Genetic distances are useful for constructing phylogenetic trees of populations as well as for estimating times of divergence between populations. In the past, many investigators have used allele frequency data obtained by protein electrophoresis and immunological methods. In recent years, many different types of molecular data such as microsatellite DNA and random amplified polymorphic DNA (RAPD) data are used, but the basic principle of computing genetic distances and constructing phylogenetic trees is the same. Here, only the basic methods for computing genetic distances are discussed. Some results from recent studies of the evolution of human populations are also presented.

Commonly Used Distance Measures

Rogers’ Distance

Suppose that there are $q$ alleles at a locus, and let $x_i$ and $y_i$ be the frequencies of the $i$th allele in populations $X$ and $Y$, respectively. Each allele frequency may take a value between $0$ and $1$. Therefore, it is possible to represent populations $X$ and $Y$ in a $q$-dimensional space. The distance between the two populations in the space is then given by

$$D_R = \left[ \frac{1}{2} \sum_{i=1}^{q} (x_i - y_i)^2 \right]^{1/2}.$$  

This distance takes a value between $0$ and $\sqrt{2}$, the latter value being obtained when the two populations are fixed for different alleles. This property is not very desirable. Therefore, Rogers proposed the following measure, which takes a value between $0$ and $1$.  

$$D_R = \left[ \frac{1}{2} \sum_{i=1}^{q} (x_i - y_i)^2 \right]^{1/2}.$$  

When allele frequency data are available for many different loci, the average of this value over all loci is used. Note, however, that this measure has one deficiency. When the two populations are both polymorphic but share no common alleles, $D_R$ is given by

$$\left[ \frac{1}{2} \sum_{i=1}^{q} (x_i - y_i)^2 \right]^{1/2}.$$  

This value can be much smaller than $1$ even if the populations have entirely different sets of alleles.

Bhattacharya’s Distance and its Modifications

Representing two populations on the surface of a multidimensional hypersphere, Bhattacharya suggested that the extent of differentiation of populations be measured in terms of the angle ($\theta$) between the two lines projecting from the origin to the two populations ($X$ and $Y$) on the hypersphere. When there are $q$ alleles, we consider a $q$-dimensional hypersphere with radius $1$ and let each axis represent the square root of the allele frequency, that is, $\bar{x}_i = \sqrt{x_i}$ and $\bar{y}_i = \sqrt{y_i}$. Therefore,

$$\sum_{i=1}^{q} \bar{x}_i \bar{y}_i = 1.$$  

Elementary geometry shows that in the case of $q$ alleles the angle $\theta$ is given by

$$\cos \theta = \frac{\sum_{i=1}^{q} \bar{x}_i \bar{y}_i}{\sqrt{\sum_{i=1}^{q} \bar{x}_i^2 \sum_{i=1}^{q} \bar{y}_i^2}}.$$  

Bhattacharya proposed that the distance between two populations be measured by

$$\theta^2 = \left[ \arccos \left( \sum_{i=1}^{q} \sqrt{x_i y_i} \right) \right]^2 = \frac{1}{2} \sum_{i=1}^{q} \frac{(x_i - y_i)^2}{(x_i + y_i)}.$$  

This measure takes a value between $0$ and $1$. When there are allele frequency data for many loci, the average of this quantity is used as a genetic distance measure in the case of $D_R$. Using a similar concept of a multidimensional space, Cavalli-Sforza and Edwards proposed the following $d_C$ distance for a locus:

$$d_C = \frac{2}{\pi} \sqrt{1 - \cos \theta}.$$  

For multi-locus data, the average of this quantity for many loci is used. However, computer simulations have shown that the following distance measure is quite efficient in recovering the true topology of an evolutionary tree when it is reconstructed from allele frequency data.

$$D_A = \left( 1 - \frac{1}{L} \sum_{k=1}^{L} \sqrt{\sum_{l=1}^{q} (x_{ik} - y_{ik})^2} \right)$$  

where $q_k$ and $L$ are the number of alleles at the $k$th locus and the number of loci examined, respectively, and the subscript $lk$ refers to the $i$th allele at the $k$th locus. This measure takes a value between $0$ and $1$, the latter value being obtained when the two populations share no common alleles. Since the maximum value of $D_A$ is $1$, $D_A$ is nonlinearly related to the number of gene substitutions. When $D_A$ is small, however, it increases approximately linearly with evolutionary time. The standard error of $D_A$, or the difference in $D_A$ between two pairs of populations can be computed by the bootstrap method if it is based on many loci. In this case, a bootstrap sample will represent a different set of loci, which have been chosen at random with replacement. Similarly, the standard errors of average $D_R$, $\theta^2$, and $d_C$ can be computed by the bootstrap.
**F**<sub>ST</sub> Distance

The allele frequencies of different populations may differentiate by genetic drift alone without any selection. When a population splits into many populations of effective size *N* in a generation, the extent of differentiation of allele frequencies in subsequent generations can be measured by Wright's *F*<sub>ST</sub>. When there are only two populations but allele frequency data are available from many different loci, it is possible to develop a statistic whose expectation is equal to *F*<sub>ST</sub>. One such statistic is given by

\[ F_{ST} = \left[ \frac{(\bar{x}_X + \bar{x}_Y) / 2 - \bar{j}_{XY}}{(1 - \bar{j}_{XY})} \right] \]  

where \( \bar{j}_X, \bar{j}_Y, \) and \( \bar{j}_{XY} \) are unbiased estimates of the means \( j_X \) and \( j_Y \) of \( \sum x_i^2 \), \( \sum y_i^2 \), and \( \sum x_i y_i \) over all loci, respectively. For a single locus, unbiased estimates of \( \sum x_i^2 \), \( \sum y_i^2 \), and \( \sum x_i y_i \) are given by

\[ \bar{j}_X = \frac{2m_X \sum x_i^2 - 1}{(2m_X - 1)} \]  
\[ \bar{j}_Y = \frac{2m_Y \sum y_i^2 - 1}{(2m_Y - 1)} \]  
\[ \bar{j}_{XY} = \frac{\sum x_i y_i}{\bar{j}_X \bar{j}_Y} \]

where \( m_X \) and \( m_Y \) are the numbers of diploid individuals sampled from populations \( X \) and \( Y \), respectively, and \( x_i \) and \( y_i \) are the sample frequencies of allele \( A_i \) in populations \( X \) and \( Y \). Therefore, \( \bar{j}_X, \bar{j}_Y, \) and \( \bar{j}_{XY} \) are the means of \( j_X \), \( j_Y \), and \( j_{XY} \) over all loci, respectively. The expectation of \( F_{ST} \) is given by \( E(F_{ST}) = 1 - \frac{2}{N \ln 2} \), where \( N \) is the number of generations after population splitting. Therefore, we have

\[ D_t = -\ln(1 - F_{ST}) \]

which is expected to be proportional to \( t \) when the number of loci used is large \( E(D_t) = t / (2N) \). This indicates that when evolutionary time is short and new mutations are negligible, one can estimate \( t \) by \( 2ND_t \) if \( N \) is known. In practice, however, new mutations always occur, and this will disturb the linear relationship between \( D_t \) and \( t \) when a relatively long evolutionary time is considered. \( N \) is also usually unknown.

**Nei's Distance**

In 1972, Nei developed a genetic distance measure called the standard genetic distance, whose expected value is proportional to evolutionary time when both effects of mutation and genetic drift are taken into account. It is estimated by

\[ D = -\ln I \]

where

\[ I = \frac{\bar{j}_{XY}}{\sqrt{\bar{j}_X \bar{j}_Y}} \]

The variances of \( I \) and \( D \) can be computed by the bootstrap method.

When the populations are in mutation-drift balance throughout the evolutionary process and all mutations result in new alleles following the infinite-allele model, the expectation of \( D \) increases in proportion to the time after divergence between two populations. That is,

\[ E(D) = 2\alpha T \]

where \( \alpha \) is the rate of mutation or gene substitution per year and \( T \) is the number of years after divergence of the two populations. Therefore, if we know \( \alpha \), we can estimate the divergence time from \( D \).

The \( \alpha \) value varies with genetic locus and the type of data used. For the genetic loci that are commonly used in protein electrophoresis, it has been suggested that \( \alpha \) is approximately \( 10^{-7} \) per locus per year. If this is the case, the time after divergence between two populations is estimated by

\[ T = 5 \times 10^6 D \]

**(δρ)**<sup>2</sup> Distance

Microsatellite DNA loci are segments of repeated DNA with a short repeat length, usually two to six nucleotides. Thus, an allele for a CA repeat locus may be represented by CACACACACACA, where the dinucleotide CA is repeated 7 times. Microsatellite loci are believed to be subject to a mutational change following the slippage model of duplication or deletion of repeat units. Therefore, new alleles are supposed to be generated by following the stepwise mutation model. Microsatellite loci are usually highly polymorphic with respect to the number of repeats, and, therefore, they are useful for studying phylogenetic relationships of populations. In 1995, Goldstein proposed that the following distance measure be used for microsatellite DNA data:

\[ (\delta \rho)^2 = \sum_{k=1}^{L} (\mu_{SK} - \mu_{rk})^2 / L \]

where \( \mu_{SK} (=\Sigma x_{ik}) \) and \( \mu_{rk} (=\Sigma y_{rk}) \) are the mean numbers of repeats at the \( k \)th locus in populations \( X \) and \( Y \), respectively. The expectation of \( (\delta \rho)^2 \) is given by \( E((\delta \rho)^2) = 2\alpha T \), where \( \alpha \) is the mutation rate per year. Therefore, \( T \) can be estimated by \( (\delta \rho)^2 / (2\alpha) \).

In practice, however, there are a number of problems with this method. First, the \( \alpha \) value apparently varies considerably with locus and organism, and it is not a simple matter to estimate \( \alpha \) for each locus. Second, the variance or the coefficient of variation of \( (\delta \rho)^2 \) is very large compared to that of other distance measures such as \( d_c \) and \( D_Y \). Therefore, a large number of loci must be used to obtain a reliable estimate of \( T \) even if \( \alpha \) is known. Third, there is evidence that the actual mutational pattern is irregular and deviates considerably from the stepwise mutation model on which this distance measure is based.

**Genetic Distance and Phylogenetic Trees**

A linear relationship of a distance measure with evolutionary time is important for estimating the time of divergence between two populations. It is also a nice property for constructing phylogenetic trees, other things being equal. In practice, however, different distance measures have different variances, and for this reason a distance measure that is linear with time is not necessarily better than a nonlinear distance in obtaining true
trees (topologies). A number of authors have studied this problem by using computer simulation. The general conclusions obtained from these studies are as follows:

1. For all distance measures, the probability of obtaining the true topology \( P_t \) is very low when the number of loci used is less than 10, but gradually increases with increasing number of loci.
2. Distance measures \( D_A \) and \( d_C \) are generally more efficient in obtaining the true topology than other distance measures under many different conditions.
3. When the total number of individuals to be studied is fixed, it is generally better to examine more loci with a smaller number of individuals per locus than fewer loci with a large number of individuals in order to have a high \( P_t \) value, as long as the number of individuals per locus is greater than about 25.

**Evolutionary Relationships of Human Populations**

Figure 1 shows the phylogenetic trees obtained by using the \( D_A \) and \( (\bar{d}_P)^2 \) distances, respectively, from allele frequency data for 25 microsatellite loci. The tree obtained by \( D_A \) distances (Figure 1(a)) shows that Africans (Pygmies and Bantu) first separated from the rest of the human groups and that the bootstrap values for the interior branches connecting Africans and chimpanzees and non-Africans and chimpanzees are both very high. (A bootstrap value for an interior branch is an indicator of the accuracy of the presence of the branch.) This result supports the currently popular view that modern humans originated in Africa. The same tree shows that Europeans first diverged from the other non-African people and then the group of New Guineans and native Australians separated from the remaining group. The first separation of Europeans from the rest of non-Africans is well supported by a high bootstrap value, but the next separation of New Guineans and Australians is less clear, because the bootstrap value for one of the two interior branches involved is only 53%. To clarify this aspect of evolutionary relationships, it seems necessary to examine many more loci. Figure 1(b) shows the tree obtained by \( (\bar{d}_P)^2 \) distances. The topology of this tree is very different from that for \( D_A \) distances and is poorly supported by the bootstrap test. This unreliable tree was obtained mainly because the sampling error of \( (\bar{d}_P)^2 \) is very large, as mentioned earlier.

**Software**

The computation of distance values from allele frequencies and phylogenetic reconstruction can be made by using the software called "POPTREE2", of which the website is http://www.med.kagawa-u.ac.jp/~genomelt/takezaki/poptree2/index.html.

*See also:* Genetic Drift; Phylogeny; Trees.

**Further Reading**


**Relevant Websites**

http://www.med.kagawa-u.ac.jp – Faculty of Medicine, Kagawa University, POPTREE2.