Jap. J. Human Genet. **23**, 341–369, 1978

THE THEORY OF GENETIC DISTANCE AND EVOLUTION OF HUMAN RACES[1]

(The Japan Society of Human Genetics Award Lecture)

Masatoshi NEI

Center for Demographic and Population Genetics, University of Texas at Houston, Houston, Texas 77025, U.S.A.

Summary 1. Theoretical works on Nei's genetic distance and its extensions are discussed. New formulae for the sampling variances of genetic distance estimates are presented. Formulae for the genetic identity of genes at the electrophoretic level when the mutation rate varies from locus to locus are also presented.

2. Empirical data suggests that the rate of gene substitution or mutation rate per locus varies considerably among protein loci, and if this factor is taken into account, the rate of decline of genetic identity (I) is no longer constant but decreases with evolutionary time. Using both the infinite-allele model and the stepwise mutation model, the numerical relationship between I and evolutionary time is presented. This relationship may be used for estimating the time after divergence between populations. The value of genetic distance or genetic identity is also affected considerably by the bottleneck effect. The bottleneck effect generally accelerates the increase of genetic distance with time, and the effect remains for a long time after the population size returns to the original level. A method for correcting for this effect is presented.

3. Application of the theory of genetic distance to data on protein polymorphism in man indicates that the genetic variation between the three major races, Caucasoid, Negroid, and Mongoloid, is much smaller than the variation within them, despite the fact that there is a conspicuous difference in some morphological characters such as pigmentation, facial structure, and hair texture. It is proposed that the differentiation of these morphological characters was brought about by relatively strong natural selection through a small number of gene substitutions, whereas general protein loci are subject to little or very weak selection. Analysis of blood group gene frequency data gives essentially the same result as those from protein loci, though they are likely to have been affected by nonrandom

Received August 12, 1978

[1] Presented at the 22nd Annual Meeting of the Japan Society of Human Genetics at Ube, Japan, November 10–12, 1977.

sampling of the loci. It is also shown that at the protein level the racial differences in man correspond to those between local races in other organisms.

4. Rough estimates of the number of codon differences between an individual of man and his various relatives are presented. It seems that the mean number of codon differences between man and chimpanzee is about 10 times larger than that between second degree relatives in Caucasians or Japanese, but about 1/19 of that between man and horse.

5. Genetic distance estimates suggest that among the three major races of man the first divergence occurred about 120,000 years ago between Negroid and a group of Caucasoid and Mongoloid and then the latter group split into Caucasoid and Mongoloid around 60,000 years ago. It is also shown that the genetic identity between man and chimpanzee corresponds to a divergence time of 4–6 million years if the assumption of constant rate of amino acid substitution is correct.

6. Methods of constructing a phylogenetic tree from genetic distance estimates are discussed. For constructing the topology of a tree, Fitch and Margoliash's method is quite efficient. For estimating branch lengths, however, Nei's method of averaging distances seems to be better.

7. A phylogenetic tree for twelve races of man is constructed by using gene frequency data for 11 protein and 11 blood group loci. This tree roughly agrees with what we expect intuitively from the morphological characters and the historical record of these races.

INTRODUCTION

In the last ten years the study of population genetics and evolution has been revitalized by the introduction of molecular techniques. Stimulated by the pioneering works by Harris (1966) and Lewontin and Hubby (1966), many authors have examined the extent of protein polymorphism in natural populations of various organisms including man. It is now clear that virtually all species are highly polymorphic at the protein level. On the other hand, comparative studies of proteins from various organisms have shown that the amino acid sequence of protein is subject to constant change in the evolutionary process, and the rate of amino acid substitution is roughly constant per year (Zuckerkandl and Pauling, 1962, 1965; Margoliash and Smith, 1965; Dayhoff, 1969). To explain these observations about protein polymorphism in populations and amino acid substitution in evolution, Kimura (1968) proposed the so-called neutral mutation or mutation-drift hypothesis. The last ten years witnessed a great deal of controversy over this hypothesis (see for example Nei (1975) and Ayala (1976)).

There is another line of study which has been affected tremendously by molecular techniques; it is the study of phylogenetic relationships of organisms or popu-

lations. Previously, the phylogeny of organisms was studied mainly by using fossil records or morphological characters. However, fossil records are available only for a limited group of organisms, and even if they are available they generally do not reveal the detailed evolutionary relationship of closely related species or races (Carlson *et al.*, 1978). In fact, for the study of racial evolution, fossil records are virtually useless even in man, where fossil records are probably most abundant. On the other hand, morphological characters are always measurable, and thus it is possible to construct a phylogenetic tree for any group of species or populations in terms of intuitive or numerical taxonomy. However, the evolutionary change of morphological characters is so complex that it is not generally directly related to the evolutionary time. Compared with morphological characters, macromolecules such as protein and DNA show a very simple pattern of evolutionary change, and the amount of change in amino acid sequences of proteins or nucleotide sequences of DNA is roughly proportional to the evolutionary time. Because of this simple relationship between amino acid substitution and evolutionary time, the phylogenetic tree constructed from data on amino acid substitution is considered to be much more reliable than that based on morphological characters.

In practice, however, data on amino acid differences for a single protein are not very useful for constructing a phylogenetic tree for closely related species or races, since these species or races often do not show any difference in amino acid sequence. In this case, an adequate evaluation of genetic difference between species or races can be made only when a large number of proteins is examined. At the present time, sequencing amino acids in proteins is time-consuming and expensive, so that racial or interspecific genetic differences are studied mainly by electrophoresis. Furthermore, these differences are generally so small that they can be described only in terms of gene frequencies rather than the presence or absence of a particular protein type. How can we then relate these data to the evolutionary process? This was exactly the problem I faced in late 1969. First, I solved this problem, assuming that the polymorphism within populations is negligible (Nei, 1971). But later, I developed a more general theory in which the intrapopulational polymorphism was taken into account (Nei, 1972).

However, this was not the end of the work but just the beginning. Application of this theory to human data produced an interesting finding. Furthermore, data analysis has demanded a more realistic and elaborate mathematical theory, which is still being improved in collaboration with my colleagues. In this paper, I would like to review this theoretical work and then discuss the results of application of the theory to human data. I shall present some new results from both theoretical work and data analysis.

GENETIC DISTANCE

Measures of genetic distance

Genetic distance is the degree of genetic difference between a pair of populations as measured by some numerical method. There are many different measures of genetic distance. Some of these are direct applications of earlier measures of morphological distances which have been used in the classical numerical taxonomy. For example, the measures proposed by Sanghvi (1953), Steinberg *et al.* (1967) and Balakrishnan and Sanghvi (1968) are all direct applications of Mahalanobis' (1936) D^2 statistic to gene frequency data. Bhattacharyya's (1946) measure, which is essentially the same as Cavalli-Sforza and Edwards' (1967) distance measure, can also be regarded as an extension of Mahalanobis' D^2 statistic for the case of multinomially distributed characters (see Nei (1977a) and Smith (1977) for recent reviews of genetic distance). In these theories populations are represented as points in a multidimensional space and the genetic distance between two populations is measured by the geometric distance between the corresponding points in the space. Thus, the principle of triangle inequality is very important, but little attention is paid to the relationship between the distance measure and the evolutionary process. The absolute values of these measures do not have any particular biological meaning, and only the relative values are important for finding the genetic relationship among populations.

Compared with these measures, Nei's (1972) distance measure is based on an entirely different concept. Namely, it is intended to measure the number of codon substitutions per locus that have occurred after divergence of the two populations under consideration. Thus, the absolute value of this measure has a clear-cut biological meaning. Theoretically, Nei's method can be applied to any pair of taxa, whether they are local populations, species, or genera, if enough data are available. Of course, the current techniques of studying gene frequencies, such as electrophoresis and immunological reaction, cannot detect all codon differences, so that we are forced to deal with only those codon differences that are detectable by the current techniques, though some correction for undetectable codons can be made under certain circumstances, as will be discussed later. Furthermore, there are some other statistical problems which make it difficult to estimate the exact number of codon differences. For these reasons, I have proposed three different measures of genetic distance, *i.e.*, the *minimum*, *standard*, and *maximum estimates* of codon differences per locus (Nei, 1973a).

Definition of Nei's distance measures

Consider two populations, X and Y, in which multiple alleles are segregating at a locus. Let x_i and y_i be the frequencies of the i-th allele in X and Y, respectively. The probability of identity of two randomly chosen genes is $j_X = \sum x_i^2$ in

population X, whereas it is $j_Y = \sum y_i^2$ in population Y. The probability of identity of two genes, chosen at random, one from each of the two populations, is $j_{XY} = \sum x_i y_i$. Note that the identity of genes defined in this way requires no assumptions about selection, mutation, and migration. We designate by J_X, J_Y and J_{XY} the arithmetic means of j_X, j_Y, and j_{XY} over all loci, including monomorphic ones, respectively. Clearly, $D_{X(m)} = 1 - J_X$, $D_{Y(m)} = 1 - J_Y$, and $D_{XY(m)} = 1 - J_{XY}$ are all equal to the proportion of different genes between two randomly chosen genomes from the respective populations. In other words, $D_{X(m)}$ and $D_{Y(m)}$ are minimum estimates of codon differences per locus between two randomly chosen genomes from populations X and Y respectively, whereas $D_{XY(m)}$ is a minimum estimate of codon differences per locus between two randomly chosen genomes, one from each of X and Y. ($D_{X(m)}$ and $D_{Y(m)}$ are equal to average heterozygosity.) Therefore,

$$D_m = D_{XY(m)} - (D_{X(m)} + D_{Y(m)})/2 \tag{1}$$

is a minimum estimate of net codon differences per locus between X and Y when the intrapopulational codon differences are subtracted. I have called D_m the *minimum genetic distance.*

The drawback of D_m is that $D_{X(m)}$, $D_{Y(m)}$, and $D_{XY(m)}$ are the proportions of different genes between two randomly chosen genomes, so that they are not proportional to the number of codon differences. Thus, D_m may be a gross underestimate of the number of net codon differences when $D_{XY(m)}$ is large. If individual codon changes are independent and follow a Poisson distribution, the mean number of net codon differences may be given by

$$D = -\ln I, \tag{2}$$

where

$$I = J_{XY}/\sqrt{J_X J_Y} \tag{3}$$

is the normalized identity of genes (or genetic identity) between X and Y. I have called D the *standard genetic distance.* It is noted that D can be written as $D = D_{XY} - (D_X + D_Y)/2$, where $D_{XY} = -\ln J_{XY}$, $D_X = -\ln J_X$, and $D_Y = -\ln J_Y$. As will be seen later, if the rate of gene (codon) substitution per year is constant, D is linearly related to the time after divergence between two populations. Also, under certain migration models D is linearly related to the geographical distance or area (Nei, 1972).

If the rate of codon changes varies from locus to locus, D still may be an underestimate of codon differences. In this case the mean number of net codon differences may be estimated by

$$D' = -\ln I', \tag{4}$$

where $I' = J'_{XY}/\sqrt{(J'_X J'_Y)}$ in which J'_{XY}, J'_X, and J'_Y, are the geometric means of j_{XY}, j_X, and j_Y, respectively, over different loci. In practice, however, D' is affected considerably by sampling errors of gene frequencies at the time of popula-

tion survey as well as by random genetic drift. These factors are expected generally to inflate the estimate of the mean number of net codon differences. Therefore, I call D' the *maximum genetic distance*. If any of the values of $j_{XY}/\sqrt{j_X j_Y}$ for individual loci is small, D' can be a gross overestimate. In fact, if there is a single locus at which there is no common allele between two populations, D' is infinitely large.

Recently, we developed a somewhat different formula for this case, assuming that the rate of codon substitution varies among loci following the gamma distribution with coefficient of variation 1 (Nei *et al.*, 1976a). It is given by

$$D_v = (1-I)/I. \tag{5}$$

The rationale of this formula will be discussed later. This distance measure seems to be superior to D', since it is not affected by sampling error so strongly.

Estimation of genetic distance

Theoretically, the genetic distance between two populations is defined in terms of population gene frequencies for all loci. In practice, however, it is virtually impossible to examine all genes in the populations. Therefore, we must estimate the genetic distance by sampling a certain number of individuals from the populations and examining a certain number of loci. Let us now consider how to estimate genetic distance from actual data, following Nei and Roychoudhury (1974a) and Nei (1978).

Clearly, there are two sampling processes involved in this case, *i.e.*, sampling of loci from the genome and sampling of individuals (genes) from the population. In the following we assume that r loci are chosen at random and n individuals ($2n$ genes) are examined for each locus. Let $\hat{x}_i$ and $\hat{y}_i$ be the frequencies at the i-th allele at a locus in the samples of $2n$ genes from populations X and Y, respectively. The usual method of estimating genetic distance is to replace x_i and y_i in (1), (3), or (4) by $\hat{x}_i$ and $\hat{y}_i$, respectively.

However, when sample size is small, this method gives a biased estimate (Nei, 1973a, 1978). The unbiased (or less biased) estimates of D_m, D, D', and D_v may be obtained by replacing $\sum x_i^2$, $\sum y_i^2$, and $\sum x_i y_i$ in the formulae for genetic distance by the unbiased estimates of these quantities. The unbiased estimates of $\sum x_i^2$, $\sum y_i^2$, and $\sum x_i y_i$ are given by $\hat{j}_X = (2n\sum \hat{x}_i^2 - 1)/(2n-1)$, $\hat{j}_Y = (2n\sum \hat{y}_i^2 - 1)/(2n-1)$, and $\hat{j}_{XY} = \sum x_i y_i$, respectively, whereas the unbiased estimates ($\hat{J}_X$, $\hat{J}_Y$, and $\hat{J}_{XY}$) of J_X, J_Y, and J_{XY} are the respective averages of $\hat{j}_X$, $\hat{j}_Y$, and $\hat{j}_{XY}$ over loci. For example, the unbiased (or less biased) estimates of D_m and D ($\hat{D}_m$ and $\hat{D}$, respectively) may be obtained by

$$\hat{D}_m = (\hat{J}_X + \hat{J}_Y)/2 - \hat{J}_{XY}, \tag{6}$$

and

$$\hat{D} = -ln[\hat{J}_{XY}/\sqrt{\hat{J}_X \hat{J}_Y}]. \tag{7}$$

The sampling variances of $\hat{D}_m$ and $\hat{D}$ have also been worked out (Nei and Roychoudhury, 1974a; Nei, 1978), but I shall not discuss them here. The only sampling variances I would like to present are those of $\hat{I}$ and $\hat{D}_v$ (estimates of I and D_v respectively), which are not given in my earlier papers. They are approximately given by

$$V(\hat{I})=\frac{\hat{J}_{XY}^2}{4\hat{J}_X^3\hat{J}_Y}V(\hat{J}_X)+\frac{\hat{J}_{XY}^2}{4\hat{J}_X\hat{J}_Y^3}V(\hat{J}_Y)+\frac{1}{\hat{J}_X\hat{J}_Y}V(\hat{J}_{XY})$$
$$+\frac{\hat{J}_{XY}^2}{2\hat{J}_X^2\hat{J}_Y^2}\mathrm{Cov}(\hat{J}_X,\hat{J}_Y)-\frac{\hat{J}_{XY}}{\hat{J}_X^2\hat{J}_Y}\mathrm{Cov}(\hat{J}_X,\hat{J}_{XY})$$
$$-\frac{\hat{J}_{XY}}{\hat{J}_X\hat{J}_Y^2}\mathrm{Cov}(\hat{J}_Y,\hat{J}_{XY}), \tag{8}$$

$$V(\hat{D}_v)=\frac{\hat{J}_Y}{4\hat{J}_X\hat{J}_{XY}^2}V(\hat{J}_X)+\frac{\hat{J}_X}{4\hat{J}_Y\hat{J}_{XY}^2}V(\hat{J}_Y)+\frac{\hat{J}_X\hat{J}_Y}{\hat{J}_{XY}^4}V(\hat{J}_{XY})$$
$$+\frac{1}{2\hat{J}_{XY}^2}\mathrm{Cov}(\hat{J}_X,\hat{J}_Y)-\frac{\hat{J}_Y}{\hat{J}_{XY}^3}\mathrm{Cov}(\hat{J}_X,\hat{J}_{XY})$$
$$-\frac{\hat{J}_X}{\hat{J}_{XY}^3}\mathrm{Cov}(\hat{J}_Y,\hat{J}_{XY}), \tag{9}$$

where V(•) and Cov(•,•) refer to the variance and covariance respectively. For example,

$$V(\hat{J}_X)=\sum_{k=1}^{r}(\hat{J}_{X(k)}-\bar{J}_X)^2/[r(r-1)],$$

$$Cov(\hat{J}_X,\hat{J}_Y)=\sum_{k=1}^{r}(\hat{J}_{X(k)}-\bar{J}_X)(\hat{J}_{Y(k)}-\bar{J}_Y)/[r(r-1)].$$

It is noted that in many vertebrate species the single-locus identity ($I_j=j_{XY}/\sqrt{j_X j_Y}$) shows a U-shaped distribution and is often mostly either 1 or 0. In this case the variance of $\hat{I}$ is approximately given by (Nei, 1971).

$$V(\hat{I})=\hat{I}(1-\hat{I})/r \tag{10}$$

We have developed a computer program for computing the unbiased estimates of genetic distance and their sampling errors. It is available by writing to the author.

In planning a survey of gene frequencies to estimate genetic distance it is important to know how many loci and how many individuals per locus should be examined when the total number of genes to be surveyed is fixed. This problem has been studied by Nei and Roychoudhury (1974a) and Nei (1978) by decomposing the variance of genetic distance into the variance among loci and the variance due to sampling of genes within loci. The results obtained indicate that the interlocus variance is much larger than the intra-locus variance unless n is extremely small, and thus it is important to study a large number of loci rather than a large number

of individuals per locus for reducing the variance of the estimate of genetic distance.

As will be mentioned later, the genetic distance can be used for estimating the time after separation of two populations under certain assumptions. In this case the standard error of the estimate of separation time may be computed from the variance of genetic distance considered above. The variance can also be used for testing the difference between two estimates of genetic distances if independent sets of loci are used for computing the two distance estimates. In practice, however, it is customary to use the same set of loci for computing distance estimates for all pairs of populations. In this case, the variances obtained from (8) and (9) are not appropriate for testing the statistical difference between distance estimates. This is because they include the variance due to the differences in the initial gene frequencies among loci at the time of population differentiation (Li and Nei, 1975). At the present time, there seems to be no method to eliminate this component from the total variance.

So far we have been interested in the genetic distance defined as the number of codon differences per locus, so that a large number of loci are required for estimating this quantity. However, collection of gene frequency data is time-consuming, and under certain circumstances only a few loci are available for the study of gene differences. In this case the estimate of genetic distance may deviate considerably from the real value. When local populations within the same species are compared, this deviation is expected to be generally upward, since gene frequencies are studied more often with highly polymorphic loci than with less polymorphic loci, and monomorphic loci in these populations almost always have the same allele. However, if one is interested only in relative values of genetic distance among several populations, the estimate of distance based on a few polymorphic loci would still be useful. As relative distances, all the measures discussed here can be used for any case because they depend on no assumptions about the evolutionary forces.

GENETIC DIFFERENTIATION OF POPULATIONS

Genetic differentiation of populations occurs only when the populations are partially or completely isolated from each other. Let us now consider the process of genetic differentiation of populations in terms of the distance measures considered above.

Complete isolation: General case

When two populations are reproductively isolated, they tend to accumulate different genes due to mutation, selection, and genetic drift. If we make a certain assumption, this problem can be studied by a simple mathematical model. The assumptions we make are as follows: (1) A population splits into two populations (X and Y) at a certain evolutionary time and thereafter no migration occurs between

the two populations. (2) Populations X and Y are in equilibrium with respect to the effects of mutation, selection, and random genetic drift, so that the average gene identities (J_X and J_Y) within populations remain constant. This assumption seems to be satisfactory in most natural populations, since closely related populations or species generally show the same degree of heterozygosity. (3) All new mutations are different from the alleles existing in the populations (infinite-allele model). This assumption seems to be satisfactory if alleles are identified at the codon (amino acid) level but probably not if they are studied by electrophoresis. I shall discuss the effect of violation of this assumption later. (4) The rate of gene substitution per locus per year (α) remains constant and is the same for all loci. The first part of this assumption seems to be roughly correct at the amino acid level (*e.g.*, Nei, 1975; Fitch, 1976; Wilson *et al.*, 1977), but the second part is certainly incorrect. However, the effect of varying rates of gene substitution among loci can be corrected, as will be seen later. It can be shown that α is equal to the mutation rate per year (v) if all mutations are neutral, whereas it is equal to $4Nsv$ if mutant genes are advantageous and semidominant, where N is the effective population size and s is the selective advantage of a mutant gene (Kimura and Ohta, 1971).

Under the above assumptions, Nei (1972, 1975) has shown that the genetic identity at the t-th generation is

$$I_A = I_0 e^{-2\alpha t}, \tag{11}$$

where is the I_0 value of I at time 0. Therefore, we have

$$D = 2\alpha t + D_0, \tag{12}$$

where $D_0 = -ln\, I_0$. In general, I_0 is close to 1, so that $D_0 \simeq 0$. It is therefore clear that D measures the accumulated number of gene (codon) substitutions per locus between the two populations.

As mentioned earlier, however, the assumption that α is the same for all loci is certainly wrong. Indeed, Nei *et al.* (1976a) have shown that the rate of amino acid substitution (per protein) varies considerably with protein and is distributed roughly as a gamma distribution with coefficient of variation 1. They also showed that the subunit molecular weights of the proteins that are often used for electrophoresis also show a gamma distribution. Furthermore, studies on the variances of single-locus heterozygosity and genetic distance in various organisms (more than one hundred different organisms) have suggested that the distribution of the rate of gene substitution or mutation rate roughly follows the gamma distribution with coefficient of variation 1 (Nei *et al.*, 1976b; Fuerst *et al.*, 1977; Chakraborty *et al.*, 1978). Zouros' (1979) recent study on the relative mutation rates supports this conclusion. It is also noted that the variation of mutation rate is apparently related to the subunit molecular weight of protein (Koehn and Eanes, 1977; Nei *et al.*, 1978).

Let us therefore assume that α has the following gamma distribution

$$f(\alpha) = \frac{b^a}{\Gamma(a)} e^{-b\alpha} \alpha^{a-1},$$

where $a=\bar{\alpha}^2/V(\alpha)$ and $\beta=\bar{\alpha}/V(\alpha)$, in which $\bar{\alpha}$ and $V(\alpha)$ are the mean and variance of α. We know that the expected genetic identity for a locus with a given value of α in the t-th generation after population splitting is $e^{-2\alpha t}$ if we assume $I_0=1$. Therefore, the expected genetic identity over all loci is

$$\bar{I}_A=\int_0^\infty f(\alpha)e^{-2\alpha t}d\alpha=\left(\frac{a}{a+2\bar{\alpha}t}\right)^a. \tag{13}$$

When the coefficient of variation ($a^{-1/2}$) is 1,

$$\bar{I}_A=1/(1+2\bar{\alpha}t). \tag{14}$$

Therefore, the mean number of gene substitutions per locus ($2\bar{\alpha}t$) is given by D_v in (5). Namely,

$$D_v=2\bar{\alpha}t. \tag{15}$$

Mathematically, $D_v>D$, but the difference between (11) and (14) is small when t is relatively small (see Table 1).

Formula (12) or (15) enables us to estimate the time after divergence between two populations, if α is known. Using the average rate of amino acid substitution for 22 proteins that are often used for electrophoresis, Nei (1975) estimated $\bar{\alpha}$ to be 10^{-7} for electrophoretic data. Therefore, t is estimated by

$$t=5\times 10^6\times D_v. \tag{16}$$

It should be emphasized, however, that the above value of α is based on a number of assumptions and thus (16) gives only a very rough estimate of divergence time. If our estimate of $\bar{\alpha}$ improves in the future, (16) should be modified accordingly.

Table 1. Evolutionary time and genetic identity under the infinite-allele model (I_A,$\bar{I}_A$) and the stepwise mutation model (I_E,$\bar{I}_E$). I_A,$\bar{I}_A$, I_E, and $\bar{I}_E$ were obtained by formulae (11), (14), (17), and (19), respectively. In this computation the rate of gene substitution ($\alpha=\nu$) was assumed to be 10^{-7} per year (see text). The accumulated number of codon substitutions may be estimated by $2\alpha t$, if the observed value of I is given.

Time ($\times 10^3$ yrs)	I_A	$\bar{I}_A$	I_E	$\bar{I}_E$	Time ($\times 10^6$ yrs)	I_A	$\bar{I}_A$	I_E	$\bar{I}_E$
10	.998	.998	.998	.998	1	.819	.833	.827	.845
50	.990	.990	.990	.990	2	.670	.714	.697	.745
100	.980	.980	.980	.981	3	.549	.625	.599	.674
200	.961	.961	.961	.962	4	.449	.556	.524	.620
300	.942	.943	.943	.945	5	.368	.500	.466	.577
400	.923	.926	.925	.928	6	.301	.455	.420	.542
500	.905	.909	.907	.913	7	.247	.417	.383	.513
600	.887	.893	.890	.898	8	.202	.385	.353	.488
700	.869	.877	.874	.884	9	.165	.357	.329	.466
800	.852	.862	.858	.870	10	.135	.333	.309	.447
900	.835	.847	.842	.857	20	.018	.200	.207	.333

It should also be emphasized that formula (11) or (14) is valid only when a large number of loci is studied since each event of gene substitution is subject to a large stochastic error. Nei and Tateno (1975) studied the distribution of single-locus gene identity ($I_j=j_{XY}/\sqrt{j_X j_Y}$) under the assumption of neutral mutations by using computer simulation. The results obtained show that when $2\bar{\alpha}t$ is small, I_j shows an inverse J-shaped distribution, whereas it shows a U-shaped distribution when $2\bar{\alpha}t$ is moderately large. Therefore, to obtain a reliable estimate of I a large number of loci must be studied. This is true even if gene substitution is mediated by natural selection (Chakraborty *et al.*, 1977). The mathematical formulae for obtaining the stochastic variance of genetic distance under the assumption of neutral mutations have been obtained by Li and Nei (1975).

It should be mentioned that formula (11) or (16) is not valid for large t when it is applied to electrophoretic data. This is because the effect of back mutations becomes important as the genetic distance increases. This problem can be studied by using Ohta and Kimura's (1973) stepwise model of neutral mutations, though some authors (*e.g.* Johnson, 1974) disagree about the appropriateness of this model to electrophoretic data. Nei and Chakraborty (1973), Li (1976a), and Chakraborty and Nei (1976, 1977) have already studied the expected genetic identity under the stepwise mutation model. The exact formula for the genetic identity for electrophoretic data (I_E) obtained by Li (1976a) is rather complicated, but Chakraborty and Nei (1977) have shown that for practical purposes Nei and Chakraborty's (1973) earlier formula can be used unless the average heterozygosity in the population is extremely high. Nei and Chakraborty's formula for genetic identity of the t-th generation can be rewritten as

$$I_E=e^{-2vt}\sum_{r=0}^{\infty}(vt)^{2r}/(r!)^2, \tag{17}$$

where v is the mutation rate per generation. We note that $\alpha=v$ in this case since we are dealing with neutral mutations.

When v varies from locus to locus following the gamma distribution, the average value of I_E is given by

$$\bar{I}_E=\frac{b^a}{\Gamma(a)}\sum_{r=0}^{\infty}\frac{t^{2r}}{(r!)^2}\frac{\Gamma(a+2r)}{(b+2t)^{a+2r}}. \tag{18}$$

At the present time, we do not know very well about the a value for the stepwise mutation model. However, if we use $a=1$ as before, we have

$$\bar{I}_E=\frac{1}{1+2\bar{v}t}\left[1+\sum_{r=0}^{\infty}\frac{(2r)!}{(r!)^2}\left(\frac{\bar{v}t}{1+2\bar{v}t}\right)^{2r}\right]. \tag{19}$$

Table 1 shows the values of genetic identity for the four different models, *i.e.*, formulae (11), (14), (17) and (19). In this table calendar year rather than generation is used as a unit of time with $\alpha=10^{-7}$. It is clear that the genetic identity is virtually the same for all four models for the first one million years. Therefore, if the observed value of I is larger than about 0.82, formula (16) may be used for estimating diver-

gence time. However, if the divergence time increases further, the difference among the models becomes pronounced. In this case formula (16) should not be used for estimating divergence time since the assumption of the same mutation rate for all loci is certainly incorrect. The formula for I_E is also expected to give an underestimate, since in this case too the same mutation rate for all loci is assumed. Therefore, for estimating divergence time, formula (14) or (19) should be more appropriate. At any rate, the numerical values in Table 1 can be used for knowing a rough estimate of divergence time if the genetic identity value is available.

Complete isolation: Short-term evolution

In general the above theory does not apply to nonprotein data such as those for blood groups, since the relationship between the codon substitution in a gene and the phenotypic change may not be so simple as that for protein loci (Nei, 1975). However, if we consider a very short period of evolutionary time, all of our measures of genetic distance are approximately linearly related to evolutionary time. In this case we can neglect the effect of mutation. In the absence of selection, the values of J_X, J_Y, and J_{XY} in generation t ($J_X(t)$, $J_Y(t)$, and $J_{XY}(t)$, respectively) can be written as

$$J_X(t)=J_Y(t)=1-(1-J(0))\,(1-\frac{1}{2N})^t \approx J(0)+(1-J(0))t/(2N)$$

$$J_{XY}(t)=J_{XY}(0)=J_X(0)=J_Y(0)=J(0) \tag{20}$$

where $t \ll 2N$ is assumed (see Nei, 1975, p. 124). Therefore, we have

$$D_M=(1-J(0))t/(2N) \tag{21}$$

$$D \approx \frac{1-J(0)}{J(0)}\,\frac{t}{2N} \tag{21b}$$

$$D_v=\frac{1-J(0)}{J(0)}\,\frac{t}{2N}. \tag{21c}$$

Thus, as long as $t \ll 2N$, our distance measures can be used even for nonprotein loci. This seems to be true whether there is selection or not (Chakraborty *et al.*, 1977). In most human populations $t \ll 2N$ appears to hold.

In this connection it should be noted that the quantity that has a simple relationship with evolutionary time is the second moment of gene frequency (Wright, 1931), and thus the genetic distance defined as a geometric distance in a multidimensional space is not proportional to evolutionary time. In my view the linear relationship with evolutionary time is one of the most important properties any genetic distance measure should have. Incidentally, Cavalli-Sforza's (1969) measure f_θ has this property when $t \ll 2N$, but Cavalli-Sforza and Edwards' (1967) d does not.

In our mathematical formulation we assumed that the two populations in question have remained in equilibrium with respect to the effects of mutation, selection, and random genetic drift. This assumption, however, may not always

be satisfied. In fact, there are many cases in which one or both of the populations have gone through bottlenecks. The bottleneck effect on genetic distance has been studied by Chakraborty and Nei (1974; 1977). They have shown that the genetic distance increases rapidly in the presence of bottleneck, and the rate of increase is higher when the bottleneck size is small than when this is large. However, if the population size returns to the original level, the bottleneck effect gradually disappears, though it takes a long time for the effect to disappear completely.

Under certain circumstances it is possible to make a correction for the bottleneck effect. In the case where only one of the two populations has gone through a bottleneck, the following genetic identity may be computed.

$$I=J_{XY}/J_X, \tag{22}$$

where J_X is the mean homozygosity (gene identity) for the population whose size has remained constant. If we use this I in (2) or (5), then D or D_v is linearly related to evolutionary time under the infinite-allele model (Chakraborty and Nei, 1974). In the case where both populations go through bottlenecks, a similar correction can be made if there is a third population the size of which is known to have remained more or less the same as that of the foundation stock of the two populations to be compared. In this case, I may be computed by replacing J_X in (22) by the mean homozygosity for the third population.

Effects of migration

In the early stage of population differentiation gene migration usually occurs between populations. Migration retards gene differentiation considerably, and even a small amount of migration is sufficient to prevent any appreciable gene differentiation unless there is strong differential selection. The effect of migration on genetic distance has been studied by Nei and Feldman (1972), Chakraborty and Nei (1974), Slatkin and Maruyama (1975), and Li (1976b) under the assumption of no selection. Their main conclusions are as follows: (1) If there is a constant rate of migration in every generation, the genetic identity (I) eventually reaches a steady-state value, which is given by

$$I=(m_1+m_2)/(m_1+m_2+2\alpha) \tag{23}$$

approximately, if $2\alpha \ll m_1+m_2 \ll 1$. Here, α is the rate of gene substitution per locus per generation and m_1 and m_2 stand for the migration rates between two populations (m_1 and m_2 may not be the same if the sizes of the two populations are not equal). (2) The approach to the steady state value is generally very slow; the number of generations required is of the order of the reciprocal of mutation rate. Formula (23) indicates that the genetic distance between populations cannot be large unless migration rates are very small.

In the presence of migration it has been customary to study the genetic differentiation of populations in terms of Wright's (1943, 1951) F_{ST} or Malécot's (1948, 1950) kinship coefficient. It is now possible to relate these quantities to the expected

number of codon differences per gene, if the breeding pattern of the population is known (Nei, 1972, 1973b, 1975). For Wright's or Malécot's steady-state formulae to be applicable, however, the breeding pattern of the population must remain the same for a long time. Furthermore, when the number of subpopulations is small, F_{ST} or Malécot's kinship coefficient is subject to a large stochastic variance (Nei *et al.*, 1977; Nei and Chakravarti, 1977; Maruyama, 1977).

GENE DIFFERENCES BETWEEN HUMAN RACES

Let us now consider how the above method can be applied to the study of gene differences between human races. Nei and Roychoudhury (1972, 1974b) studied the genic variation within and between three major races of man, *i.e.*, Caucasoids, Negroids, and Mongoloids (mostly Japanese). Surveying the literature, they collected gene frequency data of 74 protein loci for Caucasoids, 62 loci for Negroids (largely American Negroids), and 35 loci for Mongoloids. They also collected 57 blood group loci for Caucasoids, 34 for Negroids, and 21 loci for Mongoloids. As emphasized by Lewontin (1967), blood group loci are discovered only when there is polymorphism, and thus the genetic distance as well as average heterozygosity per locus tend to be an overestimate unless a large number of loci are studied. Despite this disadvantage, we used blood group data, since they are still useful for knowing the relative distances for different pairs of populations.

Intraracial and interracial genetic variations

Table 2 shows the average heterozygosities per locus for the three major races. It is clear from Table 2 that the average heterozygosity for protein loci is about 10 percent and essentially the same for all three populations. This value is close to the estimate of average heterozygosity obtained by Harris (1969) for 20 randomly

Table 2. Proportion of polymorphic loci and average heterozygosity for protein loci in the three major races of man: Caucasoids, Negroids (mostly American blacks) and Mongoloids (mostly Japanese).

Major race	Number of loci	Proportion of polymorphic loci[a]	Average heterozygosity
Protein loci			
Caucasoids	74	.31	.099±.021
Negroids	62	.40	.092±.019
Mongoloids[a]	35	.40	.098±.027
Blood group loci			
Caucasoids	57	.37	.130±.027
Negroids	34	.56	.162±.035
Mongoloids[a]	22	.71	.231±.049

[a] A locus is defined as polymorphic if the frequency of the commonest allele in the population is less than 0.99.

chosen protein loci in Caucasoids. Therefore, the loci used in this study are considered to be close to a random sample of the genome. All our protein data were obtained by electrophoresis. Since electrophoresis can detect about 1/4 of amino acid substitutions, our result suggests that the average heterozygosity at the codon level is 0.4 per locus. If we note that only about 3/4 of nucleotide substitutions result in amino acid substitution, the heterozygosity at the nucleotide level is expected to be 0.53 per locus. Namely, in an average person 53 percent of the structural loci are expected to be heterozygous at the nucleotide level. If there are 30,000 structural loci in the human genome, this means that an average person is heterozygous at about 15,000 loci. If we note that the genetic variability at the third nucleotide position in a codon is larger than the first two (Kafatos *et al.*, 1977), the actual proportion of heterozygous loci is expected to be even higher.

The average heterozygosity for blood group loci is somewhat higher than that for protein loci. However, we cannot convert this value into the proportion of heterozygous loci at the nucleotide level, since the theoretical relationship between these is not known. The average heterozygosity for Mongoloids is higher than that for Negroids, which is in turn higher than that for Caucasoids. This seems to reflect that the blood group loci used here are not random samples of all blood group loci. Since blood group loci are discovered only when there is polymorphism, the observed average heterozygosity is expected to be high when the number of loci used is small (Nei and Roychoudhury, 1974b).

Table 3 shows the estimates of genetic distances obtained from 35 common protein and 21 common blood group loci that are shared by all races. In the case of protein loci, the genetic distance between Negroids and Mongoloids is slightly higher than that between Caucasoids and Negroids. On the other hand, the distance between Caucasoids and Mongoloids is the smallest among the three pairs of groups. This suggests that Caucasoids and Mongoloids are more closely related to each other than to Negroids. The results from blood group loci are somewhat inconsistent with this conclusion. That is, Caucasoids and Negroids are more closely related than Caucasoids and Mongoloids, while Negroids and Mongoloids are least related. This is partly due to the fact that all blood group gene frequency data for Negroids come from American Negroids who have had racial admixture

Table 3. Estimates of minimum, standard and maximum genetic distances between Caucasoid, Negroid, and Mongoloid populations.

	Caucasoid/Negroid		Caucasoid/Mongoloid		Negroid/Mongoloid	
	Proteins (35 Loci)	Blood groups (21 Loci)	Proteins (35 Loci)	Blood groups (21 Loci)	Proteins (35 Loci)	Blood groups (21 Loci)
Minimum	.014±.006	.021±.008	.010±.004	.025±.009	.017±.008	.070±.034
Standard	.017±.007	.027±.012	.011±.005	.034±.014	.019±.009	.095±.049
Maximum	.021±.010	.031±.012	.012±.005	.043±.016	.026±.013	.144±.075

with Caucasoids, whereas 24 percent of protein data come from African Negroids rather than American Negroids.

The estimates of genetic distance from blood group loci are considerably larger than those from protein loci. This apparently reflects nonrandom sampling of blood group loci when the number of loci is small, as discussed earlier. The genetic distances between Negroids and Mongoloids are about three times larger than the values between the other pairs of races. The large value is mainly due to the Duffy locus, in which the single-locus distance ($d=(j_X+j_Y)/2-j_{XY}$) is .63. The gene frequency of fy^a is .05 in Negroids and .84 in Mongoloids. If this locus is excluded from the data, the minimum and maximum distances become .041 and .077, respectively.

Earlier I mentioned that the average minimum codon differences between two genes that are randomly chosen from the same population is equal to the average heterozygosity. Therefore, the ratio of D_m to $(D_X+D_Y)/2$ gives an idea about the extent of genetic divergence of populations relative to the within-population variation. This ratio is 0.08–0.15 for protein loci and 0.09–0.3 for blood group loci. Therefore, the interracial variation is small compared with the intraracial variation (Nei and Roychoudhury, 1972). The decomposition of the total variation into the intrapopulational components can also be made by a more elaborate statistical method (analysis of gene diversity) developed by Nei (1973b, 1977b). Application of this method to the present protein data shows that the proportion of interracial gene diversity among the total variation is only 7 percent (Nei, 1975). Using a different method, Lewontin (1972) obtained a similar result with respect to blood group loci.

Molecular evolution vs. morphological evolution

The above results clearly indicate that at the molecular level the genetic variation among the three major races of man is much smaller than that within the races. This is so despite the fact that the phenotypic differences such as pigmentation and facial structure among these groups are conspicuous. We note that the distributions of these characters for the three major races are virtually nonoverlapping, whereas the differences in protein loci are essentially due to gene frequency shifts. To explain this difference between the evolutionary changes of proteins and morphological characters, Nei and Roychoudhury (1972) hypothesized that the genes controlling morphological characters have been subject to stronger natural selection than the genes for general protein loci. Recently, a conspicuous difference between molecular evolution and morphological evolution was also observed in the comparison of man and chimpanzee (King and Wilson, 1975). In the evolution of these organisms a much faster divergence seems to have occurred in certain morphological characters than in general protein loci. Their explanation for this difference was somewhat different from ours, though they are not mutually exclusive. They proposed the hypothesis that the evolutionary change of morphological charac-

ters occurs mostly by mutations at regulatory gene loci which are supposed to have profound effect on morphological characters, whereas gene substitution at structural gene loci do not affect morphological characters very much. This hypothesis is attractive, but at the present time it is not clear how to distinguish regulatory genes from structural genes. King and Wilson have not defined regulatory genes precisely, but their regulatory genes seem to refer to any genes that affect transcription or translation of other genes, whether they are actually structural genes or not (Wilson *et al.*, 1977). This creates a difficulty in identifying regulatory genes in experimental studies. Furthermore, it is quite likely that a large part of the genome in higher organisms function both as structural genes and regulatory genes in the process of morphogenesis. This is because morphogenesis is so well coordinated that production of a protein or enzyme would often stimulate or restrict the transcription or translation of other genes, as is probably the case with protein hormones. If this is so, it would not be very meaningful to distinguish between "regulatory genes" and "nonregulatory genes." It seems to me that the important thing in the study of evolutionary change of morphological characters is to identify the genes that control morphological characters at the molecular level and find out the relationship between nucleotide substitution and morphological change. However, Nei's (Nei and Roychoudhury, 1972; Nei, 1975) hypothesis and King and Wilson's have one common feature. Namely, the rapid evolutionary change of a morphological character may be brought about by a small number of gene substitutions.

Comparison with other organisms

Estimates of genetic distance between various ranks of taxa such as races, subspecies, species, and genera have been obtained in many different organisms. Table 4 gives a summary of these estimates (Nei, 1975). It is clear that the magnitude of genetic distance between local races is essentially the same for all organisms and similar to those for human races. Namely, in terms of genetic distance human races are equivalent to local races in other organisms (Nei and Roychoudhury, 1972). Table 4 shows that the genetic distance value generally increases as the rank of taxa to be compared becomes higher, as expected. One notable exception is the genetic distance between man and chimpanzee. These two organisms belong to different families according to the present classification. Yet the standard genetic distance is only 0.62 (King and Wilson, 1975), which corresponds to genetic distances between different species in the same genus in other organisms. The closeness between man and chimpanzee has also been indicated by immunological studies (Goodman, 1961; Sarich and Wilson, 1967) and comparisons of amino acid sequences of some proteins (Wilson and Sarich, 1969; King and Wilson, 1975). Therefore, it seems to be certain that man and chimpanzee are genetically much closer than the current taxonomy suggests.

As mentioned earlier, the unit of Nei's standard genetic distance is the number

Table 4. Estimates of genetic distance from electrophoretic data in various organisms. Adapted from Nei (1975).

	Taxa	No. of taxa	No. of loci	$\hat{D}=-ln\,\hat{I}$
A.	Local races			
	Man	3	35	.011– .019
	Rodents	13	18–41	.000– .058
	Drosophila	12	11–24	.001– .010
B.	Subspecies			
	Rodents	16	27–41	.004– .262
	Lizards	4	23	.335– .351
	Fish	9	17	.062– .218
	Drosophila	11	12–25	.028– .234
C.	Species			
	Mammals	7	14–27	.12– .63
	Lizards	4	23	1.32–1.75
	Drosophila	45	13–28	.05–2.54
D.	Genera			
	Fish	5	16	1.1–2.8(∞)
E.	Man-Chimp (Families)		42	.62
F.	Man-Horse (Orders)			(18)[a]

[a] This was estimated from amino acid sequence data (Nei, 1975).

of codon substitutions per gene that are detectable by the technique used. Electrophoresis is expected to detect only about 1/4 of codon differences between homologous proteins. Therefore, a rough estimate of codon differences may be obtained by multiplying the genetic distance values in Table 3 by 4 if the D value is relatively small. When D is large or I is small, the number of codon differences may be estimated by using formulae (14) and (19) (see also Table 1).

Gene differences between relatives

In this connection one might ask the question: what is the mean number of codon differences per locus between an individual of man and his various relatives? The number of codon differences per locus between two genomes, one from each of two "unrelated individuals" in a population, may be estimated by the average heterozygosity, *i.e.* $1-J$. When J is not close to 1, however,

$$D_X=(1-J)/J \tag{24}$$

is a better estimate. The rationale for this is as follows: Consider a cistron composed of n codons, and let δ_i be the probability that the i-th codon is different between two randomly chosen cistrons. If δ_i is independent of δ_i for any pair of i and j, the probability that two randomly chosen cistrons have an identical codon sequence is $P=\prod_{i=1}^{n}(1-\delta_i)\approx e^{-\Sigma\delta_i}=e^{-D_c}$, where D_c is the mean number of codon differences per cistron (Kimura, 1969). Let us now assume that D_c varies with locus following the gamma distribution with coefficient of variation one. Then, the mean

Table 5. Rough estimates of the number of codon differences (substitutions) per 100 structural genes (about 40,000 codons) between two randomly chosen genomes. In these computations electrophoresis was assumed to detect only one quarter of codon differences, and no consideration was given to synonymous codon differences.

Two genomes taken from	$\hat{J}$ or $\hat{J}_{XY}$[b]	No. of codon differences per 100 cistrons
Contribution due to new mutation		0.0024
First degree relatives in man[a]		33
Second degree relatives in man[a]		39
Unrelateds within races[a]	0.900	44
Unrelateds in Yanomama Indians	0.961	16
Between races[a]	0.880	49
Man and Chimpanzee	0.511	383
Man and Horse		7,200

[a] Caucasoids, Negroids, and Japanese.

[b] All of these values were obtained by electrophoresis.

of P over loci ($\bar{P}$) is $1/(1+\bar{D}_c)$, where $\bar{D}_c$ is the mean of D_c over loci. Therefore, $\bar{D}_c$ may be estimated by equating $\bar{P}$ to J, *i.e.*, by (24). When D_X is obtained from electrophoretic data, it should be multiplied by 4. On the other hand, the mean number of codon differences per locus (D_F) between relatives with kinship coefficient F may be obtained by

$$D_F=(1-F)D_X. \tag{25}$$

Some results obtained by this formula are given in Table 5. In this table, estimates of the mean numbers of codon differences between man and other organisms are also presented. The value between different races or between man and chimpanzee was estimated by $D_{XY}=(1-J_{XY})/J_{XY}$, where J_{XY} is the average gene identity between two randomly chosen genomes, one from each of the two populations under consideration, whereas the value between man and horse was obtained from amino acid sequence data.

It is clear that the mean number of codon differences per locus is larger for remote relatives than for close relatives. For example, the mean number of codon differences between man and chimpanzee is about 10 times larger than that between second-degree relatives in Caucasians, Negroes, and Japanese, but is about 1/19 of that between man and horse.

In this connection it should be noted that the mean number of codon differences per locus for a given degree of relatives is not the same for all human races. Some races such as the Yanomama Indians in South America have a lower average heterozygosity than Caucasians or Japanese. This means that unrelated individuals in the former are genetically more similar than those in the latter. In fact, if we use data for 15 protein loci obtained by Weitkamp *et al.* (1972) and Weitkamp and Neel (1972) in Yanomama Indians, the D_X value becomes 16 per 100 genes. This

value is smaller than that for first-degree relatives in Caucasians or Japanese. In other words, two unrelated individuals in Yanomama Indians are genetically closer with each other than two first-degree relatives in Caucasians or Japanese.

Table 5 also gives an estimate of codon differences due to new mutations ($2v$ per generation). This estimate was obtained by considering the number of rare alleles maintained by mutation, selection, and genetic drift (Nei, 1977c), and refers to the differences at the amino acid level. It is about 1/18,000 of the value for unrelateds in Caucasians or Japanese.

Evolutionary time

For many years anthropologists have asked the question: when did the three major races of man diverge from each other. Coon (1962) suggests that the divergence of these races occurred about 500,000 years ago when man was not yet *Homo sapiens* but *Homo erectus*. He speculates that *Homo erectus* evolved into *Homo sapiens* independently in each of his five human "subspecies," Caucasoid, Negroid, Mongoloid, Australoid, and Capoid. This speculation is mainly based on the evolutionary change of dentition. It is, however, quite unlikely that the same evolutionary change occurred in five "subspecies" independently. On the other hand, Cavalli-Sforza and Bodmer (1971) have suggested, on the basis of their study of the differentiation of blood group gene frequencies and ecological conditions in the Pleistocene, that the divergence between Negroid and Mongoloid occurred about 50,000 years ago and some time later but probably before 20,000 years ago Caucasoid was formed either from Mongoloid or Negroid or both.

Our data in Table 3 can be used to answer this question. Namely, using formula (16), we can estimate the time of divergence between two populations from genetic distance data. However, some caution is necessary, since the genetic distance estimates involving Negroid in Table 3 are largely based on gene frequency data from American Negroes rather than African Negroes (Nei and Roychoudhury, 1974b) and 20 percent of the genome of American Negroes had been derived from Caucasians (Reed, 1969). The correction for the effect of this type of gene migration on genetic distance can be made by the method of Nei (Nei, 1974; Nei and Roychoudhury, 1974b). At any rate, the corrected estimates of genetic distance and the estimates of the time after divergence between these populations are pre-

Table 6. Estimates of standard genetic distance and divergence time between the three major races of man after correction for migration of Caucasoid genes into Negroid gene pool. 35 common protein loci were used.

Comparison	Standard distance	Divergence time (yrs)
Caucasoid/Negroid	.023	115,000
Caucasoid/Mongoloid	.011	55,000
Negroid/Mongoloid	.024	120,000

sented in Table 6. The estimates of divergence time given in this table are considered to be minimal, since in the early stage of population differentiation there must have been some migration. Therefore, the three major races of man appear to have been isolated for at least 50–100 thousand years (Nei and Roychoudhury, 1974b). Table 6 also suggests that among the three racial groups the first divergence occurred about 120,000 years ago between Negroids and the group of Caucasoids and Mongoloids and then the latter group split into two around 60,000 years ago.

As mentioned earlier, man and chimpanzee are genetically much closer than their morphological characters suggest. Studying the immunological distance between these two organisms, Sarich and Wilson (1966, 1967) suggested that they were separated about 4–5 million years ago in contrast with the view of many anthropologists, who believed at that time that the separation occurred about 30 million years ago. This generated a great deal of controversy among anthropologists, which is still going on. It is therefore interesting to apply our method to this problem. King and Wilson (1975) studied the genetic distance between man and chimpanzee and obtained $\hat{D} \equiv -ln\ \hat{I} = 0.62$. This corresponds to $\hat{I} = 0.538$. The values of $\bar{I}_A$ and $\bar{I}_E$ in Table 1 therefore suggest that the divergence time is 4–6 million years, which agrees with Sarich and Wilson's estimate. Of course, our estimate has a large standard error, so that much reliance cannot be given at the present time. It is desirable to use a large number of loci for studying this problem.

Recently a number of authors examined whether "genetic distance clock" works well or not for estimating evolutionary time. For example, Gorman and Kim (1977) studied the genetic distance between two species of fish, *Abudefduf saxatilis* and *A. troschelii*, from the Atlantic and Pacific coasts of the Panama Isthmus, using 28 protein loci. These two species were apparently formed when the Panama Isthmus separated the Atlantic and Pacific Oceans about 2–5 million years ago. Their genetic distance estimate is $\hat{D} = 0.32$, from which we obtain $\hat{I} = 0.726$. From the values of $\bar{I}_A$ and $\bar{I}_E$ in Table 1, we see that this $\hat{I}$ value corresponds to a divergence time of about 2 million years. Therefore, there is a good agreement between theory and data. In general, the "genetic distance clock" seems to work relatively well as a rough approximation when it is applied properly (Nei, unpublished). Sarich (1977) also has shown that there is a high correlation between genetic distance and albumin immunological distance, which is supposed to be linearly related with evolutionary time. It should be noted, however, that the standard error of the time estimate is generally very large.

PHYLOGENY OF HUMAN POPULATIONS

Methods of constructing a phylogenetic tree

The phylogeny of human populations based on genetic distance measure was first constructed by Cavalli-Sforza and Edwards (1964). Since then a large number of authors have applied their method to study the phylogeny of various groups of

populations (*e.g.* Cavalli-Sforza, 1966; Fitch and Neel, 1969; Imaizumi *et al.*, 1973). In these studies, however, the number of loci used was rather small, and little attention was given to the magnitude of errors introduced by stochastic changes of gene frequencies. For example, Cavalli-Sforza and Edwards (1964) used gene frequency data for only five blood group loci. The necessity of a large number of loci for constructing a phylogenetic tree was later pointed out by Kidd and Cavalli-Sforza (1971) in a computer simulation study and by Nei and Roychoudhury (1974a) and Li and Nei (1975) in their theoretical studies. In practice, however, the number of loci at which gene frequencies have been studied for many human races is quite limited at the present time. Therefore, the phylogenetic tree constructed should be regarded to be still tentative.

In the previous section we emphasized the importance of using randomly chosen loci to estimate genetic distance, including both polymorphic and monomorphic loci. For the purpose of constructing a phylogenetic tree, however, one can use only polymorphic loci, if the magnitudes of genetic distances are small as those for human populations. Furthermore, if we consider a relatively short evolutionary time, we can use not only protein loci but also nonprotein loci, since in this case the effect of mutation is negligible and the genetic distance is still approximately linear with evolutionary time.

There are several methods of constructing a phylogenetic tree from genetic distance data. The simplest one is the so-called unweighted pair-group method (UPG) developed by Sokal and Sneath (1963) with the additional assumption of constant rate of evolution (Nei, 1975). Our computer simulation (Tateno and Nei, unpublished) has shown that despite its simplicity, it generally gives a quite reasonable tree. Chakraborty (1977) has shown that this method gives the least-squares estimates of branch lengths if the topology of the tree is known. Ideally, however, it is desirable to compute the average squared deviations of estimated distances from observed distances (or Fitch and Margoliash's (1967) percent standard deviation) for all possible trees and choose the one that minimizes the averaged squared deviation. Unfortunately, the number of possible trees is so large even for a relatively small number of populations used, that it is virtually impossible to compute average squared deviations for all possible trees (Cavalli-Sforza and Edwards, 1967). Because of this difficulty, Fitch and Margoliash (1967) devised an ad hoc method by which about 40 different trees are usually tested.

Farris' (1972) method is not based on the principle of minimizing the average squared deviation but on the principle of minimum evolutionary paths. However, this method seems to be quite efficient in obtaining a correct topology of a tree *when the rate of gene substitution is constant.* Our computer simulation has shown that the rate of recovery of the true tree is considerably higher for this method than in the UPG method or Fitch and Margoliash's method. For estimating the branch lengths of a topology, however, the Farris' method does not seem to be very good (Prager and Wilson, 1978). In fact, our simulation has shown that the average

squared deviation of estimated distances from the true (known) branch lengths is larger in this method than in Fitch and Margoliash's method or the UPG method. Recently, Tateno and Nei (unpublished) improved some aspects of Farris' method and showed that the modified Farris' method gives a tree of which the average squared deviation is smaller than Farris' method. However, when the rate of gene substitution fluctuates from time to time or when there is migration between populations, these methods often give a quite erroneous tree. In this respect Fitch and Margoliash's (1967) method seems to be better than the Farris' method and its modification.

Although our work on the methodology of making a phylogenetic tree has not been completed, it seems to us that for obtaining the topology of a tree Fitch and Margoliash's method is quite efficient. However, once the topology is determined, the branch lengths of the tree should be determined by Nei's (1975) method of averaging distances. This is justified because the time after divergence is necessarily the same for any pair of populations. Furthermore, application of Nei's method often eliminates negative branches existing in Fitch and Margoliash's tree.

Data analysis

In the last decade a large amount of gene frequency data has been accumulated for many newly discovered protein loci. In collaboration with Drs. Paul Fuerst and Shozo Yokoyama I have surveyed the literature and collected gene frequency data for many different races with the aim of constructing a phylogenetic tree of human races. At the present time, however, the number of loci at which gene frequency data are available rapidly declines as the number of races included increases. When the twelve races listed in Table 7 are considered, there are 11 protein loci (AK, ADA, Acp_1, Gc, Tf, $Hb\alpha$, $Hb\beta$, Hp, PGM_1, PGM_2, 6PGD) and 11 blood group loci (ABO, MNSs, P, Lutheran, Kell, Secretor, Duffy, Kidd, Diego, Wright,

Table 7. Estimates of standard genetic distances between twelve races of man. Eleven protein loci and eleven blood group loci, the majority of which were polymorphic, were used.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
English (1)											
Italians (2)	.0008										
Indians (3)	.0106	.0111									
Japanese (4)	.0326	.0319	.0204								
Chinese (5)	.0427	.0427	.0292	.0047							
Ainu (6)	.0476	.0494	.0300	.0068	.0087						
New Guineans (7)	.0594	.0575	.0631	.0388	.0347	.0531					
Micronesians (8)	.0343	.0324	.0297	.0068	.0086	.0228	.0228				
North Amerinds (9)	.0291	.0280	.0411	.0377	.0462	.0534	.0534	.0355			
South Amerinds (10)	.0524	.0502	.0664	.0654	.0647	.0445	.0445	.0448	.0292		
Ghana (11)	.0418	.0427	.0628	.1023	.1251	.1384	.1285	.0951	.0778	.0722	
Bantu (12)	.0393	.0409	.0614	.0992	.1228	.1344	.1221	.0914	.0760	.0730	.0048

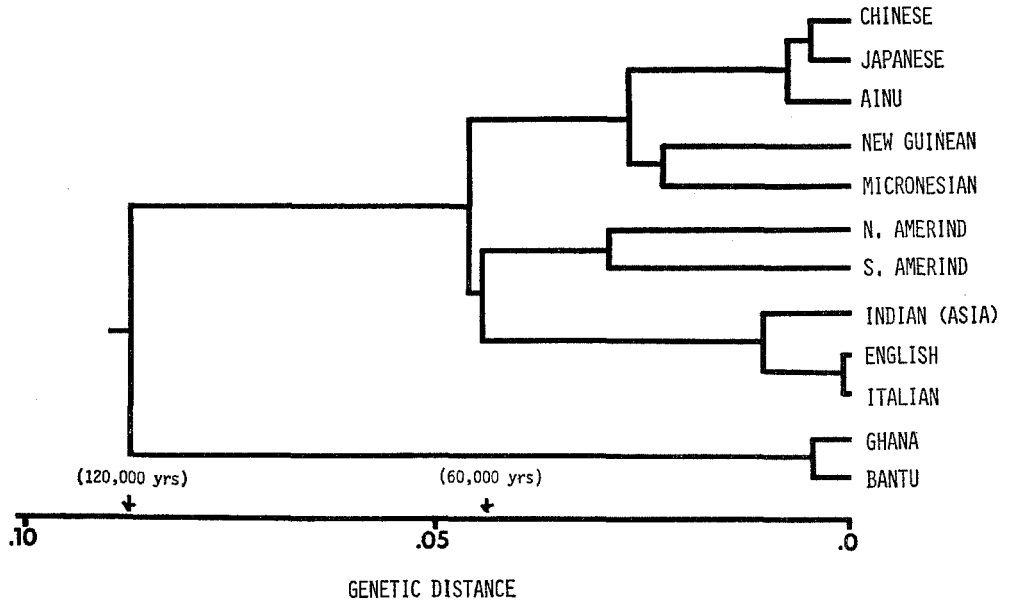

Fig. 1. The phylogenetic tree for 12 human races. This tree is based on gene frequency data for 11 protein and 11 blood group loci.

Rh) available. The genetic distance estimates obtained from these gene frequency data are presented in Table 7. Using these estimates, I constructed a phylogenetic tree, which is given in Fig. 1. In the construction of this tree I used Fitch and Margoliash's method for determining the topology and Nei's method for determining the branch lengths.

This tree roughly agrees with what we expect intuitively from the morphological characters and historical records of these races. The pattern of divergence of Caucasoids, Negroids, and Japanese is essentially the same as that given by Nei and Roychoudhury (1974b) (see also Table 6). The races which are considered to belong to the same racial group are generally clustered together. In general, this tree agrees much better with what we intuitively believe than Cavalli-Sforza's (1966) tree. However, there are some anomalies. For example, American Indians are closer to Caucasoids than to Mongoloids, despite the fact that they apparently diverged from Mongoloids and moved to the American Continents through the Bering Strait about 30,000 years ago. This anomaly seems to have occurred partly because American Indians have received Caucasian genes to a considerable extent. Another possible reason is that American Indians have had a rather small effective population size, which increases both the mean and variance of genetic distance.

The Ainu of Hokkaido, Japan, are often called a racial isolate, since they have different physical characteristics compared with their neighboring populations. Figure 1, however, shows that they are closer to Japanese and Chinese than to any other race. This supports Omoto's (1975) conclusion that the Ainu are not really

a "racial island." Of course, the closeness of Ainu to Japanese in Fig. 1 is partly due to the recent admixture between these two races. Geneological data suggests that about 40 percent of the genomes of the present Ainu come from Japanese. However, Omoto's (1975) phylogenetic tree constructed with the correction for this factor also shows that Ainu are closer to Japanese and Chinese than to other races.

Some further comments should be made on the effect of gene migration in the phylogenetic tree in Fig. 1. In this tree Indians are clustered with the English and Italians. Intuitively, this appears to be reasonable, since Indians are supposed to belong to the Caucasoid group. However, genetic distance values in Table 7 suggest that Indians are also rather close to Japanese and Chinese compared with the relationship between the English-Italian group and the Japanese-Chinese group. Namely, Indians are somewhere between the two groups, and this is apparently caused by gene admixture between neighboring populations. In this respect, the phylogeny in Fig. 1 is somewhat misleading. It is noted that in Roychoudhury's (1977) recent phylogenetic tree based on 29 protein loci (16 polymorphic loci) Indians were first clustered with the Japanese and then with the English. Table 7 also shows that the English and Italians are related to both Mongoloids and Negroids probably because of gene migration. From these observations, it is clear that in order to make a reliable tree a proper consideration should be given to the effect of migration. Unfortunately, there is no objective method to treat this effect at the present time.

Recently a number of authors have produced a phylogenetic tree or dendrogram for populations within the same race. In this case the effect of migration is more important, and the dendrogram constructed does not necessarily represent the historical pathways of the populations. Nevertheless, such a dendrogram seems to be useful for understanding the genetic relationship of the populations.

As mentioned earlier, the phylogenetic trees of human races constructed at the present time are far from complete. Fortunately, however, gene frequency data are rapidly accumulating. In the near future we will be able to obtain a much better picture of the biological history of the present human races. It may also be possible for us to predict or even guide the evolutionary course of our own species in the future.

Acknowledgments I wish to thank all of my colleagues who have cooperated with me in the study of genetic distance and genetic differentiation of human races. They include Drs. A.K. Roychoudhury, R. Chakraborty, W.-H. Li, Y. Tateno, A. Chakravarti, P.A. Fuerst, and S. Yokoyama. This study was supported by NIH grants GM 19513 and 20293 and NSF DEB 76-06069.

REFERENCES

Ayala, F.J. ed. 1976. *Molecular Evolution.* Sinauer Assoc. Inc., Sunderland, Massachusetts.

Balakrishnan, V. and Sanghvi, L.D. 1968. Distance between populations on the basis of attribute data. *Biometrics* **24**: 859–865.

Bhattacharyya, A. 1946. On a measure of divergence between two multinomial populations. *Sankhya* **7**: 401–406.

Carlson, S.S., Wilson, A.C., and Maxon, R.D. 1978. Reply to L. Radinsky's comment: Do albumin clocks run on time? *Science* **200**: 1183–1184.

Cavalli-Sforza, L.L. 1966. Population structure and human evolution. *Proc. R. Soc. London Ser. B* **164**: 362–379.

Cavalli-Sforza, L.L. 1969. Human diversity. *Proc. 12th Intl. Congr. Genet. (Tokyo)* **3**: 405–416.

Cavalli-Sforza, L.L. and Bodmer, W.F. 1971. *The Genetics of Human Populations.* Freeman, San Francisco.

Cavalli-Sforza, L.L. and Edwards, A.W.F. 1964. Analysis of human evolution. In: *Genetics today, Proc. 11th Intl. Cong. Genet., The Hague.* Pergamon Press, Oxford, pp. 923–933.

Cavalli-Sforza, L.L. and Edwards, A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. *Am. J. Hum. Genet.* **19**: 233–257.

Chakraborty, R. 1977. Estimation of time of divergence from phylogenetic studies. *Can. J. Cytol. Genet.* **19**: 217–223.

Chakraborty, R., Fuerst, P.A., and Nei, M. 1977. A comparative study of genetic variation within and between populations under the neutral mutation hypothesis and the model of sequentially advantageous mutation. *Genetics* **86**: s10–s11.

Chakraborty, R., Fuerst, P.A., and Nei, M. 1978. Statistical studies on protein polymorphism in natural populations. II. Gene differentiation between populations. *Genetics* **88**: 367–390.

Chakraborty, R. and Nei, M. 1974. Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. *Theor. Popul. Biol.* **5**: 460–469.

Chakraborty, R. and Nei, M. 1976. Hidden genetic variability within electromorphs in finite populations. *Genetics* **84**: 385–393.

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

Coon, C.S. 1962. *The Origin of Races.* Alfred A. Knopf, New York.

Dayhoff, M.O. ed. 1969. *Atlas of Protein Sequence and Structure,* Vol. 4. Natl. Biomed. Res. Found., Washington, D.C.

Farris, J.S. 1972. Estimating phylogenetic trees from distance matrices. *Am. Nat.* **106**: 645–668.

Fitch, W.M. 1976. Molecular evolutionary clocks. In: *Molecular Evolution,* F.J. Ayala, ed., Sinauer Assoc. Inc., Sunderland, Mass., pp. 160–178.

Fitch, W.M. and Margoliash, E. 1967. Construction of phylogenetic trees. *Science* **155**: 279–284.

Fitch, W.M. and Neel, J.V. 1969. The phylogenic relationships of some Indian tribes of Central and South America. *Am. J. Hum. Genet.* **21**: 384–397.

Fuerst, P.A., Chakraborty, R., and Nei, M. 1977. Statistical studies on protein polymorphism in natural populations. I. Distribution of single locus heterozygosity. *Genetics* **86**: 455–483.

Goodman, M. 1961. The role of immunochemical differences in the phyletic development of human behavior. *Human Biology* **33**: 131–162.

Gorman, G.C. and Kim, Y.J. 1977. Genotypic evolution in the face of phenotypic conservativeness: *Abudefduf* (Pomacentridae) from the Atlantic and Pacific sides of Panama. *Copeia* **1977**: 694–697.

Harris, H. 1966. Enzyme polymorphisms in man. *Proc. R. Soc. London Ser. B* **164**: 298–310.

Harris, H. 1969. Enzyme and protein polymorphism in human populations. *Br. Med. Bull.* **25**: 5–13.

Imaizumi, Y., Morton, N.E., and Lalouel, J.M. 1973. Kinship and race. In: *Genetic structure of populations,* N.E. Morton, ed., University of Hawaii, Honolulu, pp. 228–233.

Johnson, G.B. 1974. On the estimation of effective number of alleles from electrophoretic data. *Genetics* **78**: 771–776.

Kafatos, F.C., Efstratiadis, A., Forget, B.G., and Weissman, S.M. 1977. Molecular evolution of

human and rabbit β-globin mRNAs. *Proc. Natl. Acad. Sci.* **74**: 5618–5622.

Kidd, K.K. and Cavalli-Sforza, L.L. 1971. Number of characters examined and error in reconstruction of evolutionary trees. In: *Mathematics in the Archaeological and Historical Sciences*, F.R. Hodson, D.G. Kendall, and P. Tautu, eds., University of Edinburgh Press, Edinburgh, pp. 335–346.

Kimura, M. 1968. Evolutionary rate at the molecular level. *Nature* **217**: 624–626.

Kimura, M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutation. *Genetics* **61**: 893–903.

Kimura, M. and Ohta, T. 1971. *Theoretical Aspects of Population Genetics.* Princeton University Press, Princeton, New Jersey.

King, M. and Wilson, A.C. 1975. Evolution at two levels in humans and Chimpanzees. *Science* **188**: 107–116.

Koehn, R.K. and Eanes, W.F. 1977. Subunit size and genetic variation of enzymes in natural populations of *Drosophila. Theor. Popul. Biol.* **11**: 330–341.

Lewontin, R.C. 1967. An estimate of average heterozygosity in man. *Am. J. Hum. Genet.* **19**: 681–685.

Lewontin, R.C. 1972. The apportionment of human diversity. *Evol. Biol.* **6**: 381–398.

Lewontin, R.C. and Hubby, J.L. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of *Drosophila pseudoobscura. Genetics* **54**: 595–609.

Li, W.-H. 1976a. Electrophoretic identity of proteins in a finite population and genetic distance between taxa. *Genet. Res.* **28**: 119–127.

Li, W.-H. 1976b. Effect of migration on genetic distance. *Am. Natur.* **110**: 841–847.

Li, W.-H. and Nei, M. 1975. Drift variances of heterozygosity and genetic distance in transient states. *Genet. Res.* **25**: 229–248.

Mahalanobis, P.C. 1936. On the generalized distance in statistics. *Proc. Natl. Acad. Sci. India* **2**: 49–55.

Malécot, G. 1948. *Les Mathématiques de l'hérédité.* Masson et Cie, Paris.

Malécot, G. 1950. Quelques schémas probabilistes sur la variabilité des populations naturelles. *Ann. Univ. Lyon, Sci. Sect. A* **13**: 37–60.

Margoliash, E. and Smith, E.L. 1965. Structural and functional aspects of cytochrome c in relation to evolution. In: *Evolving Genes and Proteins*, V. Bryson and H.J. Vogel, eds., Academic Press, New York, pp. 221–242.

Maruyama, T. 1977. *Stochastic Problems in Population Genetics.* Springer-Verlag, Berlin, Heidelberg, New York.

Nei, M. 1971. Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identiy. *Am. Nat.* **105**: 385–398.

Nei, M. 1972. Genetic distance between populations. *Am. Nat.* **106**: 283–292.

Nei, M. 1973a. The theory and estimation of genetic distance. In: *Genetic Structure of Populations*, N.E. Morton, ed., University of Hawaii, Honolulu, pp. 45–54.

Nei, M. 1973b. Analysis of gene diversity in subdivided populations. *Proc. Natl. Acad. Sci.* **70**: 3321–3323.

Nei, M. 1974. A new measure of genetic distance. In: *Genetic Distance*, J.F. Crow and C. Denniston, eds., Plenum Press, New York and London, pp. 63–76.

Nei, M. 1975. *Molecular Population Genetics and Evolution.* North Holland, Amsterdam and New York.

Nei, M. 1977a. Genetic distance. In: *Genetics*, E. Matsunaga and K. Omoto, eds., Yuzankaku Publ., Tokyo, pp. 29–62.

Nei, M. 1977b. F-statistics and analysis of gene diversity in subdivided populations. *Ann. Hum. Genet., London* **41**: 225–233.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* **89**: 583–590.

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

Nei, M., Chakraborty, R., and Fuerst, P.A. 1976a. Infinite allele model with varying mutation rate. *Proc. Natl. Acad. Sci.* **73**: 4164–4168.

Nei, M. and Chakravarti, A. 1977. Drift variances of F_{ST} and G_{ST} statistics obtained from a finite number of isolated populations. *Theor. Popul. Biol.* **11**: 307–325.

Nei, M., Chakravarti, A., and Tateno, Y. 1977. Mean and variance of F_{ST} in a finite number of incompletely isolated populations. *Theor. Popul. Biol.* **11**: 291–306.

Nei, M. and Feldman, M.W. 1972. Identity of genes by descent within and between populations under mutation and migration pressures. *Theor. Popul. Biol.* **3**: 460–465.

Nei, M., Fuerst, P.A., and Chakraborty, R. 1976b. Testing the neutral mutation hypothesis by distribution of single locus heterozygosity. *Nature* **262**: 491–493.

Nei, M., Fuerst, P.A., and Chakraborty, R. 1978. Subunit molecular weight and genetic variability of proteins in natural populations. *Proc. Natl. Acad. Sci.*, **75**: 3359–3362.

Nei, M. and Roychoudhury, A.K. 1972. Gene differences between Caucasian, Negro, and Japanese populations. *Science* **177**: 434–436.

Nei, M. and Roychoudhury, A.K. 1974a. Sampling variances of heterozygosity and genetic distance. *Genetics* **76**: 379–390.

Nei, M. and Roychoudhury, A.K. 1974b. Genic variation within and between the three major races of man, Caucasoids, Negroids, and Mongoloids. *Am. J. Hum. Genet.* **26**: 421–443.

Nei, M. and Tateno, Y. 1975. Interlocus variation of genetic distance and the neutral mutation theory. *Proc. Natl. Acad. Sci.* **72**: 2758–2760.

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

Omoto, K. 1975. Genetic affinities of the Ainu as assessed from data on polymorphic traits. In: *Anthropological and Genetic Studies on the Japanese*, S. Watanabe, S. Kondo, and E. Matsunaga, eds., University of Tokyo Press, Tokyo, pp. 296–303.

Prager, E.M. and Wilson, A.C. 1978. Construction of phylogenetic trees for proteins and nucleic acids: comparison of alternative matrix methods. *J. Mol. Evol.* **11**: 129–142.

Reed, T.E. 1969. Caucasian genes in American Negroes. *Science* **165**: 841–858.

Roychoudhury, A.K. 1977. Gene diversity in Indian populations. *Hum. Gent.* **40**: 99–106.

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

Sarich, V.M. 1977. Rates, sample sizes, and the neutrality hypothesis for electrophoresis in evolutionary studies. *Nature* **265**: 24–28.

Sarich, V.M. and Wilson, A.C. 1966. Quantitative immunochemistry and the evolution of primate albumins: micro-complement fixation. *Science* **154**: 1563–1566.

Sarich, V.M. and Wilson, A.C. 1967. Immunological time scale for hominid evolution. *Science* **158**: 1200–1203.

Slatkin, M. and Maruyama, T. 1975. The influence of gene flow on genetic distance. *Am. Nat.* **109**: 597–601.

Smith, C.A.B. 1977. A note on genetic distance. *Ann. Hum. Genet.* **40**: 463–479.

Sokal, R.R. and Sneath, P.H.A. 1963. *Principles of Numerical Taxonomy*. Freeman, San Francisco.

Steinberg, A.G., Bleibtreu, H.K., Kurczynski, T.W., Martin, A.O., and Kurcyzynski, E.M. 1967. Genetic studies on an inbred human isolate. In: *Proc. 3rd Intl. Cong. Human Genetics*. The Johns Hopkins Press, Baltimore, pp. 267–289.

Weitkamp, L.R., Arends, T., Gallango, M.L., Neel, J.V., Schultz, J., and Shreffler, D.C. 1972.

Nei, M. 1977c. Estimation of mutation rate from rare protein variants. *Am. J. Hum. Genet.* **29**: 225–232.

The genetic structure of a tribal population, the Yanomama Indians. III. Seven serum protein systems. *Ann. Hum. Genet.* **35**: 271–279.

Weitkamp, L.R. and Neel, J.V. 1972. The genetic structure of a tribal population, the Yanomama Indians. IV. Eleven erythrocyte enzymes and summary of protein variants. *Ann. Hum. Genet.* **35**: 433–444.

Wilson, A.C., Carlson, S.S. and White, T.J. 1977. Biochemical evolution. *Annu. Rev. Biochem.* **46**: 573–639.

Wilson, A.C. and Sarich, V.M. 1969. A molecular time scale for human evolution. *Proc. Natl. Acad. Sci.* **63**: 1088–1093.

Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16**: 97–159.

Wright, S. 1943. Isolation by distance. *Genetics* **28**: 114–138.

Wright, S. 1951. The genetic structure of populations. *Ann. Eugenics* **15**: 323–354.

Zouros, E. 1979. Mutation rates, population sizes, and amounts of electrophoretic variation of enzyme loci in natural populations. *Genetics*, in press.

Zuckerkandl, E. and Pauling, L. 1962. Molecular disease, evolution, and genic heterogeneity. In *Horizones in Biochemistry*, M. Kasha and B. Pullman, eds., Academic Press, New York, pp. 189–225.

Zuckerkandl, E. and Pauling, L. 1965. Evolutionary divergence and convergence in proteins. In *Evolving Genes and Proteins*, V. Bryson and H. J. Vogel, eds., Academic Press, New York, pp. 97–166.