8. Genetic Distance and Molecular Phylogeny

Masatoshi Nei

Genetic distance is the degree of gene difference (genomic difference) between species or populations that is measured by some numerical method. Thus, the average number of codon or nucleotide differences per gene is a measure of genetic distance. There are various kinds of molecular data that can be used for measuring genetic distance. When the two species to be compared are distantly related, data on amino acid or nucleotide sequences are used (e.g. Dayhoff 1972). In this case, the genetic polymorphism within species is usually ignored, since its effect on the total genetic distance is small. When two closely related species or populations are compared, however, the effect of polymorphism cannot be neglected, and one has to examine many proteins or genes from the same populations. Sequencing of amino acids or nucleotides for many proteins or genes is time-consuming and expensive, so that more efficient molecular techniques are needed for studying the genetic relationship of closely related organisms.

There are two such techniques available now. One is protein electrophoresis, which has been used extensively for the last fifteen years in evolutionary studies. This technique does not produce data on amino acid or nucleotide sequence differences, but the gene frequency data generated by this technique can be used to estimate genetic distance. The other is the restriction enzyme technique, which is now used by an increasing number of investigators (e.g., Brown et al. 1979, Avise et al. 1979, Shah and Langley 1979, Brown 1983, Avise and Lansman 1983). This technique does not produce data on nucleotide sequence differences, but these differences can be estimated by using some statistical methods (Nei and Li 1979, Kaplan and Langley 1979, Gotoh et al. 1979, Nei and Tajima 1983). Unfortunately, this technique is still time-consuming, and the accuracy of the estimates of genetic distance obtained is not necessarily high.

In this chapter we first consider statistical methods for estimating genetic distance for closely related organisms with special consideration of electrophoretic data. We then describe mathematical models that are important for understanding the process and mechanism of genetic differentiation of populations in terms of certain genetic distance measures. Finally, we discuss the empirical relationship between genetic distance and evolutionary time and outline several problems concerning the reconstruction of phylogenetic trees by using genetic distances.

GENETIC DISTANCE

Measures of Genetic Distance

In the past several decades various measures of genetic distance have been proposed. Some are direct applications of earlier measures of morphological distances which have been used in classical numerical taxonomy. For example, the measures proposed by Sanghvi (1953), Steinberg et al. (1967), Balakrishnan and Sanghvi (1968), and Siciliano et al. (1973) are all direct applications of Mahalanobis's (1936) D^2 statistic to gene frequency data. Bhattacharyya's (1946) measure, which is essentially the same as Cavalli-Sforza and Edwards's (1967), can also be regarded as an extension of Mahalanobis's D^2 statistic for the case of discrete characters.

In these theories populations are represented as points in 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 genetic distance and evolutionary change of populations. 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. In some distance measures such as Latter's (1973) ϕ^*, the distance is related to Wright's F_{ST}, which is in turn related to evolutionary time under the assumption of no mutation. However, these measures are not a direct measure of the amount of gene differences between populations.

Compared with these measures, Nei's (1972) distance measure is based on an entirely different concept. It is intended to measure the number of gene or 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, protein electrophoresis cannot detect all codon differences, so that we are forced to deal with only those codon differences that are detectable by the technique. 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: the minimum, standard, and maximum estimates of codon differences per locus (Nei 1973b).

Definition of Nei's Distance Measures

Consider two populations, X and Y, in which l alleles are segregating at a locus. Let x_i and y_i be the frequencies of the ith allele in X and Y, respectively. The probability of identity of two randomly chosen genes is $j_X = \Sigma x_i^2$ in population X and $j_Y = y_i^2$ in population Y. The probability of identity of two genes chosen at random, one from each of the two populations, is $j_{XY} = \Sigma x_i y_i$. Here, Σ indicates summation over all alleles. For example, $x_i^2 = x_1^2 + x_2^2 + \cdots + x_l^2$, and $\Sigma x_i y_i = x_1 y_1 + x_2 y_2 + \cdots + x_l y_l$. 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 respective arithmetic means of j_X, j_Y, and j_{XY} over all loci in the genome, including monomorphic ones. Clearly, $D_{X(m)} \equiv 1 - J_X$, $D_{Y(m)} \equiv I - J_Y$, and $D_{XY(m)} \equiv 1 - J_{XY}$ are all equal to the proportion of different genes (alleles) 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,

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

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*. It is noted that if we denote by x_{ij} and y_{ij} the frequencies of the ith allele at the jth locus in populations X and Y, respectively, D_m can also be written as

$$\begin{aligned} D_m &= \frac{1}{2R}\sum_{j=1}^{R}\sum_{i=1}^{l_j}(x_{ij} - y_{ij})^2 \\ &= \sum_{j=1}^{R}\frac{d_j}{R}\,, \end{aligned} \tag{2}$$

where l_j and R are the number of alleles at the jth locus and the number of loci in the genome, respectively, and $d_j \equiv \Sigma_{i=1}^{l_j}(x_{ij} - y_{ij})^2/2$ is the distance at the jth locus.

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 (substitutions) may be given by

$$D = -\ln I\,, \tag{3}$$

where

$$I = \frac{J_{XY}}{\sqrt{J_X J_Y}} \tag{4}$$

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, it is linearly related to the time since divergence between the two populations. Also, in certain migration models it is linearly related to the geographical distance (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

$$D' = -\ln I' , \tag{5}$$

where $I' = J'_{XY}/\sqrt{(J'XJ'_Y)}$, in which J'_{XY}, J'_X, and J'_Y are the geometric means of j_{XX}, 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 population 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' becomes infinite.

Nei at al. (1976) 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. It is given by

$$D_v = \frac{1 - I}{I} . \tag{6}$$

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

Estimation of Genetic Distance

Theoretically, the genetic distance between two populations is defined in terms of the poplation gene frequencies for all loci in the genome. In practice, however, it is virtually impossible to examine all genes in the populations for all loci. 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 (1978a).

Clearly, there are two sampling processes involved in this case: 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 of the ith allele at a locus in 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 Eqs. (1), (4), or (5) by $\hat{x}_i$ and $\hat{y}_i$, respectively.

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

$$\hat{D}_m = \left(\frac{\hat{J}_X + \hat{J}_Y}{2}\right) - \hat{J}_{XY} \tag{7}$$

and

$$\hat{D} = -\ln\left[\frac{\hat{J}_{XY}}{\sqrt{\hat{J}_X \hat{J}_Y}}\right] . \tag{8}$$

Obviously, Eq. (8) is valid only when the number of loci (r) is large (Li and Nei 1975).

The sampling variances of $\hat{D}_m$, $\hat{D}$, $\hat{D}_v$, and $\hat{I}$ can be computed by the methods given by Nei and Roychoudhury (1974a) and Nei (1978a, b). To compute the variance of $\hat{D}_m$, we first note that the unbiased estimate of d_j in (2) is given by

$$\hat{d}_j = \frac{2n_X \Sigma_i \hat{x}_{ij}^2 - 1}{2(2n_X - 1)} + \frac{2n_Y \Sigma_i \hat{y}_{ij}^2 - 1}{2(2n_Y - 1)} - \Sigma_i \hat{x}_{ij}\hat{y}_{ij} . \tag{9}$$

Therefore, the variance of $\hat{D}_m$ is

$$V(\hat{D}_m) = \frac{\sum_{j=1}^{r}(\hat{d}_j - \hat{D}_m)^2}{r(r-1)} . \tag{10}$$

The variances of $\hat{D}$, $\hat{D}_m$, and $\hat{I}$ are more complicated, and usually a computer is required for the computation unless the populations are highly monomorphic. Such a computer program is available upon request.

When $\hat{I}$ is lower than 0.9 for all population pairs and average heterozygosity is low for all populations, the variances [$V(\hat{I})$ and $V(\hat{D})$] of $\hat{I}$ and $\hat{D}$ are approximated by

$$V(\hat{I}) = \frac{\hat{I}(1 - \hat{I})}{r} \tag{11}$$

$$V(\hat{D}) = \frac{1 - \hat{I}}{\hat{I}r} \tag{12}$$

(Nei 1971, 1978b). The reason for this is that in this case single-locus genetic identity $I_j = \hat{j}_{XY}/\sqrt{(\hat{j}_X\hat{j}_Y)}$ usually takes a value close to 1 or 0, so that $\hat{I}_j$ approximately follows the binomial distribution (Ayala et al. 1974).

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 (1978a) 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 intralocus 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 to reduce the variance of the estimate of genetic distance.

Under certain assumptions genetic distance can be used to estimate the time after separation of two populations. 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 to test 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 (10) and its equivalent formulae for $\hat{D}$ and $\hat{D}_v$ are not appropriate for testing the difference between two distance estimates. This is because they include the variance resulting from the differences in the initial gene frequencies among loci at the time of population differentiation (Li and Nei 1975). However, the difference between a pair of D_m's can be tested in the following way. If we note $\hat{D}_m = \Sigma \hat{d}_j/r$, the difference between a pair of $\hat{D}_m$'s, say $\hat{D}_{m1}$ and $\hat{D}_{m2}$, can be written as

$$\begin{aligned}\hat{D}_{m1} - \hat{D}_{m2} &= \frac{\Sigma(\hat{d}_{j1} - \hat{d}_{j2})}{r} \\ &= \frac{\Sigma\delta_j}{r}\,,\end{aligned} \tag{13}$$

where $\delta_j = \hat{d}_{j1} - \hat{d}_{j2}$ is the difference in $\hat{d}$ for the jth locus. Therefore, the difference between $\hat{D}_{m1}$ and $\hat{D}_{m2}$ is tested by the ordinary t–test for δ_j. Strictly speaking, δ_j is not normally distributed, but the above test would given an approximate significance level, since the t–test is known to be robust. It should be noted that a significant difference between $\hat{D}_{m1}$ and $\hat{D}_{m2}$ also implies a significant difference between the corresponding standard distances

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 though its variance inevitably becomes large.

MATHEMATICAL MODELS OF POPULATION DIFFERENTIATION

The 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 genetic 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. With certain assumptions, this problem can be studied by a simple mathematical model. The assumptions we make are as follows:

- A population splits into two populations (X and Y) at a certain evolutionary time and thereafter no migration occurs between the two populations.
- 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 many natural populations, since closely related populations or species generally show the same degree of heterozygosity. In some cases, of course, the bottleneck effect seems to be important, and this effect will be considered later.

- 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.
- 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 tth year is

$$I_A = I_0 e^{-2\alpha t} , \tag{14}$$

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

$$D = 2\alpha t + D_0 , \tag{15}$$

where $D_0 = -\ln I_0$. In the present model $I_0 = 1$, so that $D_0 = 0$. It is clear from (15) 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 incorrect. Nei et al. (1976a) have shown that the rate of amino acid substitution per polypeptide 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 follow a gamma distribution. Furthermore, studies on the variances of single-locus heterozygosity and genetic distance in various organisms (more than one hundred different species) 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's (1979) study on the relative mutation rates supports this conclusion. It is also noted that variation of the mutation rate is apparently related to the subunit molecular weight of protein (Koehn and Eanes 1977, Ward 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 $b = \bar{\alpha}/V(\alpha)$, in which $\bar{\alpha}$ and $V(\alpha)$ are the mean and variance of α. The expected genetic identity is then given by

$$\begin{aligned} \bar{I}_A &= \frac{E(\Sigma j_{xy})}{\sqrt{E(\Sigma j_x)E(\Sigma j_y)}} \\ &= \frac{\Sigma E(j_i)e^{-2\alpha_i t}}{\Sigma E(j_i)} \\ &\simeq \int_0^\infty f(\alpha)e^{-2\alpha t}d\alpha = \left(\frac{a}{a-2\bar{\alpha}t}\right)^a , \end{aligned} \tag{16}$$

where $E(j_i)$ is the expected homozygosity at the ith locus, and Σ stands for the summation for all loci in the genome. Equation (16) is expected to give an overestimate of the true value of expected genetic identity, since it is based on the assumption of no correlation between $E(j_i)$ and $\exp(-2\alpha_i t)$ though in practice there should be a positive correlation. In the case of neutral alleles, the effect of the above assumption on $\bar{I}_A$ can be evaluated, but unless $2\alpha_i$ is very large, the effect does not seem to be important (Griffiths 1980).

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

$$\bar{I}_A = \frac{1}{1+2\bar{\alpha}t} . \tag{17}$$

Therefore, the mean number of gene substitutions per locus ($2\bar{\alpha}t$) can be estimated by D_v in (6):

$$D_v \equiv 2\bar{\alpha}t = \frac{1-\bar{I}_A}{\bar{I}_A} . \tag{18}$$

Mathematically, $D_v > D$, but the difference between (14) and (17) is small when t is relatively small (see Table 8.1). However, note that, because of the assumption we have made above, Eq. (18) is expected to give an underestimate of $2\alpha t$ when $2\alpha t$ is large.

Table 8.1 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 Eqs. (14), (17), (20), and (22), respectively. In this computation the rate of gene substitution ($\alpha = v$) was assumed to be 10^{-7} per year (see text).

Time ($\times 10^3$ yrs)	I_A	$\bar{I}_A$	I_E	$\hat{I}_E$	Time ($\times 10^6$ yrs)	I_A	$\bar{I}_A$	I_E	$\hat{I}_E$
10	0.998	0.998	0.998	0.998	1	0.819	0.833	0.827	0.845
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	0.835	0.847	0.842	0.857	20	0.018	0.200	0.207	0.333

Recently, Hillis (1984) claimed that the effect of variation of substitution rate among loci can be taken care of if we redefine I in Eq. (3) as the mean of $j_{XY}/\surd(j_X h_Y)$ over loci. However, a theoretical basis for this claim has not been found.

Either formula (15) or (18) 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 α to be 10^{-7} for electrophoretic data (see the following section). Therefore, assuming $D_0 = 0$, t may be estimated by

$$t = 5 \times 10^6 \times D\,. \qquad (19)$$

It should be emphasized, however, that the above value of α is based on a number of assumptions, and thus (19) gives only a very rough estimate of divergence time.

It should be mentioned that Eqs. (15) and (18) are valid only when a large number of loci are studied, since each event of gene substitution is subject to large stochastic errors. Nei and Tateno (1975) studied the distribution of single-locus gene identity [$I_j = j_{XY}/\surd(j_X j_Y)$] under the assumption of neutral mutations by using computer simulation. The results obtained show that I_j shows an inverse J-shaped distribution when $2\bar{\alpha}t$ is small, 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).

Strictly speaking, Eq. (15) is not appropriate for electrophoretic data, even if the mutation rate is the same for all loci. This is because at the electrophoretic level the effect of back mutations becomes important as t increases. This problem can be studied by using Ohta and Kimura's (1973) stepwise model of neutral mutations, although some authors (Ramshaw et al. 1979, Fuerst and Ferrell 1980, McCommas 1983) have questioned the appropriateness of this model to electrophoretic data. Nei and Chakraborty (1973), Li (1976b), and Chakraborty and Nei (1976, 1977) have studied the expected genetic identity under the stepwise mutation model. The exact formula for the genetic identity for electrophoretic data (I_E) is rather complicated (Li 1976b), but for practical purposes we can use the following equation:

$$I_E = e^{-2vt} \sum_{r=0}^{\infty} \frac{(vt)^{2r}}{(r!)^2} , \tag{20}$$

where v is the mutation rate per generation (Nei 1978b). 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

$$\bar{I}_E = \frac{b^a}{\Gamma(a)} \sum_{r=0}^{\infty} \frac{t^{2r}}{(r!)^2} \frac{\Gamma(a + 2r)}{(b + 2t)^{a+2r}} \tag{21}$$

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

$$\bar{I}_E = \frac{1}{1 + 2\bar{v}t} \left[1 - \sum_{r=1}^{\infty} \frac{(2r)!}{(r!)^2} \left(\frac{\bar{v}t}{1 + 2\bar{v}t} \right)^{2r} \right] , \tag{22}$$

where $\bar{v}$ is the mean of v over all loci. This formula is expected to give an overestimate of the expected genetic identity when $2\bar{v}t$ is large, as in the case of Eq. (16).

Table 8.1 shows the values of genetic identity for the four different models, i.e., Eqs. (14), (17), (20), and (22). In this table, calendar year rather than generation is used as a unit of time, with $\bar{\alpha} = \bar{v} = 10^{-7}$. The relationship between genetic distance $D = -\log_e I$ and evolutionary time is also given in Figure 8.1 for four different models. 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, Eq. (19) may be used for estimating divergence time. However, if the divergence time increases further, the difference between the models becomes pronounced. In this case, Eq. (19) should not be used for estimating divergence time, since the assumption of the same mutation rate for all loci is not necessarily valid. The formula for I_E is also expected to give an underestimate, since in this case the same mutation rate is assumed for all loci. Therefore, for estimating divergence time, Eq. (17) or (22) seems to be more appropriate, although the applicability of the stepwise mutation model used in (22) is yet to be confirmed. At any rate, the numerical values in Table 8.1 or Figure 8.1 can be used for getting 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 loci such as those for blood groups, since the relationship between codon substitution and phenotypic change at these loci 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

$$\begin{aligned} J_X(t) &= J_Y(t) = 1 - [1 - J(0)]\left(1 - \frac{1}{2N}\right)^t \\ &\approx J(0) + [1 - J(0)]\left(\frac{t}{2N}\right) , \\ J_{XY}(t) &= J_{XY}(0) = J_X(0) = J_Y(0) = J(0) , \end{aligned} \tag{23}$$

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

$$D_m \doteq 1 - J(0)\left(\frac{t}{2N}\right) . \tag{24a}$$

$$D \cong \frac{1 - J(0)}{J(0)}\left(\frac{t}{2N}\right) , \tag{24b}$$

$$D_v \cong \frac{1 - J(0)}{J(0)}\left(\frac{t}{2N}\right) . \tag{24c}$$

Thus, as long as $t << 2N$, our distance measures can be used even for nonprotein loci. In most human populations, $t << 2N$ appears to hold.

In a computer simulation, Reynolds et al. (1983) showed that in the absence of mutation, D can increase nonlinearly with time even for a relatively short evolutionary time. However, this happened because they started with unrealistic initial allele frequencies (200 different alleles in a population of 100 diploid individuals) and traced the genetic change of populations until a substantial amount (about 40 percent) of genetic variability was lost. When the mutation-drift balance is maintained with the infinite-allele model, the relation-

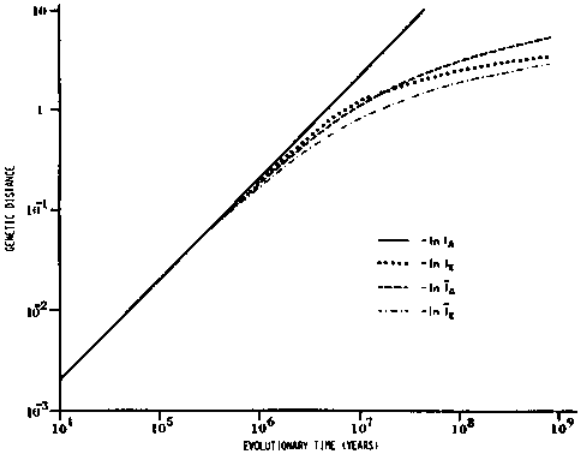

Fig. 8.1 Relationship between evolutionary time and genetic distance for four different models.

ship between *D* and *t* is approximately linear for a considerable period of evolutionary time, as seen from Fig. 8.1.

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 a genetic distance measure should have. Unfortunately, many distance measures such as Manhattan distance, the *d* of Cavalli-Sforza and Edwards (1967), Roger's (1972) distance, and that of Siciliano et al. (1973) [which is identical with that of Thorpe (1979)] do not have this property even under the effects of genetic drift and mutation alone (Nei 1976, Nei et al. 1983). In the absence of mutation, Latter's (1973b) distance ϕ^* can be related to evolutionary time by $t = -2N\ln(1 - \phi^*)$. In the presence of mutation (infinite-allele model), the expectation of ϕ^* becomes

$$E(\phi *) = \frac{J(\infty) + [J(0) - J(\infty)]e^{-2v+(1/2N)t} - J(0)e^{-2vt}}{1 - J(0)e^{-2vt}} ,$$

where $J(\infty) = 1/(4Nv + 1)$. When $J(\infty) = J(0)$, as is usually assumed, it reduces to

$$E(\phi *) = \frac{J(\infty)(1 - e^{-2vt})}{1 - J(\infty)e^{-2vt}} .$$

Therefore, $-2N\ln(1 - \phi^*)$ is no longer linear with evolutionary time. This limits the utility of ϕ^* and does not support the contention of Reynolds et al. (1983) that ϕ^* is preferable to *D* when evolutionary time is relatively short.

In our mathematical formulation, we assumed that the average heterozygosities of the two populations in question have remained constant throughout the entire evolutionary process. 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 in detail by Chakraborty and Nei (1974, 1977). They have shown that the genetic distance increases rapidly in the presence of bottlenecks and the rate of increase is higher when the bottleneck size is small than when it 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 (Fig. 8.2).

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:

$$I = \frac{J_{XY}}{J_X}, \tag{25}$$

where J_X is the mean homozygosity (gene identity) for the population whose size has remained constant. If we use this I in (3) or (6), 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 have gone 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

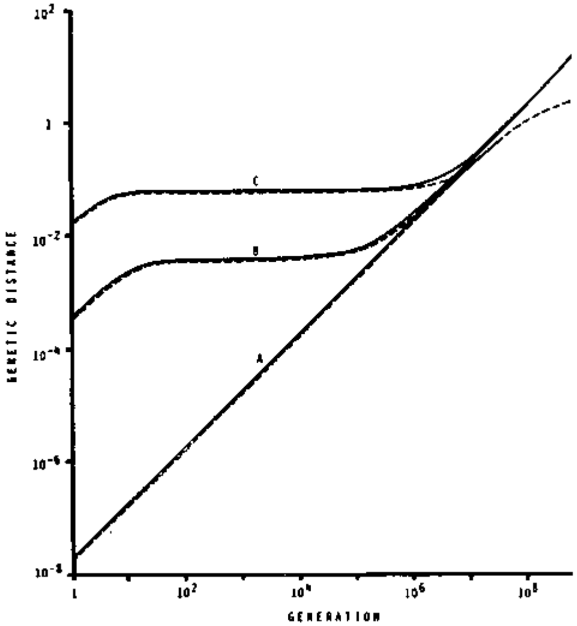

Fig. 8.2 Bottleneck effects on genetic distance. Solid lines represent the genetic distance for the infinite-allele model, whereas broken lines represent the distance for the stepwise mutation model. Computations have been made under the assumption that one isolated population (or species) is established through a bottleneck of size N_0 and thereafter population size increases to the level of the parental population following the sigmoid curve. The genetic distance in the ordinate represents the distance between this population and its parental population, which has undergone independent evolution. A: no bottlenecks. B: N_0 = 100. C: N_0 = 10.

the two populations under investigation. In this case, I may be computed by replacing J_X in (25) 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 (1976a) under the assumption of no selection. Their main conclusions are as follows:

- 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

$$I = \frac{m_1 + m_2}{m_1 + m_2 + 2\alpha} \tag{26}$$

approximately, if $2\alpha << m_1 + m_2 << 1$. 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).

- The approach to the steady state value is generally very slow; the number of generations required is of the order of the reciprocal of the mutation rate. Formula (26) indicates that the genetic distance between populations cannot be large unless migration rates are very small.
- If we know α, Eq. (26) can be used for estimating the maximum amount of migration between two populations. That is, if we write $m = (m_l + m_2)/2$, Eq.(26) becomes $I = m/(m + \alpha)$. Therefore, $m = \alpha I/(1-I)$. This formula has been used by a number of authors (e.g., Chakraborty and Nei 1974, Nei 1975, Larson et al. 1984).

EMPIRICAL RELATIONSHIP BETWEEN GENETIC DISTANCE AND EVOLUTIONARY TIME

In the previous section I discussed the theoretical relationship between genetic distance and the time since divergence between two populations. This relationship can be used for estimating the divergence time from genetic distance data. In practice, however, the underlying assumptions are not always satisfied, so that the straightforward application of the theory can be misleading. Recently, Thorpe (1982) and Avise and Aquadro (1982) surveyed published data on the relationship between genetic distance and evolutionary time and concluded that there is no solid evidence for a universal electrophoretic clock. However, they did not examine the possible causes of seemingly nonuniversal molecular clocks. In reality, genetic distance is affected by various factors other than evolutionary time, and these factors should be considered when assessing the relationship between genetic distance and evolutionary time.

In recent years a number of authors estimated evolutionary times from genetic distances and compared them with the estimates from other sources such as fossil records, separation of lands and seas, and island formation. In these studies most authors used Eq. (15) or $t = kD$, where k is a proportionality constant and given by $k=(2\alpha)^{-1}$. Table 8.2 shows the results of these studies. It should first be emphasized that the estimates of evolutionary times in this table are not as certain as they might suggest; the numerical values are presented only to give a rough idea. Some results such as those of Highton and Larson (1979) are not included because of the uncertainty of these estimates. Despite this reservation, however, Table 8.2 shows that the agreement between the estimates from genetic distance and other sources is reasonably good in most studies, particularly if we consider the large standard error of genetic distances. However, there is one problem. That is, the proportionality constant used is not always the same, and indeed, there is a 20–fold difference between the largest and smallest k values. Therefore, the data in this table suggest that there is no molecular clock that is universal to all organisms. Before rushing to this conclusion, however, we must examine each set of data carefully, taking into account the other factors that might have affected genetic distance estimates.

Proteins Used

The first factor to be considered is the set of proteins used for electrophoresis. As discussed by Nei (1971, 1975), the rate of gene substitution per locus per year in Eq. (15) or (18) may be expressed as

$$\alpha = nc\lambda\ , \tag{27}$$

where n is the average number of amino acids per polypeptide, c is the proportion of amino acid substitutions that are detectable by electrophoresis, and λ is the average rate of amino acid substitution per year. The value of $\alpha = 10^{-7}$ in Eq. (19) is obtained by using $n = 400$, $c = 0.25$, and $\lambda = 1 \times 10^{-9}$ (see Nei 1975, p. 33, for the rationale). In practice, however, different investigators use different sets of proteins, though large-scale electrophoretic surveys usually contain many commonly used proteins (Avise and Aquandro 1982). For example, the n value (515) for the proteins used by Nevo et al. (1974) was somewhat larger than that (400) of Nei and Roychoudhury (1974b). They also assumed that $\lambda = 2.1 \times 10^{-9}$. Mainly because of these differences, Nevo et al.'s $k = 1.5 \times 10^{6}$ was about three times smaller than that of Nei and Roychoudhury (1974b).

Table 8.2 Estimates of divergence time from genetic distance and other sources. Proportionality constant k for $t = kD$ is also given.

Organism	Distance	Time (years) estimated from		No. of loci	k	Source
		Distance[a]	Other sources			
Mammals						
Negroid and Mongoloid (human)	0.024	$(1.2\pm0.6)\times10^5$	$(5-20)\times10^4$	35	5×10^6	Nei and Roychoudhury (1974b)
Man and chimpanzee	0.62[b]	$(4\pm1)\times10^6$	5×10^6	44	(5×10^6)	Eq. (18)
Japanese and cont. macaques	0.11	5.4×10^5	$(4-5)\times10^5$	28	5×10^6	Nozawa et al. (1977)
Pocket gophers (*Thomomys*)	0.08	$(1.2\pm0.8)\times10^5$	$(1-2)\times10^5$	31	1.5×10^6	Nevo et al. (1974)
Pocket gophers (*Geomys*)	?[c]	2.7×10^5	Ca 3×10^5	22	8.1×10^5	Penney and Zimmerman (1976)
Woodrats (*Neotoma*)	0.18	$(1.5\pm0.9)\times10^5$	$(2-4)\times10^5$	20	8.1×10^5	Zimmerman and Nejtek (1977)
		$(9\pm4.5)\times10^5$			(5×10^6)	
Deer mice spp.	0.15	$(1.5\pm0.7)\times10^5$	$(1-5)\times10^5$	28	9.9×10^5	Gill (1976)
		$(7.5\pm3)\times10^5$			(5×10^6)	Eq. (19)
Ground squirrels (spp.)	0.56	$(5\pm1)\times10^6$	5×10^6	37	6.7×10^6	Smith and Coss (1984)
		$(4\pm1)\times10^6$			(5×10^6)	Eq. (18)
Ground squirrels (subsp.)	0.10	$(6.9\pm0.3)\times10^5$	7×10^5	37	6.7×10^6	Smith & Coss (1984)
		$(5.5\pm1)\times10^5$			(5×10^6)	Eq. (18)
Birds						
Galapagos finch	0.12	$(6\pm3)\times10^5$	$(5-40)\times10^5$	27	5×10^6	Yang and Patton (1981)
Reptiles						
Bipes spp.	0.62	$(3.1\pm1)\times10^6$	4×10^6	22	5×10^6	Kim et al. (1976)
		$(4\pm1)\times10^6$			(5×10^6)	Eq. (18)
Lizards (*Uma*)	0.28	$(5\pm2.5)\times10^6$	Ca 5×10^6	22	18×10^6	Adest (1977)
		$(1.6\pm0.8)\times10^6$			(5×10^6)	Eq. (18)

Table 8.2 (continued) Estimates of divergence time from genetic distance and other sources. Proportionality constant k for $t = kD$ is also given.

Organism	Distance	Time (years) estimated from		No. of loci	k	Source
		Distance[a]	Other sources			
Fishes						
Cave and surface fishes	0.14	$(7 \pm 4.6) \times 10^5$	$(3 - 20) \pm 10^5$	17	5×10^6	Chakraborty and Nei (1974)
Minnows	0.053	2.7×10^5	$(1 - 20) \times 10^5$	24	5×10^6	Avise et al. (1975)[d]
Panamanian fishes	0.32	5.8×10^6	$(2 - 5) \times 10^6$	28	18×10^6	Gorman & Kim (1977)
		$(1.9 \pm 0.6) \times 10^6$			(5×10^6)	Eq. (18)
Panamanian fishes	0.24[c]	3.5×10^6	$(2 - 5) \times 10^6$	31.4	18×10^6	Vawter et al. (1980)
		1.2×10^6			(5×10^6)	Eq. (18)
Echinoids						
Panamanian sea urchins	0.03 − 0.64	—	$(2 - 5) \times 10^6$	15 − 18	—	Lessios (1979)
	0.39[f]	2.4×10^6			(5×10^6)	Eq. (18)

[a] The standard error of D was not given in many papers. In this case it was computed by using Eq. (12), except for the case of small D.

[b] Data from King and Wilson (1975).

[c] It is not clear how the authors computed D.

[d] The authors used a different value of k, but $k = 5 \times 10^6$ gives a better result.

[e] Average for ten pairs of species.

[f] Average for four pairs of species.

Detectability of Protein Differences By Electrophoresis

The detectability of amino acid differences in proteins by electrophoresis depends on various biological and biochemical conditions of electrophoresis such as the tissue used and the pH and type of the gel. These conditions are not always the same for all organisms studied or for all laboratories where the experiments are done. The differences in these conditions are expected to cause some differences in the detectability of protein differences, c. Although Avise and Aquadro (1982) suggested that the technique of electrophoresis used is virtually the same for most laboratories, there is evidence that this is not necessarily the case. For example, King and Wilson (1975) reported a genetic distance of $D = 0.62$ between man and chimpanzee, whereas Bruce and Ayala (1979) and Nozawa et al. (1982) obtained $D = 0.39$ and 0.45, respectively, for the same pair of species. These differences are partly due to the differences in the proteins used. However, even when the same 15 protein loci common to the three studies were used, there were substantial differences: the D values for the data of King and Wilson, Bruce and Ayala, and Nozawa et al. were 0.83, 0.41, and 0.71, respectively (Nozawa et al. 1982). This indicates that c is not the same for all laboratories. In a study of ground squirrels, Smith and Coss (1984) speculated that the c value for their data was probably smaller than 0.25, and obtained $k = 6.7 \text{ x } 10^6$.

Sampling Errors

As mentioned earlier, a large number of loci should be used to obtain a reliable estimate of genetic distance, particularly when genetic distance is small. In practice, however, many investigators use a relatively small number of loci for various reasons. In these cases the results obtained could be misleading. For example, consider an extreme case in which two populations are virtually monomorphic and fixed for different alleles at 10% of the loci. In this case the expected genetic identity is 0.9 ($D = 0.11$), but the observed identity can be substantially larger or smaller than 0.9. Indeed, if r loci are examined, the observed value of I will be 1 (or D will be 0) with the probability of $(0.9)^r$. In the case of $r = 20$, this probability is 0.12, which is not very small. To make this probability smaller than 0.05, r must be equal to or larger than 29, whereas the minimum value of r that makes the probability less than 0.01 is $r = \log 0.01/\log(1 - I) = 2/0.0458 = 44$, where $I = 0.9$. This is a substantial number of loci. In practice, of course, the number can be a little smaller than this, because usually some loci are polymorphic and this reduces the variance of I or D to some extent. Nevertheless, we must be cautious about the effect of sampling error when the number of loci examined is small.

In a study of the genetic distance between the sea urchins of the Pacific and those of the Atlantic coast of Panama, Lessios (1979) observed an unusually low genetic distance for one of the four species pairs examined, although the Panama Isthmus is known to have been above sea level for the last

2–5 million years. From this observation, he concluded that electrophoretic data cannot be used for dating evolutionary time. However, he studied only 18 protein loci, so that this unusual result could well be due to sampling error (Vawter et al. 1980). Furthermore, if we take the average of the genetic distances for all the four species pairs, it becomes 0.39, which corresponds to a divergence time of 2.4 million years (Table 8.2). This divergence time is consistent with the estimate from geological data.

Nonlinear Relationship of D With Time

When the time since divergence between two species is long, the relationship between D and t is no longer linear, as mentioned earlier. This factor has not been taken into account properly in most of the empirical studies cited in Table 8.2. In some cases, the correction for this nonlinearity seems to improve the agreement between the estimates of divergence times from genetic distance and other sources. For example, King and Wilson (1975) obtained $D = 0.62$ between man and chimpanzee using Eq. (3). If we use Eq. (19) with $\alpha = 10^{-7}$, this gives $t = 3.1 \times 10^6$ years, which seems to be too low when compared with the estimate (5×10^6 years) obtained from fossil records and Sarich and Wilson's (1967) immunological distances. In this case, however, Eq. (18) is expected to give a better estimate of genetic distance (D_v) than Eq. (3). If we use Eqs. (18) and (19), we have $t = (4 \pm 1) \times 10^6$ years, which is no longer incompatible with the estimate from other sources. Furthermore, if we use Eq. (22), t becomes even larger. Similarly, the estimate of t by Kim et al. (1976) can be improved by using Eq. (18), as shown in Table 8.2. The same property was noted by Smith and Coss (1984) in their study of ground squirrels, though they used a proportionality constant of 6.7×10^6.

The largest proportionality constant k in Table 8.2 is that of Gorman and Kim (1977), Adest (1977), and Vawter et al. (1980). This value was obtained by comparing Sarich and Wilson's (1967) immunological distance (d_I) with D (Maxson and Wilson 1974). In this comparison, however, many D values larger than 1 were used. Therefore, the k value obtained is expected to be an overestimate. In a later study of the regression coefficient (b) of d_I on D, Maxson and Maxson (1979) obtained $b = 26$. Since d_I is (stochastically) related to evolutionary time t by $t = 6 \times 10^5 d_I$ in mammals, reptiles, and amphibians (Prager et al. 1974), we obtain $t = 16 \times 10^6 D$. In a similar study Highton and Larson (1979) obtained $b = 24$ and $t = 14 \times 10^6$D. The k values obtained in these studies are still about three times as large as that in Eq. (19), but they are again based on many D values larger than 1. If we use only the D values less than 0.5 in these studies, the k value seems to be substantially lower than 14×10^6 (see Fig. 4 of Highton and Larson 1979). (The regression coefficient should be computed by fitting $d_I = bD$ rather than $d_I = a + bD$, as in Highton and Larson.)

Nevertheless, the value of $k = 18 \times 10^6$ seems to be much better than $k = 5 \times 10^6$ in explaining Vawter et al.'s (1980) and Adest's (1977) electrophoretic data. In these cases one can use Eq. (18) to compute the expected evolutionary time under the assumption of $k = 5 \times 10^6$, but the values obtained are much smaller than the estimates obtained from other sources. The only data that can be accommodated with the value of $k = 5 \times 10^6$ are those of Gorman and Kim (1977) (Table 8.2). Since the data of Vawter et al. are based on 10 pairs of species, their results cannot be dismissed as a special case. This suggests either that the fishes and reptiles studied by Vawter et al. and Adest do not show the same evolutionary rates as those for many other organisms or that the electrophoretic technique used for these organisms did not detect protein differences so efficiently as in other organisms.

Bottleneck Effects

One of the troublesome problems in dating evolutionary time from genetic distance data is the bottleneck effect. As we have seen earlier, the bottleneck effect accelerates the increase of D temporarily. This acceleration occurs because the homozygosity [J_X or J_Y in Eq. (3)] in one or both populations increases under the bottleneck effect. The expectation of J_{XY} is not affected by the bottleneck effect. If the population size is restored to the original level after going through a bottleneck, J_X or J_Y also gradually returns to the original level. Once J_X and J_Y reach the original level, the bottleneck effect on D can no longer be detected. Usually, however, it takes a long time before this effect disappears. At any rate, in the presence of bottlenecks Eq. (19) is expected to give an overestimate of evolutionary time. On the other hand, if one tries to calibrate the evolutionary clock by using D values which are affected by a bottleneck, a relatively small value of k will be obtained.

In many cases, it is difficult to know whether a particular population or a particular pair of populations has undergone bottlenecks. In some cases, however, there is clear-cut evidence of a bottleneck effect, and we can make a correction for it. For example, in the case of cave and surface fishes of *Astyanax mexicanus* the cave populations are very small and clearly derived from the surface populations (Avise and Selander 1972). At the present time, the average homozygosities in the cave populations are virtually 1, and if we use Eq. (25) the bottleneck effect is eliminated. The genetic distance for this species in Table 8.2 has been estimated in this way (Chakraborty and Nei 1974).

In some other cases the bottleneck effect (or the effect of population size reduction) can be inferred from the level of heterozygosity or homozygosity even if we do not know the history of the populations. For example, in the case of pocket gophers *Thomomys talpoides* studied by Nevo et al. (1974), there are a number of subspecies which are fixed for different chromosome numbers. Apparently, these subspecies have been isolated from each other for a long time. Furthermore, average heterozygosity is quite low in these subspecies. Therefore, it is possible that the genetic distances among them are affected by bottlenecks. Possibly, for this reason the k value for these subspecies is lower than 5×10^6. Unfortunately, we cannot make a correction for D in this case. A similar bottleneck effect might have occurred in the other species (*Geomye* species) of pocket gophers listed in Table 8.2.

General Remarks

It is now clear that the conclusion that there is no solid evidence for a universal electrophoretic clock (Thorpe 1982, Avise and Aquadro 1982) is not so firm. Rather, if we take into account various factors that affect the relationship between D and t, electrophoretic data seem to be useful for getting a rough idea about evolutionary time. It should be noted that D is intended to measure the number of amino acid substitutions that are detectable by electrophoresis. Therefore, so long as amino acid substitution occurs at a constant rate, as seems to be the case, D should increase as evolutionary time increases. However, electrophoretic data give a less accurate estimate of evolutionary time than amino acid sequence data, since they are affected by several factors such as the detectability of protein differences and bottleneck effects. Therefore, when D is to be used as a molecular clock, it should be understood that it gives only a rough estimate of evolutionary time.

It should also be noted that the electrophoretic clock may not be the same for all groups of organisms. Even if the same electrophoretic technique is used, the detectability of protein differences may vary with the group of organisms used (Avise and Aquadro 1982). If this is the case, separate electrophoretic clocks must be used for different groups of organisms. This certainly reduces the utility of the molecular clock, but even under this restriction, the concept of a molecular clock is useful for clarifying the evolutionary relationships of closely related organisms, where no other methods for estimating evolutionary time are usually available. For some time, molecular evolutionists have thought that the rate of molecular evolution is significantly lower in birds than in mammals, reptiles, and amphibians (Prager et al. 1974, Avise and Aquadro 1982). Recent evidence, however, suggests that what is wrong may not be molecular data but fossil records (Wyles et al. 1983). If more studies are done on the relationship between genetic distance and evolutionary time, similar results may be obtained in some cases.

Finally it should be emphasized that the electrophoretic clock is mainly useful for closely related species. If D is too large (say $D \geqslant 1$), its variance becomes very large even if a substantial number of loci are studied, so that the reliability of dating declines. The high frequency of backward and parallel mutations at the electrophoretic level in the case of $D \geqslant 1$ makes the clock unreliable. Nozawa et al. (1982) noted that the estimate of the time since divergence between man and macaques was much smaller than that from fossil records (about 30 million years). I believe that man and macaques are too far distant for the electrophoretic clock to be applied.

RECONSTRUCTION OF PHYLOGENETIC TREES

Genetic distances can be used not only for estimating evolutionary times but also for constructing phylogenetic trees. For the latter purpose we do not have to know the exact relationship between genetic distance and evolutionary time. As long as *D* is approximately linear with time, the evolutionary relationship among organisms can be estimated by using genetic distances.

There are many methods for constructing a phylogenetic tree from genetic distance data, but the two most frequently used for electrophoretic data are the unweighted pair-group method (UPGMA) and Farris's (1972) distance Wagner method. UPGMA was originally proposed for phenetic classification by Sokal and Michener (1958), but it can be used for phylogeny construction as long as the distance measure used is (stochastically) proportional to evolutionary time. A simple explanation of this method is given by Nei (in press). When the topology of the tree to be made is known, this method gives least-squares estimates of branch lengths (Chakraborty 1977). On the other hand, Farris's method is intended to construct a minimum evolution tree (a tree requiring a minimum number of evolutionary changes) and supposedly requires a distance measure that satisfies the triangle inequality. A distance measure that is often used for this purpose is Rogers's (1972) distance. Tateno et al. (1982) and Nei et al. (1983) have shown that when genetic distance is subject to a large stochastic error, this method tends to give overestimates of gene substitutions, contrary to the original intention. To rectify this property, Tateno et al. (1982) modified the Farris method. A brief account of the Farris method and its modified version is given by Nei (in press).

In the last decade there has been a great deal of controversy over the relative merits of various tree-making methods. Some of the discussions are quite philosophical, whereas some others are based upon conjectures on the evolution of special groups of organisms. The main problem in this controversy is that in most cases we do not know the true evolutionary tree with which a reconstructed tree is to be compared. To address this problem Tateno et al. (1982) and Nei et al. (1983) conducted computer simulations to find out which method reconstructs the true tree with the highest probability. In simulation study one can set up a model tree, and a reconstructed tree can be compared with this model tree. It is therefore possible to know which method yields a better tree than others. In the following I present some of their results, particularly those of Nei et al. (1983).

Computer Simulation Studies

The method of computer simulation used in the study by Nei et al. (1983) can be summarized:

- The evolutionary change of populations was assumed to occur solely by mutation and genetic drift with $4Nv = 0.2$ following the infinite-allele model, where N and v are the effective population size and the mutation rate per locus per generation, respectively.

- An ancestral population, which was in equilibrium with respect to the effects of mutation and genetic drift, was split into two populations. At a later time, one of the two descendant populations was again split into two populations This process was continued until the model tree given by Figure 8.3a was completed.
- The expected number of gene substitutions per locus was proportional to evolutionary time. In Figure 8.3a the expected number (M) of gene substitutions for the shortest branch is 0.1, but we also studied the case of $M = 0.004$. $M = 0.1$ would represent a typical case of interspecific comparison, and $M = 0.004$ a case of comparison of local races or subspecies.
- Starting from the ancestral population, the gene frequency change in each population was followed, and at the end of evolution, the gene frequencies for all populations were recorded.
- Using the gene frequency data, five genetic distance measures were computed: standard genetic distance, D; minimum genetic distance, D_m; Rogers's (1972) distance, D_R; Cavalli-Sforza's (1969) distance, f_θ; and Nei et al.'s (1983) angular distance, D_A. These five distance measures were used, because the accuracy of a tree reconstructed was expected to depend on the distance measure used.
- To see the effect of the number of loci used, ten sets of distance matrices were computed for each distance measure by using the first 10 loci, first 20 loci, and so on, until all 100 loci were used.
- For each of these distance matrices, evolutionary trees were reconstructed by using UPGMA, the Farris method, and the modified Farris method.

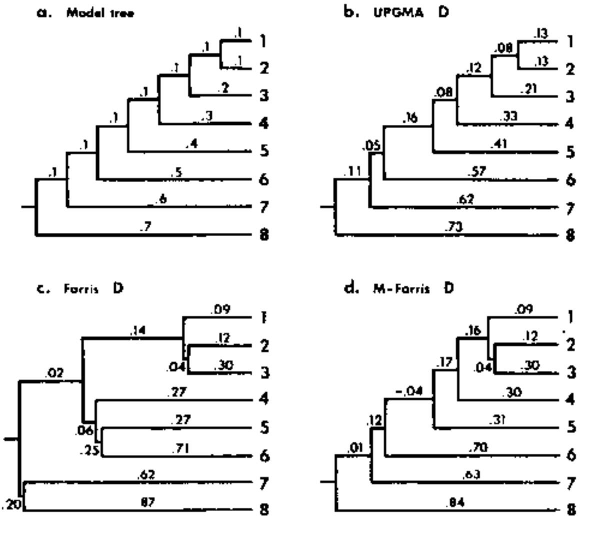

Fig. 8.3 Model tree (a) and reconstructed trees (b,c,d) by using D in one replication of computer simulation. The value given to each branch (or internode) in the model tree is the expected number of gene substitutions, whereas the corresponding number in the reconstructed tree is the estimate of branch length. $4Nv = 0.2$, $M = 0.1$, and the number of loci used is 50. (b) $d_T = 0$, $S_E = 0.073$; (c) $d_T = 8$, $S_E = 0.250$; (d) $d_T = 2$, $S_E = 0.138$.

- ● Each tree reconstructed was compared with the model (true) tree, and the extents of topological errors and errors in the estimates of branch length were evaluated.
- ● This simulation was repeated ten times for each value of $M = 0.1$ and $M = 0.004$.

There are two different criteria for measuring the deviation of a reconstructed tree from the model tree. One is the degree of distortion of the topology of a reconstructed tree, and the other is the amount of deviation of patristic (estimated) branch lengths from true lengths. The extent of the topological errors was measured by Robinson and Foulds's (1981) distortion index (d_T), which is roughly twice the number of interchanges of OTUs (operational taxonomic units) that are required for converting the topology of a reconstructed tree into that of the true tree. When the topology of a reconstructed tree is correct, d_T takes a value of 0. On the other hand, the amount of errors of the estimates of branch lengths was measured by the following average deviation of patristic distances from the expected distances.

$$S_E = \left[\frac{2\sum_{i>j}(D_{ij} - D'_{ij})}{n(n-1)}\right]^{1/2}, \tag{28}$$

where D_{ij} and D_{ij}' are the patristic distance and expected distance between OTUs i and j, respectively. The patristic distance between OTUs i and j is the sum of all branches connecting these two OTUs in the reconstructed tree, whereas the expected distance is the sum of branches connecting the same pair of OTUs in the model tree. The results obtained may be summarized as follows:

Topological errors. Figure 8.3 shows examples of reconstructed trees from one replication of computer simulation. In this case Nei's standard distances based on gene frequency data for 50 loci were used. The topology of the tree reconstructed by UPGMA is identical with that of the true tree, so that $d_T = 0$. On the other hand, the tree reconstructed by the Farris method has many topological errors, d_T being equal to 8, whereas d_T for the modified Farris tree is 2. Figure 8.4 shows the phylogenetic trees produced by using Rogers's distance and Cavalli-Sforza's f_Θ from the same set of gene frequency data. UPGMA again gives the correct topology ($d_T = 0$) for both D_R and f_Θ, whereas the other two methods give trees with several topological errors. Therefore, from the comparison of d_T alone, we can conclude that UPGMA is better than the other two evolutionary tree-making methods. However, this is the result from one replication of computer simulation, and to make a general conclusion we must consider the results from all ten replications.

The average distortion indices ($\bar{d}_T$) over ten replications for the case of $M = 0.1$ are given in A of Fig. 8.5 in relation to the number of loci used (r). It is seen that $\bar{d}_T$ is very large when $r = 10$ but rapidly declines as r increases. However, the decrease of $\bar{d}_T$ with increasing r is nonlinear, and the rate of decrease is not large when r is equal to or larger than 30. When 30 loci are used,

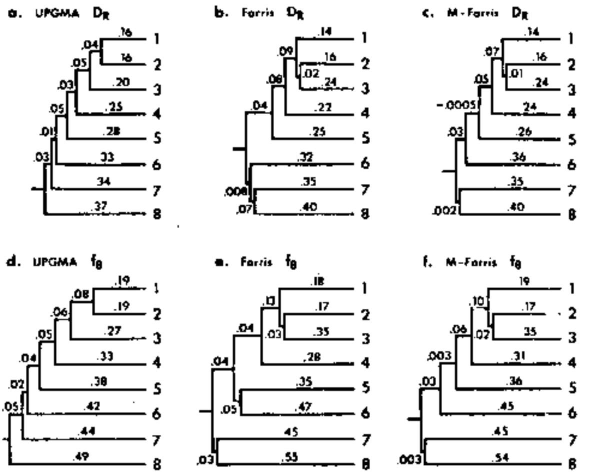

Fig. 8.4 Constructed trees by using D_R and f_θ in the same replication of computer simulation as that of Fig. 8.3. The value given to each branch is the estimate of branch length. $4Nv = 0.2$, $M = 0.1$, and the number of loci used is 50.

(a) $d_T = 0$, $S_E = 0.448$;
(b) $d_T = 6$, $S_E = 0.388$;
(c) $d_T = 4$, $S_E = 0.449$;
(d) $d_T = 0$, $S_E = 0.270$;
(e) $d_T = 6$, $S_E = 0.243$;
(f) $d_T = 4$, $S_E = 0.278$.

$\bar{d}_T$ is already about 2 for UPGMA, which means that the amount of error of a reconstructed tree is about one interchange of OTUs from the true tree. As r increases further, $\bar{d}_T$ decreases very slowly, and even with $r = 100$, $\bar{d}_T$ is not 0, except in one case. This result suggests that in the construction of a phylogenetic tree for a group of species at least 30 loci should be used.

Figure 8.5 also shows that UPGMA and the modified Farris method generally give a smaller value of $\bar{d}_T$ than the Farris method for all distance measures and for all r values. This is true even if Rogers's distance (D_R), which satisfies the triangle inequality, is used. It is also noted that the differences in $\bar{d}_T$ among different distance measures are generally small, though D_A tends to give a better topology than other measures. The better performance of D_A compared with others seems to be due to the fact that D_A has a relatively small coefficient of variation.

The value of $\bar{d}_T$ for the case of $M = 0.004$ is generally larger than that for $M = 0.1$, as expected (Fig. 8.5.). To make $\bar{d}_T$ equal to 2, a large number of loci must be used. However, the rate of decrease of $\bar{d}_T$ with increasing r again declines around $r = 30$. In this case UPGMA always shows a smaller value of $\bar{d}_T$ than the other two methods, and the modified Farris method tends to show a smaller value than the Farris method. This is true irrespective of the distance measure used.

Errors of the estimates of branch lengths. Another important criterion of the accuracy of a reconstructed tree is the deviation of estimates of branch lengths from true branch lengths. We have seen that in the example trees of Figs. 8.3 and 8.4 the topology of the tree reconstructed by UPGMA is correct irrespective of the distance measure used. However, the estimates of branch lengths are considerably different from each other. Comparison of these trees with the true tree (Fig. 8.3a) indicates that D gives a better result for estimating branch lengths than the other distance measures. Indeed, the S_E value for D is

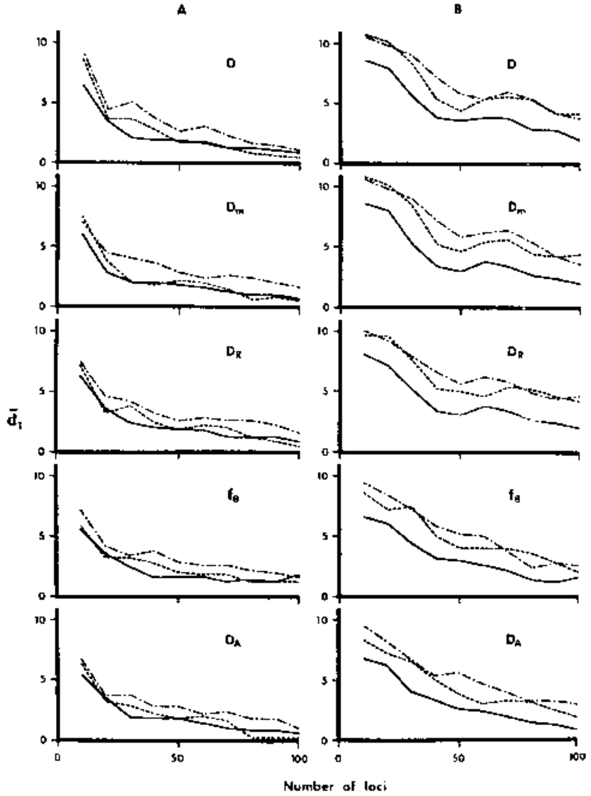

Fig. 8.5 Relationships between $\bar{d}_T$ and the number of loci used for the cases of $M = 0.1$ (A) and $M = 0.004$ (B). Solid line, UPGMA; chain line, Farris method; broken line, modified Farris method.

0.073, whereas the S_E for D_R and f_Θ is 0.448 and 0.270, respectively. Therefore, with this criterion the tree produced by UPGMA using D is the best among the three. One might think that the Farris method would give a good tree when D_R is used. That this is not the case can be seen from the comparison of Fig. 8.4b with Fig. 8.3c. Compared with D, D_R generally gives a tree in which the part near the root is condensed and the other part is elongated. This is because D_R is not proportional to the expected number of gene substitutions (Fig. 8.6). A similar pattern is observed for D_m, f_Θ, and D_A, although the results for D_m and D_A are not shown here.

The average values ($\bar{S}_E$) of S_E over all replications for the cases of 20 loci, 60 loci, and 100 loci examined are presented in Table 8.3. The value of $\bar{S}_E$ varies considerably with the tree-making method and the distance measure used. The smallest value is obtained when UPGMA with D is used. This supports our visual conclusion from Figs. 8.3 and 8.4. When D is used, the modified Farris method also shows a relatively small value of $\bar{S}_E$. In contrast, the $\bar{S}_E$ for the Farris method is nearly twice as large as that of UPGMA. In the case of D, $\bar{S}_E$ decreases as the number of loci used increases, as expected. When the other distance measures are used, UPGMA no longer gives the smallest value

Table 8.3 Means of average deviations of partristic distances from expected distances ($\bar{S}_E$). $4Nv$ = 0.2 and M = 0.1. These results are based on 10 replications. All results should be divided by 10^3.

	D	D_m	D_R	f_θ	D_A
			20 loci		
UPGMA	252 ± 26	539 ± 12	457 ± 11	305 ± 17	426 ± 13
Farris	461 ± 66	497 ± 13	417 ± 13	272 ± 16	386 ± 14
Modified Farris	291 ± 31	539 ± 11	458 ± 10	309 ± 16	427 ± 12
			60 loci		
UPGMA	136 ± 10	540 ± 5	456 ± 4	295 ± 11	426 ± 5
Farris	225 ± 12	511 ± 7	430 ± 5	268 ± 12	397 ± 5
Modified Farris	161 ± 9	541 ± 5	457 ± 4	297 ± 10	427 ± 5
			100 loci		
UPGMA	122 ± 8	534 ± 5	452 ± 4	296 ± 8	420 ± 5
Farris	204 ± 13	510 ± 5	427 ± 5	271 ± 9	395 ± 6
Modified Farris	140 ± 6	535 ± 5	452 ± 4	297 ± 8	421 ± 5

of $\bar{S}_E$. This is because the other distance measures are not linearly related with time (Fig. 8.6). However, the $\bar{S}_E$ values for the other distance measures are always larger than those for D in UPGMA. This result indicates that UPGMA, in

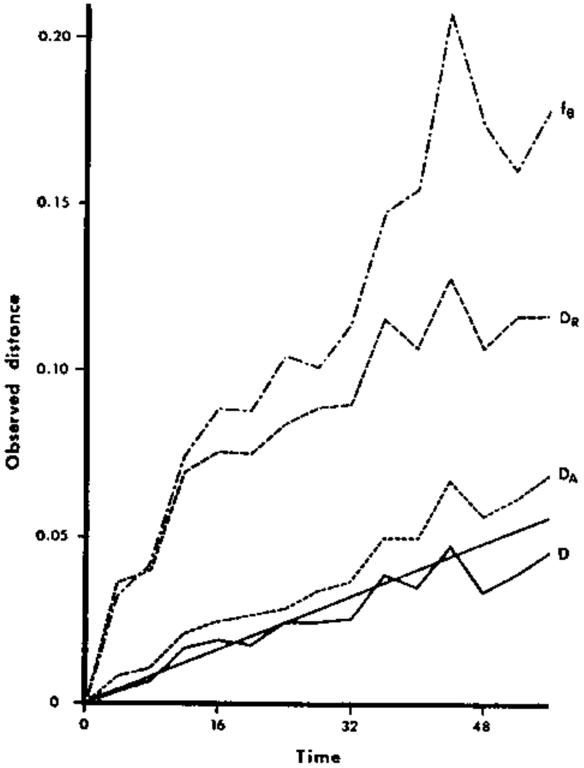

Fig. 8.6 Relationships between genetic distance and evolutionary time in one replication of computer simulation. $4Nv$ = 0.2, M = 0.004, and the number of loci used is 20. The straight line represents the expected value of D. The expectations of D_R, f_θ and D_A are not linear with time. Time is measured in the unit of $2N$ generations.

combination with D, gives the best estimates of branch lengths. In the case of $M = 0.004$ the $\bar{S}_E$ values are considerably smaller than those for $M = 0.1$, but essentially the same conclusion has been obtained (see Nei et al. 1983).

General Remarks

We have seen that both the topology and branch lengths of a reconstructed tree are often quite wrong unless a large number of loci are used. In the study of phylogenetic relationships of related species many authors have used 20–40 genetic loci. The study by Nei et al. (1983) indicates that even if 30 loci are used and M is as large as 0.1, some parts of a reconstructed tree are incorrect with a high probability. In their study only 8 OTUs were used because of the limited computer time available, but the error in reconstructed trees is expected to increase disproportionately as the number of OTUs increases (Tateno et al. 1982).

One important factor for determining the accuracy of a reconstructed tree is the branch lengths of the true tree. If there are many branches of which the true distances are as small as 0.004, the reconstructed tree is usually incorrect even if 100 loci are used. This result is discouraging, but we must accept it since it is due to the stochastic nature of gene substitution. Clearly, we cannot be overconfident about evolutionary trees reconstructed from electrophoretic data. Nevertheless, a large part of the topology of a reconstructed tree seems to be correct if 30 or more loci are used. In many cases even this approximate phylogenetic tree is useful for studying various evolutionary problems.

The accuracy of a reconstructed tree also depends on a tree-making method and distance measure used. In general, UPGMA, in combination with D, is the best among the three tree-making methods examined. Some authors have used Fitch and Margoliash's (1967) method for tree-making. This method usually requires more computer time, yet the efficiency of recovering the correct tree is not so high as UPGMA (Tateno et al. 1982).

It is interesting to see that the simple UPGMA, which was originally proposed for phenetic taxonomy, shows the best performance. The reason for this seems to be that the genetic distance based on a relatively small number of loci is subject to a large stochastic error, and the procedure of distance-averaging used in UPGMA reduces this error to a considerable extent. It should be noted, however, that this conclusion is only for electrophoretic data and does not necessarily apply to other types of data such as amino acid sequences for distantly related organisms (Blanken et al. 1982, Nei in press).

In the past many authors have used the Farris method because in this method the unequal rates of gene substitution in different branches can be taken into account. However, a large part of the seemingly different rates of gene substitution are apparently caused by stochastic errors in gene frequency changes. The Farris method cannot distinguish these stochastic errors from the true variation in substitution rate and thus is susceptible to errors in tree-making. Furthermore, in the presence of stochastic errors the Farris method often gives overestimates of branch lengths (Tateno et al. 1982). Nevertheless, the Farris method seems to be superior to UPGMA in obtaining a correct unrooted topology when the rate of gene substitution varies substantially with evolutionary lineage and stochastic errors are relatively small (Tateno et al. 1982).

Some numerical taxonomists (e.g., Farris 1981) claim that the genetic distance measures used in phylogeny construction should satisfy the triangle inequality. They give two arguments for this. First, when one wants to represent the species or populations concerned in a multidimensional space and measure the geometric distances between them, it is necessary to use a measure that obeys this principle. Second, if every estimate of genetic distance between OTUs represents the sum of the actual number of gene substitutions for all relevant branches of the true tree, then the triangle inequality should hold. Representation of populations in a multidimensional space is mathematically interesting, but it is not necessary for tree-making. Furthermore, the *geometric* distance between populations measured in this way is not proportional to the number of gene substitutions, and thus it is inappropriate for measuring *genetic* distance (Nei 1978b). (Genetic distance, as defined earlier, is the extent of gene difference between two populations.)

Their second argument looks reasonable at first sight, but it is not realistic. Although D is not a metric in individual cases, its expectation is a metric. Thus, if a very large number of loci are used, the genetic distance between any pair of OTUs will represent the sum of the numbers of gene substitutions for all relevant branches, at least theoretically. In practice, it is virtually impossible to examine hundreds or thousands of loci for phylogeny construction at the present time. Therefore, we must estimate the genetic distance from a smaller number of loci, and in the process of this estimation the metricity of D is disturbed by statistical errors. Nevertheless, it is possible to construct a reasonably good phylogenetic tree by using D, as shown here and previously by Nei et al. (1983). Actually metricity is not really required for tree-making so long as a proper distance measure and a proper tree-making method are used. In this connection it should be noted that usual estimates of nucleotide or amino acid substitutions are not metrics either, because they are estimated statistically by taking into account back mutations, parallel mutations, and multiple mutations, and essentially the same argument as the above applies to these estimates (see Tateno et al. 1982).

Recently, Farris (1981) criticized Sarich and Wilson's (1967) immunological distance and Nei's (1972) distance for their nonmetricity and claimed that any nonmetric distance would not show clocklike behavior. However, he did not consider that the molecular clock is stochastic rather than deterministic and subject to errors due to backward and parallel mutations. Furthermore, his criticism of D is based on gene frequency differences for one locus. In practice, D is designed to be used for many loci and should not be used for one locus (Nei 1972). Note also that, contrary to Farris's assumption, the amount of gene frequency difference between two populations is not proportional to evolutionary time whether there is selection or not. Only when the dynamics of gene frequency changes in populations is taken into account properly can one develop a distance measure that is useful in evolutionary studies. D has been developed exactly in this way.

Rogers's (1972) distance (D_R) has often been used in conjunction with the Farris method, because it satisfies the triangle inequality. The fact that this distance does not give negative branches when the Farris method is used seems to have been attractive to some workers. However, metricity of distance itself does not give any advantage in tree-making, as mentioned above. Just like D, this distance may occasionally decrease in the evolutionary process because of stochastic errors (Fig. 8.6), and thus D_R does not necessarily show the true genetic relationship among OTUs. Furthermore, as mentioned earlier, it is not linear with evolutionary time.

In addition to its nonlinear relationship with evolutionary time, D_R also has theoretical defect: it is not necessarily 1 even when the two populations concerned have no shared alleles. This occurs when the populations are polymorphic. For example, when there are five nonshared alleles in each population and all allele frequencies are equal, i.e., 1/5, we have $D_R = 1/\sqrt{(5)} = 0.45$. On the other hand, if the two populations are fixed for different alleles, D_R becomes 1. From the evolutionary point of view, this is a poor property. A similar property is observed with D_m.

In this section we have been concerned with the reconstruction of phylogenetic trees from gene frequency data. In recent years evolutionists (e.g., Brown et al. 1979, Avise et al. 1979b, Shah and Langley 1979) have started to use the restriction endonuclease technique to study the genetic differences between species or populations. In this case the number of nucleotide differences per nucleotide site can be estimated by the statistical methods of Nei and Li (1979), Kaplan and Langley (1979), Kaplan and Risko (1981), and Nei and Tajima (1983). The estimates obtained by these methods have a statistical property similar to that of Nei's D. Therefore, the conclusions obtained in this paper seem to apply to these estimates as well. In this case, however, we must use a large number of restriction endonucleases to make a reliable phylogenetic tree (Li 1986).