ± .005 gene substitutions per $N$ generations, averaged over the long-term evolutionary

history of individual populations of size $N$ and mutation rate given by $N\mu = 0.05$ [6].

The genetic distance between contemporary populations with a common origin is expected to increase at a rate approximately equal to

$$\gamma(t) = 1 - \exp(-2r^*t/N), \tag{18}$$

where $r^*$ is the expected number of mutational changes per $N$ generations in the proteins concerned. Figure 6 illustrates the values of $\gamma$ observed in the present

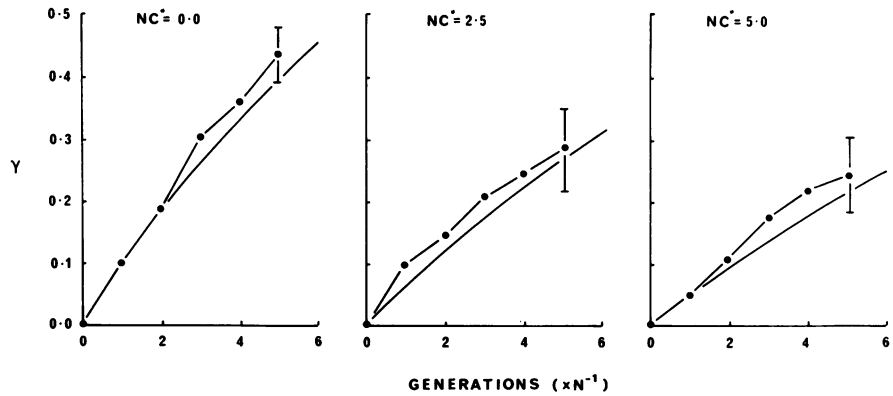


FIG. 6.—Average values of the parameter $\gamma$ which estimates the mean rate of accumulation of mutational changes at a locus, based on the observed gene frequencies in pairs of contemporary isolated populations of common origin. The means have been obtained by substituting average values of $H$ and $H_B$ in equation (8). The theoretical curves are based on equation (18) with $r^* = 0.050$, 0.033, and 0.024, respectively, for the regimes $NC^* = 0.0$, 2.5, and 5.0. The values of $r^*$ for $NC^* = 2.5$ and 5.0 were determined empirically in an earlier study of long-term evolutionary changes in single isolated populations [6].

study for a random sample of loci, both polymorphic and monomorphic, by comparison with theoretical curves given by equation (18) using $r^* = 0.050$, 0.033, and 0.024, respectively, for the regimes $NC^* = 0.0$, 2.5, and 5.0. Unlike parameters related to the coefficient of kinship, the values of $\gamma$ can be seen to reflect accurately the effects of natural selection in reducing the average rate of gene substitution.

*Genetic Distances Separating the Races of Man*

Our discussion has so far been restricted to computer populations whose evolution and differentiation is completely documented, enabling firm conclusions to be reached as to the behavior of a number of measures of genetic distance over evolutionary time. We shall now consider the magnitude and interpretation of these same distance measures derived from the available data on genetic variation in human populations.

Two sets of calculations have been carried out: (1) comparisons of Caucasoids, Negroids, and Mongoloids based on estimates of $\phi_S$, $f_\theta$, $\phi_Y$, and $\phi^*$ from a sample of 15 polymorphic loci; and (2) a comparison of Caucasoids and Negroids in terms

of $\phi^*$ and $\gamma$, estimated from a sample of 40 loci, both polymorphic and monomorphic. The results are presented in tables 2 and 3.

The gene frequency data for the calculations were obtained primarily from two sources [1, 22]. The total sample of 40 loci includes those listed by Cavalli-Sforza

TABLE 2

ESTIMATES OF GENETIC DISTANCE BETWEEN CAUCASOIDS AND NEGROIDS BASED ON A SAMPLE OF 40 LOCI, BOTH POLYMORPHIC AND MONOMORPHIC

| RACIAL COMPARISON | MEAN HETEROZYGOSITY | GENETIC DISTANCE MEASURE† | |
|---|---|---|---|
| | | $\phi^*$ | $\gamma$ |
| Causacoid-Negroid ..... | .164 ± .006 | .214 ± .065 | .054 ± .024 |

NOTE.—Sources of the data are given in the text.

† $\phi^*$ is Latter's measure of kinship, defined by equations (5) and (6); $\gamma$ is Latter's measure of mutational divergence, defined by equations (7) and (8).

TABLE 3

ESTIMATES OF GENETIC DISTANCE BETWEEN THE MAJOR RACIAL GROUPS BASED ON 15 POLYMORPHIC LOCI

| RACIAL COMPARISON | GENETIC DISTANCE MEASURE† | | | |
|---|---|---|---|---|
| | $\phi_S$ | $f_\theta$ | $\phi_Y$ | $\phi^*$ |
| Caucasoid-Negroid ..... | .094 ± .032 | .257 ± .100 | .102 ± .040 | .220 ± .075 |
| Caucasoid-Mongoloid ..... | .061 ± .017 | .175 ± .051 | .067 ± .022 | .138 ± .042 |
| Negroid-Mongoloid ..... | .125 ± .048 | .323 ± .122 | .155 ± .058 | .310 ± .094 |

NOTE.—Sources of the gene frequency data are given in the text.

† $\phi_S$ is a derivative of Sanghvi's measure, defined by equations (1) and (14); $f_\theta$ is Cavalli-Sforza's measure, defined by equation (2); $\phi_Y$ is Yasuda's measure, defined by equation (3); $\phi^*$ is Latter's measure of kinship, defined by equations (5) and (6).

and Bodmer [22, table 11.7] with complete data for all three racial groups, plus additional loci listed by Harris [1, table 8.5]. The complete set of loci used for the calculations of table 2 is as follows: ABO, MNS, P, Rh, Lu, K, Le, Fy, Jk, Di, Xg, Hp, Inv, Tf, Gc, Lp, red-cell acid phosphatase, 6-phosphogluconate dehydrogenase, phosphoglucomutase ($PGM_1$, $PGM_2$, $PGM_3$), adenylate kinase, pseudocholinesterase, peptidases A, B, C, and D, adenosine deaminase, phosphohexose isomerase, malate dehydrogenase, isocitrate dehydrogenase, red-cell hexokinase, lactate dehydrogenase, methemoglobin reductase, red-cell pyrophosphatase, pyruvate kinase, placental acid phosphatase, nucleoside phosphorylase, triosephosphate isomerase, and a red-cell oxidase.

Gene frequencies for these loci have been taken from Cavalli-Sforza and Bodmer [22] and Harris [1] for all loci except Gc: for this locus the data of Bearn, Bowman,

and Kitchin [23] for the Swedish, Nigerian Habe, and Chinese populations were used. Two loci not included in the sample were the $\beta$-lipoprotein locus Ag (because the frequencies given for Negroes by Cavalli-Sforza and Bodmer could not be verified) and G6PD (because of its known association with malaria). It is important to point out that Cavalli-Sforza and Bodmer's data for each of the three racial groups refers to the largest single sample of data showing Hardy-Weinberg equilibrium, wherever possible representing populations from northern Europe for Caucasoids, central or southern Africa for Negroids, and China for Mongoloids.

The polymorphic loci used for the calculation of table 3 are those in the above sample for which accurate data are available for all racial groups, with a frequency of heterozygotes greater than 10% in at least one of the three. These are ABO, MNS, P, Rh, Le, Fy, Jk, Xg, Hp, Inv, Gc, Lp, red-cell acid phosphatase, 6-phosphogluconate dehydrogenase, and $PGM_1$.

A comparison of the distance measures of tables 2 and 3 with the computed population distances depicted in figures 1–6 leads to the following conclusions: (1) The six measures of distance between representative Caucasoid and Negroid populations indicate that the two races have undergone the equivalent of approximately $0.5N$ generations of independent evolution. The Negroid-Monogoloid separation is of the order of $0.7N$ generations. The parameter $N$ here denotes the harmonic mean of effective population size since the separation of the races concerned. The value of $N$ appears to be of the order of 5,000–10,000 individuals, based on calculations which assume the commonly occurring blood group, serum protein, and enzyme variants to be very nearly selectively neutral. This corresponds to a time scale of something like 50,000–100,000 years [24]. (2) The measures $f_\theta$ and $\phi^*$ in table 3 are in excellent agreement, as expected from the computer results of figures 2 and 4. It must be concluded that $\phi_S$ and $\phi_Y$ grossly underestimate the actual coefficient of kinship in human populations, being subject to the same bias as the statistic $\theta$ defined by equation (15).

## CONCLUSIONS

1. Alternative measures of the genetic distance between contemporary computer populations of known evolutionary history have been shown to give appreciably different estimates of the coefficient of kinship at all levels of inbreeding.

2. No one measure provides a completely satisfactory estimate of kinship, but the parameters $f_\theta$ suggested by Cavalli-Sforza [3] and $\phi^*$ proposed by Latter [7] give useful approximations to the value of $\phi$ specified by equation (11).

3. The parameters $\phi_Y$ proposed by Yasuda [12], and $\phi_S$ based on Sanghvi's measure of distance $G_S^2$ [2] have been shown to seriously underestimate the coefficient of kinship. Both measures are based on a form of Wahlund's expression which is inappropriate for the comparison of only two populations.

4. All five parameters related to the coefficient of kinship, namely, $\phi_S$, $f_\theta$, $\phi_Y$, $\theta$, and $\phi^*$, have been shown to be insensitive to natural selection for an intermediate level of gene or enzyme activity.

5. The rate of *mutational* divergence, on the other hand, is estimated by the

parameter $\gamma$ given by equation (8) and is closely related to the true *rate of gene substitution* observed in the computer populations, whether in the presence or absence of natural selection.

6. An analysis of gene frequency data from the three major racial groups gives a set of distance estimates which is entirely consistent with the foregoing conclusions. The calculations indicate that the Caucasoid-Negroid divergence is the equivalent of approximately $0.5N$ generations of independent evolution, where $N$ denotes the harmonic mean of effective population size since the separation of the races.

SUMMARY

The early stages of genetic differentiation between completely isolated contemporary populations have been studied by means of computer simulation. The model assumes a constant effective population size, $N$, with continual mutation to new alleles at a rate compatible with observed levels of heterozygosity in man and other animals. Loci subject to natural selection for an optimal level of gene or enzyme activity have been studied as well as strictly neutral loci.

Four measures of genetic distance have been shown to lead to appreciably different estimates of the coefficient of kinship. Cavalli-Sforza's $f_\theta$ and Latter's $\phi^*$ give useful approximations to the true value, particularly at low levels of inbreeding.

A measure of mutational divergence between isolated populations ($\gamma$) has also been studied and shown to reflect accurately the true rate of gene substitution due to the effects of drift, mutation, and natural selection for optimal enzyme activity.

An analysis of gene frequency data from the three major racial groups indicates that the separation between Caucasoids and Negroids is the equivalent of approximately $0.5N$ generations of independent evolution, where $N$ denotes the harmonic mean of effective population size since the separation of the races. The value of $N$ appears to be of the order of 5,000–10,000 individuals, corresponding to a time scale of roughly 50,000–100,000 years.

## REFERENCES

1. HARRIS H: *The Principles of Human Biochemical Genetics.* Amsterdam, North Holland, 1970
2. BALAKRISHNAN V, SANGHVI LD: Distance between populations on the basis of attribute data. *Biometrics* 24:859–865, 1968
3. CAVALLI-SFORZA LL: Human diversity, in *Proceedings 12th International Congress Genetics,* vol 3, 1969, pp 405–416
4. MORTON NE: Human population structure. *Ann Rev Genet* 3:53–74, 1969
5. SELANDER RK: Biochemical polymorphism in populations of the house mouse and old-field mouse. *Sympos Zool Soc London* 26:73–91, 1970
6. LATTER BDH: Selection in finite populations with multiple alleles. III. Genetic divergence with centripetal selection and mutation. *Genetics* 70:475–490, 1972
7. LATTER BDH: The island model of population differentiation: a general solution. *Genetics.* In press, 1973
8. SANGHVI LD: Comparison of genetical and morphological methods for a study of biological differences. *Amer J Phys Anthrop* N.S. 11:385–404, 1953

9. WORKMAN PL, NISWANDER JD: Population studies on southwestern Indian tribes. II. Local genetic differentiation in the Papago. *Amer J Hum Genet* 22:24–49, 1970
10. CAVALLI-SFORZA LL, EDWARDS AWF: Phylogenetic analysis: models and estimation procedures. *Evolution* 21:550–570, 1967
11. CAVALLI-SFORZA LL, ZONTA LA, NUZZO F, BERNINI L, DE JONG WWW, MEERA KHAN P, RAY AK, WENT LN, SINISCALCO M, NIJENHUIS LE, VAN LOGHEM E, MODIANO G: Studies on African Pygmies. I. A pilot investigation of Babinga Pygmies in the Central African Republic. *Amer J Hum Genet* 21:252–274, 1969
12. YASUDA N: An extension of Wahlund's principle to evaluate mating type frequency. *Amer J Hum Genet* 20:1–23, 1968
13. NEI M: Genetic distance between populations. *Amer Nat* 106:283–292, 1972
14. LATTER BDH: Selection in finite populations with multiple alleles. II. Centripetal selection, mutation and isoallelic variation. *Genetics* 66:165–186, 1970
15. YOSHIDA A: Amino acid substitution (histidine to tyrosine) in a glucose-6-phosphate dehydrogenase variant (G6PD Hektoen) associated with over-production. *J Molec Biol* 52:483–490, 1970
16. HARRIS H: Enzyme polymorphisms in man. *Proc Roy Soc* [*Biol*] 164:298–310, 1966
17. LEWONTIN RC: An estimate of average heterozygosity in man. *Amer J Hum Genet* 19:681–685, 1967
18. SELANDER RK, YANG SY: Protein polymorphism and genic heterozygosity in a wild population of the house mouse (*Mus musculus*). *Genetics* 63:653–667, 1969
19. PRAKASH S, LEWONTIN RC, HUBBY JL: A molecular approach to the study of genic heterozygosity in natural populations. IV. Patterns of genic variation in central, marginal and isolated populations of *Drosophila pseudoobscura*. *Genetics* 61:841–858, 1969
20. MORTON NE, HARRIS DE, YEE S, LEW R: Pingelap and Mokil Atolls: migration. *Amer J Hum Genet* 23:339–349, 1971
21. MORTON NE, YEE S, HARRIS DE, LEW R: Bioassay of kinship. *Theor Pop Biol* 2:507–524, 1971
22. CAVALLI-SFORZA LL, BODMER WF: *The Genetics of Human Populations*. San Francisco, W. H. Freeman, 1971
23. BEARN AG, BOWMAN BH, KITCHIN FD: Genetic and biochemical considerations of the serum group-specific component. *Cold Spring Harbor Symp Quant Biol* 29:435–442, 1964
24. LATTER BDH: Measures of genetic distance between individuals and populations, in *The Genetics of Population Structure*, edited by MORTON NE, Honolulu, Univ. Hawaii Press. In press, 1973