24. Measures of Genetic Distance

Genetic distance is a fundamental concept in systematics and evolutionary biology:

– A genetic distance between two DNA sequences, two individuals, or two species (or other taxa) is a quantitative estimate of how divergent they are genetically.

– Genetic distances are important in understanding patterns of relatedness, patterns of migration and gene flow (interbreeding), rates of genetic change, phylogeny reconstruction, etc.

There are many different measures of genetic distances that have been used in studies of natural populations, but most can be classified into two categories: geometric and probabilistic. Until only recently, all distances measures used have been based on the infinite-alleles model of mutation and drift; however, genetic-distance measures for allozyme or microsatellite-DNA data should be based on the stepwise-mutation model.

1) The most widely used measure of geometric distance: Rogers’ (1972)

For a locus with $k$ alleles, let $p_{i,x}$ and $p_{i,y}$ be the frequencies of the $i$th allele in populations X and Y, respectively. Then

$$D = \sqrt{\frac{1}{2} \sum_{i=1}^{k} (p_{i,x} - p_{i,y})^2}$$

is the Euclidean distance between the populations, scaled to vary from 0–1 (0 = genetically identical, 1 = fixed for two different alleles). This is simply the straight-line distance between two points on a plot, of which the axes are the frequencies of the alleles and the points represent the two populations.

Examples for two alleles at one locus:
Note that, for two alleles (giving a two-dimensional space), the points representing different allele frequencies line on a straight line of slope –1. (Why?) Multiplying by the $\frac{1}{2}$ term within the square-root radical is equivalent to dividing the Euclidean distance by $\sqrt{2}$, which is the maximum distance between two points (at opposite corners of the allele space). This division scales the distance to lie between 0–1, inclusive.

A comparable measure of genetic similarity (or identity, $I$) is

$$ I = 1 - D. $$

2) The most widely used probabilistic distance, based on measures of inbreeding: Nei’s (1972)

For two randomly chosen alleles from the gene pool of $X$, the probability of their being identical is the sum of squared probabilities (assumed to be equal to the relative frequencies), summed across $k$ alleles (coefficient of inbreeding):

$$ j_X = \sum_{i=1}^{k} p_{i,X}^2. $$

For population $Y$, it is

$$ j_Y = \sum_{i=1}^{k} p_{i,Y}^2. $$

The probability of identity of one allele drawn from $X$ and another drawn from $Y$ is the sum of cross-products (coefficient of kinship):

$$ j_{XY} = \sum_{i=1}^{k} (p_{i,X} \cdot p_{i,Y}) $$

Because sums of squares and cross-products are additive, the $j_X$, $j_Y$, and $j_{XY}$ can be averaged across loci. In Nei’s notation,

$$ J_X = \overline{j_X}, \quad J_Y = \overline{j_Y}, \quad J_{XY} = \overline{j_{XY}} $$

where the means are across multiple loci.

Nei’s “normalized identity” of alleles between $X$ and $Y$,

$$ I = \frac{J_{XY}}{\sqrt{J_X J_Y}} $$
24. Genetic distance

is simply the product-moment correlation between alleles drawn randomly from X and Y, measure of the probability that two alleles drawn randomly from X and Y will be identical. It is a measure of genetic similarity rather than genetic distance. Nei proposed the genetic distance measure

\[ D = -\ln I, \]

which measures the accumulated number of gene substitutions per locus.

**Comparison of Rogers’ and Nei’s distances:** The following examples are for 2 alleles at 1 locus. The allele frequencies for population Y are held at \( p = 0, q = 1 \). The allele frequencies for population X vary from \( p = 0 - 1 (q = 1 - 0) \).

3) A recent probabilistic genetic-distance measure based on the stepwise-mutation model: Goldstein’s et al. (1995):

\[ D_1 = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{i,X} p_{j,Y} (i - j)^2 \]

where \( i \) and \( j \) are the “repeat scores” of two alleles, and \( p_{i,X} \) is the frequency of allele \( i \) in population X. This measure was proposed to be appropriate for microsatellite data, but is based on a model that was originally formulated for allozyme data.
Mutation-drift models:

Measures of genetic distance are based on particular **mutation-drift models** that predict the equilibrium resulting from the “birth” of alleles by mutation and the “death” of alleles via genetic drift:

(1) **Infinite-allele model** (Kimura and Crow, 1964)

- mutations occur at a constant (mean) rate, $\mu$, over time, identical for all loci;
- otherwise, Hardy-Weinberg-Castle conditions exist (random mating, no significant within-group structure);
- all alleles are selectively neutral (but the model can be adjusted to account for heterotic [selectively advantageous] combinations);
- new mutant alleles always different from existing ones in population (no reverse mutations);
- in the long run, most alleles lost via genetic drift;
- equilibrium attained when loss of alleles by random drift exactly balances the gain due to mutation;
- the equilibrium number of alleles maintained in a population is $n_e = 1 + 4N_e\mu$.
- Nei’s measure of genetic distance is linear with respect to time under this model:

$$D^e_s = -\ln \left[ J_{XY} / \sqrt{J_X J_Y} \right]$$

(2) **Stepwise-mutation model** (Ohta and Kimura, 1973)

- allelic states expressed by a series of integers ($\ldots, A_{-1}, A_0, A_1, \ldots$);
- all mutational changes are stepwise negative or positive (to one of the two adjacent states), but there is no limit to the number of possible alleles;
- “positive” and “negative” mutational events can cancel one another, leading back to the original state (i.e., allows for bidirectional mutation);
- equilibrium attained when loss of alleles by random drift exactly balances the gain due to mutation;
- the equilibrium number of alleles maintained in a population is $n_e = \sqrt{1 + 8N_e\mu}$.
- the Goldstein *et al.* measure of genetic distance is linear with respect to time under this model:

$$D^s_t = \sum_{i=1}^{k} \sum_{j=1}^{k} p_{i,x} p_{j,y} (i - j)^2 ;$$

- for measures of population subdivision analogous to F-statistics under this model, see Slatkin (1995);
Comparison of the two models:

![Graph showing comparison of two models: IA and SM](image)

**References:**


