Vol. 106, No. 949 The American Naturalist May–June 1972

GENETIC DISTANCE BETWEEN POPULATIONS*

MASATOSHI NEI

Division of Biological and Medical Sciences, Brown University, Providence, Rhode Island 02912

In a study of the number of gene differences between related species, Nei (1971*a*) developed a statistical method for estimating the number of codon differences per gene and the divergence time between closely related species. This method utilizes electrophoretic data on protein identity between different species. A similar method was used independently by Kimura and Ohta (1971) for estimating the divergence time between two subspecies of the house mouse. This method, however, is not very useful for the study of gene differences between races or closely related local populations within a species, since the gene differences are not large enough for the effect of polymorphism within populations to be neglected.

Recently, Nei (1971*b*) modified his method, taking into account the effect of polymorphism within populations. He defined the normalized identity of genes between populations, which is equivalent to protein identity (Nei 1971*a*). He then related it to the accumulated number of gene differences per locus, which was now called *genetic distance*. This measure of genetic distance has several advantages over those proposed by Cavalli-Sforza and Edwards (1967), Balakrishnan and Sanghvi (1968), Hedrick (1971), and others. (1) It is related to Malecot's coefficient of kinship in a simple way. (2) It measures the accumulated number of gene substitutions per locus. (3) If the rate of gene substitutions per year is constant, it is linearly related to evolutionary time. (4) In some migration models it is linearly related to geographical distance or area.

A detailed account of genetic distance and its extensions is presented here.

IDENTITY OF GENES AND GENETIC DISTANCE

Consider two randomly mating diploid populations, X and Y, in which multiple alleles segregate at a locus. Let x_i and y_i be the frequencies of the ith alleles in X and Y, respectively. The probability of identity of two randomly chosen genes is $j_X = \Sigma x_i^2$ in population X, while it is $j_Y = \Sigma y_i^2$ in population Y. The probability of identity of a gene from X and a gene from Y is $j_{XY} = \Sigma x_i y_i$. If there is no selection and each allele is derived

* This research was supported in part by National Science Foundation grant GB–21224 and by Public Health Service grant GM-17719.

from a single mutation in an ancestral generation, the expected values of j_X and j_Y are equal to Wright's coefficients of inbreeding in X and Y, respectively, while that of j_{XY} is equal to Malecot's (1967) coefficient of kinship. The normalized identity of genes between X and Y with respect to this locus is defined as

$$I_j = j_{XY}/\sqrt{j_X j_Y}. \tag{1}$$

This quantity is unity when the two populations have the same alleles in identical frequencies, while it is zero when they have no common alleles.

The normalized identity of genes between X and Y with respect to all loci is defined as

$$I = J_{XY}/\sqrt{J_X J_Y}, \tag{2}$$

where J_X, J_Y, and J_{XY} are the arithmetic means of j_X, j_Y, and j_{XY}, respectively, over all loci, including monomorphic loci. Theoretically, it is possible to compute the arithmetic mean of I_j rather than the above quantity, but the genetic interpretation of the arithmetic mean is not simple, although numerical values of these two quantities generally turn out to be close to each other. The genetic distance between X and Y is then defined as

$$D = -\log_e I. \tag{3}$$

This definition is quite appropriate if the rate of gene substitution per locus is the same for all loci. In this case, D measures the accumulated number of gene substitutions per locus, as will be seen below. Data on amino acid substitutions in some proteins, however, indicate that the rate of gene substitution varies considerably among loci (Dayhoff 1969). If this rate is not the same for all loci, D underestimates the number of gene substitutions per locus (Nei 1971*a*). Expression (2) is a more generalized definition of I in Nei (1971*a*), which was the proportion of identical proteins between two related species. Maruyama (1970*a*) used a quantity equivalent to I in (2) in his study of the effective number of alleles in subdivided populations, but apparently he was not interested in the use of I for measuring genetic distance.

When the rate of gene substitution varies with locus and all I_j's are large, a more appropriate measure of genetic distance is given by

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

where $I' = J'_{XY}/\sqrt{J'_X J'_Y}$, in which J'_X and J'_Y, and J'_{XY} are the geometric means of j_X, j_Y, and j_{XY}, respectively. I discuss the rationale and disadvantage of this measure later. Formula (4) can also be expressed as

$$D' = \left(\sum_{j=1}^{n} d_j \right) / n, \tag{4'}$$

where d_j is the value of $-\log_e I_j$ at the jth locus and n is the number of loci examined.

DIFFERENTIATION OF ISOLATED POPULATIONS

Suppose that a population splits into two isolated populations and thereafter no migration occurs between the two populations. Let I_0 be the normalized identity of genes between the two populations at the time of completion of sexual isolation. For simplicity, assume the effective sizes (N) of the two populations are equal and that they are in a steady state with balanced mutation, selection, and genetic random drift. Following Kimura (1971), assume that each new mutation is an allele not preexisting in either population. The probability of nucleotide substitution per DNA base per year or per generation is so small that this assumption should usually be valid. Under these conditions, the value of j_X at a particular locus varies from time to time, but its expectation remains constant. When gene substitution occurs through neutral mutations, $E(j_X)$ is equal to $1/(4Nu + 1)$, where u is the rate of neutral mutations per generation (Kimura and Crow 1964). If selection is important, $E(j_X)$ takes a more complicated form (see Kimura 1969). At any rate, if a large number of loci are examined, J_X can be assumed to be constant and equal to $E(j_X)$. The same is true for J_Y.

On the other hand, the expectation of j_{XY} gradually decreases as time goes on. Let α be the rate of gene substitution per locus per year. There is increasing evidence that the rate of gene substitution per locus per year is nearly constant (Dayhoff 1969). If neutral gene substitution is important, α is equal to the rate of neutral mutations per locus per year (Kimura 1968; Nei 1969; Crow 1969; King and Jukes 1969). If selection is important, it is given by $4Nsu'$, approximately, where s is the average selective advantage of a new mutation over the original allele and u' is the mutation rate per locus per year (Kimura and Ohta 1971). If the rate of gene substitution varies with time, the average value of α can be used in the following formulation.

The expectation of j_{XY} in the tth year after sexual isolation $[E(j_{XY}^{(t)})]$ is given by

$$\begin{aligned} E(j_{XY}^{(t)}) &= E(j_{XY}^{(0)})(1 - \alpha_X)^t(1 - \alpha_Y)^t \\ &\simeq E(j_{XY}^{(0)})e^{-(\alpha_X+\alpha_Y)t}, \end{aligned} \tag{5}$$

where α_X and α_Y are the values of α for populations X and Y, respectively. In the following we denote the average of α_X and α_Y by α. Then,

$$\begin{aligned} &E(j_{XY}^{(t)})\,[E(j_X^{(t)})E(j_Y^{(t)})]^{-\frac{1}{2}} \\ &= \frac{E(j_{XY}^{(0)})}{\sqrt{E(j_X^{(0)})E(j_Y^{(0)})}}\,e^{-2\alpha t}. \end{aligned} \tag{6}$$

Therefore, if a large number of loci are examined and α is the same for all loci, we have

$$I = I_0 e^{-2\alpha t} \tag{7}$$

approximately, where I is defined in (2). The genetic distance D then becomes

$$D = 2\alpha t - \log_e I_0. \tag{8}$$

The value of I_0 would be close to one in most cases, so that the second term in the above formula can generally be neglected. It is clear that D is essentially the same as that of Nei (1971*a*) and measures the accumulated number of gene (allele) differences per locus between the two populations.

The value of I_0 can be estimated under certain conditions. If a population splits into two in a generation and the size of one or both of the two resultant populations is small, the effect of the so-called founder principle occurs. Let x_i, y_i, and z_i be the frequency of the ith alleles at a locus in resultant populations X and Y and their parental population Z, respectively. The expectation of Σx_i^2 is

$$\begin{aligned} E(\Sigma x_i^2) &= E(1 - 2 \sum\sum_{i<j} x_i x_j) \\ &= 1 - 2 \sum\sum_{i<j} (\bar{x}_i \bar{x}_j + \sigma_{ij}), \end{aligned}$$

where $\bar{x_i}$ is the expectation of x_i and equal to z_i, and σ_{ij} is the covariance between x_i and x_j and equal to $-z_i z_j/(2N_X)$, in which N_X is the size of population X. Thus, $E(\Sigma x_i^2) = J_Z + (1 - J_Z)/(2N_X)$. Similarly, $E(\Sigma y_i^2) = J_Z + (1 - J_Z)/(2N_Y)$, where N_Y is the size of population Y. On the other hand, $E(\Sigma x_i y_i) = J_Z$. Therefore, the expectation of I_0 is

$$I_0 = \frac{J_Z}{\sqrt{[J_Z + (1 - J_Z)/(2N_X)]\,[J_Z + (1 - J_Z)/(2N_Y)]}}. \tag{9}$$

For example, if $J_Z = 0.8$, $N_X = \infty$, and $N_Y = 50$, then I_0 is 0.9987. If $N_Y = 5$ and the other parameters remain the same, I_0 is 0.9877. Thus, only when the size of a new population is extremely small does the founder principle become important.

Migration could occur between the X and Y populations before complete sexual isolation is established. In such cases, if the same rate of migration occurs for a considerable period of time, I_0 may be equated to its steady-state value (discussed below).

As mentioned earlier, when α varies with locus, D' may be a better estimate of the number of gene differences than D. Since the natural logarithm of $E(j_{XY}^{(t)})\,[E(j_X^{(t)})E(j_Y^{(t)}]^{-1/2}$ at the jth locus is $-2\alpha_j t$, where α_j is the value of α at this locus and I_0 is assumed to be one, D' can be written as

$$\begin{aligned} D' &= 2(\alpha_1 + \alpha_2 + \ldots + \alpha_n)\, t/n \\ &= 2\alpha_m t, \end{aligned} \tag{10}$$

where α_m is the average value of α_j. In practice, however, this value is affected by sampling errors of gene frequencies at the time of population survey as well as by genetic random drift to a considerable extent. These

factors are expected generally to inflate the estimate of $2\alpha_m t$. If one of I_j is zero, D' becomes infinity. The effect of these factors, however, will not be large if sample size is large and all of I_j's are larger than 0.7. At any rate, the true value of $2\alpha_m t$ is expected to be somewhere between D and D'.

GENETIC DISTANCE BETWEEN INCOMPLETELY ISOLATED POPULATIONS

If migration occurs between two populations, they always share some common alleles and differentiation is hindered. Nei (in preparation) studied the temporal change of identity of genes with migration. Here, however, I examine only the steady-state relation between the identity of genes and migration rate. In the following I consider only neutral genes, since the genetics of subdivided populations with selection is not well developed.

A. Island Model

Maruyama (1970*a*) worked out the steady-state formulae for the expected values of j_X and j_{XY} in the island model with a finite number of subpopulations. Let N, n, m, and u be the size of a subpopulation, number of subpopulations, migration rate, and mutation rate per generation, respectively. Here again assume that new mutations are always different from the alleles extant in the population. Simplification of Maruyama's formula gives

$$E(j_X) = (1-u)^2[a-(a-b)(1-u)^2]/(2NG);$$

$$E(j_{XY}) = b(1-u)^2/(2NG),$$

where $a = (1-m)^2 + m^2/n$, $b = m(2-m)/n$, and

$$G = 1-(1-u)^2\left(1+a-b-\frac{a}{2N}\right)+(a-b)(1-u)^4\left(1-\frac{1}{2N}\right).$$

From these formulae we can get the value of $I = E(j_{XY})/E(j_X)$, where $E(j_Y)$ is equal to $E(j_X)$. It is

$$I = b/[a-(a-b)(1-u)^2]. \tag{11}$$

When u is much smaller than m, I is given by

$$I = \frac{m(2-m)}{m(2-m)+2nu(1-m)^2} \tag{12}$$

approximately. It is of interest to see that I is independent of N. I is unity, as it should be, when $u = 0$ or when $m = 1$. On the other hand, $I = 0$ when $n = \infty$. In reality, of course, n is always finite.

From formula (12) we obtain

$$D = \log_e\left[1+\frac{nu(1-m)^2}{m(1-m/2)}\right], \tag{13}$$

which becomes

$$D = \frac{nu(1 - m)^2}{m(1 - m/2)} \tag{14}$$

approximately, if $nu(1 - m)^2/[m(1 - m/2)]$ is small (< 0.3).

Most values of D between local populations appear to be sufficiently small. Formula (14) indicates that D is linearly related to the number of subpopulations. If the number of subpopulations is linearly related to the geographical area in which they are located, D becomes a linear function of the area.

Nei (1971*b*) gave the formula $I = 1 - H/(4NmJ)$ for Wright's (1943) island model, where H is the average heterozygosity within a subpopulation and J is $1 - H$. This formula is based on the assumptions that there are only two alternative alleles and that the number of subpopulations is infinitely large.

B. Stepping-Stone Model

Maruyama (1970*a*) also studied the identity of genes between different subpopulations in the stepping-stone model of finite length. He provided simple formulae for $E(j_X)$ and $E(j_{XY})$ in a circular model, which is asymptotically equal to Kimura and Weiss's (1964) one-dimensional stepping-stone model when the number of subpopulations (n) is large. When n is large and the number of steps (k) between two subpopulations is relatively small ($k << n$), they become approximately

$$E(j_X) = (1 + 4N\sqrt{2mu})^{-1}; \tag{15b}$$

$$E(j_{XY}) = (1 + 4N\sqrt{2mu})^{-1}(1 - \sqrt{2u/m})^k, \tag{15a}$$

where m is the migration rate such that in every generation the proportion $m/2$ of the genes in a subpopulation is exchanged with each of its two neighboring subpopulations, and N and u are the same quantities as before. Therefore, the normalized identity of genes between subpopulations which are k steps apart is given by

$$\begin{aligned} I(k) &= (1 - \sqrt{2u/m})^k \\ &\simeq e^{-\sqrt{2u/m}\,k}. \end{aligned} \tag{16}$$

Interestingly, this quantity is identical with the correlation coefficient of gene frequencies between two subpopulations when there are only two alternative alleles and n is infinitely large (Kimura and Weiss 1964). This is because the variance of gene frequencies within a subpopulation and the covariance of gene frequencies between subpopulations can be directly related to J_X and J_{XY}, respectively (see Maruyama 1970*a*, 1970*b*). Formula (16) also holds true when the geographical distribution of individuals is continuous rather than discrete. Malecot (1959) has shown that the identity

of genes by descent between two individuals living apart with distance x is given by

$$\phi(x) = ce^{-\sqrt{2u/\sigma^2}\,x},$$

where σ^2 is the variance of migration distance in a generation, and c is a constant. If new mutations are always different from the alleles extant in the population, the identity of genes by descent is equal to J_{XY}. Therefore, $I(x) = \exp(-\sqrt{2u/\sigma^2}\,x)$.

Returning to the stepping-stone model, the genetic distance between two populations which are k steps apart is given by

$$D(k) = \sqrt{2u/m}\,k. \tag{17}$$

Thus, $D(k)$ is linearly related to the geographical distance. Expression (16) indicates that the proportion of common alleles between two neighboring populations ($k = 1$) is $1 - \sqrt{2u/m}$, while they have different alleles with probability $\sqrt{2u/m}$. Thus, $D(k)$ again measures the accumulated number of allele differences per locus.

Kimura and Weiss (1964) and Weiss and Kimura (1965) also studied the correlation of gene frequencies in a two-dimensional stepping-stone model, while Malecot (1959, 1967) investigated the coefficient of kinship in a continuous two-dimensional distribution. These studies indicate that the normalized identity of genes between two subpopulations which are k_1 steps apart in one coordinate and k_2 steps apart in the other in the stepping-stone model is given by

$$I(x) = ce^{-\sqrt{2u/m}\,x}/\sqrt{x}, \tag{18}$$

where k_1 and k_2 are both assumed to be equal to or larger than five, x is $(k_1^2 + k_2^2)^{1/2}$, and c is a constant. Therefore, the genetic distance is given by

$$D(x) = \sqrt{2u/m}\,x + \tfrac{1}{2}\log_e x - \log_e c. \tag{19}$$

In this case $D(x)$ is not linear with respect to x.

NUMERICAL EXAMPLE

I have applied my method to data of Selander, Hunt, and Yang (1969) on protein variation in two subspecies of house mouse, which data were also used by Hedrick (1971) to illustrate his measure of genetic similarity. Selander, Hunt, and Yang (1969) provided gene frequency data for 36 different proteins in four populations (populations 1, 2, 3, and 4) of *Mus musculus musculus* and two populations (populations 5 and 6) of *M. m. domesticus*. The 36 proteins studied are believed to be controlled by 41 different loci, and 17 of these show polymorphism. Estimates of I and D are presented in table 1. Here I and D refer only to those alleles that are detectable by electrophoresis. The identity of genes between populations

TABLE 1
ESTIMATES OF I AND D AMONG SIX POPULATIONS OF THE HOUSE MOUSE

	M. m. musculus				*M. m. domesticus*	
POPULATION	1	2	3	4	5	6
1	(.9319)	.0178	.0256	.0210	.1941	.1959
2	.9823	(.9031)	.0147	.0094	.1701	.1713
3	.9749	.9854	(.9155)	.0057	.1906	.1859
4	.9792	.9906	.9943	(.8899)	.1337	.1292
5	.8236	.8436	.8264	.8748	(.9319)	.0018
6	.8221	.8425	.8303	.8787	.9982	(.9301)

SOURCE.—Selander, Hunt, and Yang (1969).
NOTE.—The values above the diagonal are estimates of D, and the values below the diagonal are estimates of I. The values on the diagonal are of average homozygosity (J_X).

within subspecies is quite high, and consequently genetic distance is small. The two subspecies share 82%–88% of common genes, values considerably higher than the rough estimate by Selander, Hunt, and Yang (1969) of 68%. The average genetic distance between the two subspecies is 0.1714, while that for local populations within subspecies is 0.0137. Thus the former is 12 times higher than the latter. If electrophoresis detects one-third of amino acid substitutions, the number of gene differences between the two subspecies is estimated to be 0.5 per locus.

DISCUSSION

In recent years several authors proposed different measures of genetic distance between populations. In many of them, however, it is not clear what biological unit they are going to measure. From the standpoint of genetics, the most appropriate measure of genetic distance would be the number of nucleotide or codon differences per unit length of DNA.

Theoretically, it is possible to determine the number of nucleotide differences by biochemical techniques. At the present time, however, sequencing of nucleotides is expensive and time-consuming even for a short length of DNA. DNA hybridization techniques now available are too crude to be used for detecting a small number of nucleotide differences that would occur among local populations within a species. Therefore, we are forced to use other techniques. The method proposed here is one such technique. Of course, the accuracy of my method is affected by a number of factors such as detectability of gene differences (e.g., by electrophoresis), varying rate of nucleotide substitution at different loci, etc., as discussed by Nei (1971*a*). Therefore estimates of gene differences obtained by my method should be regarded as first approximations. Nevertheless, they seem to be important, since very little is known about the gene differences between local populations.

I am concerned here with genetic distance between local populations within a species. My method does not give reliable estimates when I is close

to zero. However, if substitution of synonymous codons is disregarded, $2\alpha t$ can be written as $2n\lambda_a t$, where n is the number of amino acids that compose a polypeptide and λ_a is the average rate of amino acid substitutions per residue per year (Nei 1971*a*). Therefore, if amino acid sequence data are available for many different proteins, the value of D can be estimated even for a pair of distantly related species. This makes D a general measure of genetic distance for any pair of organisms.

I have defined the normalized identity of genes in equation (2). This concept itself may be useful in some cases. It measures the proportion of genes that are common in the two populations under investigation. Sneath (Sokal and Sneath 1963) proposed the use of J_{XY} as a measure of genetic distance. Similarly, Morton et al. (1971) proposed the use of Malecot's coefficient of kinship, which becomes equal to J_{XY} under certain conditions. Expression J_{XY} is an appropriate measure of the identity of genes between two individuals (two genomes), one from population X and the other from population Y. Since, however, populations as a whole are to be compared, the normalized identity appears to be more appropriate than J_{XY} itself.

So far I have considered only randomly mating diploid populations. However, definitions of I and D hold true in any population, if populations can be adequately defined, whether they are haploid, tetraploid, or selfing populations. This is because the definition of I and D depend solely on gene frequencies rather than on genotype frequencies. Of course, such quantities as α and $E(j_X)$ may depend on ploidy and mating system. For example, in a population of N selfing individuals, the steady-state value of $E(j_X)$ under mutation pressure will be $1/(2Nu + 1)$ instead of $1/(4Nu + 1)$.

In the present measure of genetic distance, the triangle inequality, which is often assumed in numerical taxonomy (see, e.g., Levandowsky and Winter 1971), does not hold. However, evolution does not occur so as to assure this property at least at the nucleotide level.

SUMMARY

A measure of genetic distance (D) based on the identity of genes between populations is formulated. It is defined as $D = -\log_e I$, where I is the normalized identity of genes between two populations. This genetic distance measures the accumulated allele differences per locus. If the rate of gene substitution per year is constant, it is linearly related to the divergence time between populations under sexual isolation. It is also linearly related to geographical distance or area in some migration models. Since D is a measure of the accumulated number of codon differences per locus, it can also be estimated from data on amino acid sequences in proteins even for a distantly related species. Thus, if enough data are available, genetic distance between any pair of organisms can be measured in terms of D. This measure is applicable to any kind of organism without regard to ploidy or mating scheme.

LITERATURE CITED

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

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

Crow, J. F. 1969. Molecular genetics and population genetics. 12th Int. Cong. Genet., Proc. 3:105–113.

Dayhoff, M. O. 1969. Atlas of protein sequence and structure, 1969. Nat. Biomed. Res. Found., Silver Spring, Md.

Hedrick, P. W. 1971. A new approach to measuring genetic similarity. Evolution 25: 276–280.

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

———. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. Genetics 61:893–903.

———. 1971. Theoretical foundation of population genetics at the molecular level. Theoretical Population Biol. 2:174–208.

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

Kimura, M., and T. Ohta. 1971. Protein polymorphism as a phase of molecular evolution. Nature 229:467–469.

Kimura, M., and G. Weiss. 1964. The stepping-stone model of population structure and the decrease of genetic correlation with distance. Genetics 49:561–576.

King, J. L., and T. H. Jukes. 1969. Non-Darwinian evolution. Science 164:788–798.

Levandowsky, M., and D. Winter. 1971. Distance between sets. Nature 234:34–35.

Malecot, G. 1959. Les modèles stochastiques en génétique de population. Pub. Inst. Statist. Univ. Paris 8:173–210.

———. 1967. Identical loci and relationship. 5th Berkeley Symp. Math. Statist. Probability, Proc. 4:317–332.

Maruyama, T. 1970*a*. Effective number of alleles in a subdivided population. Theoretical Population Biol. 1:273–306.

———. 1970*b*. Stepping stone models of finite length. Advance. Appl. Probability 2:229–258.

Morton, N. E., S. Yee, D. E. Harris, and R. Lew. 1971. Bioassay of kinship. Theoretical Population Biol. 2:507–524.

Nei, M. 1969. Gene duplication and nucleotide substitution in evolution. Nature 221: 40–43.

———. 1971*a*. Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. Amer. Natur. 105:385–398.

———. 1971*b*. Identity of genes and genetic distance between populations. Genetics 68:s47.

Selander, R. K., W. G. Hunt, and S. Y. Yang. 1969. Protein polymorphism and genic heterozygosity in two European subspecies of the house mouse. Evolution 23: 279–390.

Sokal, R. R., and P. H. A. Sneath. 1963. Principles of numerical taxonomy. Freeman, San Francisco.

Weiss, G., and M. Kimura. 1965. A mathematical analysis of the stepping stone model of genetic correlation. J. Appl. Probability 2:129–149.

Wright, S., 1943. Isolation by distance. Genetics 31:114–138.