Proc. Nat. Acad. Sci. USA
Vol. 71, No. 12, pp. 4847–4849, December 1974

Gene Frequency Comparisons Between Taxa: Support for the Natural Selection of Protein Polymorphisms

(evolution/population genetics/*Drosophila*/speciation/enzyme variation)

FRANCISCO J. AYALA* AND MICHAEL E. GILPIN†

* Department of Genetics, University of California, Davis, Calif. 95616; and † Department of Biology, University of California, La Jolla, Calif. 92037

Communicated by Theodosius Dobzhansky, September 13, 1974

ABSTRACT The hypothesis has been advanced that the pervasive protein variation found in natural populations of many organisms is adaptively neutral, and thus not subject to natural selection. This neutrality hypothesis predicts that at polymorphic gene loci different configurations of allelic frequencies will occur in different species. Results of an extensive study of protein variation in several species of *Drosophila* show that any two species have very similar allelic frequencies at a substantial proportion of all gene loci, while at many other loci the species have very different sets of alleles. Genetic distances have been calculated between pairs of subspecies, morphologically similar species, and morphologically different species. The distribution of genetic distances is strikingly different from the predictions of the neutrality theory. Protein variation appears to be maintained by natural selection.

Gel electrophoresis and selective enzyme assays allow identification of allelic variants at single genetic loci in single individuals. Although not all allelic variation is detectable (electrophoretic "alleles" are classes of alleles with equivalent ionic charge properties), these techniques make it possible to quantify, to a first approximation, the amount of genetic variation at a given locus. A large random sample of loci permits estimation of (*i*) the amount of genetic variation in natural populations of organisms and (*ii*) the amount of genetic differentiation between populations of the same or different species.

Gel electrophoresis and other techniques have uncovered a great deal of genetic variation in natural populations of outcrossing, sexually reproducing organisms (1–12). Since the genetic variation found is much larger than the amount that can be maintained by natural selection according to certain theoretical models, some authors have suggested that most or all genetic variation detected at the molecular level may be adaptively neutral (13, 14). Therefore, the variation would not be sustained by natural selection, but would rather be maintained as a neutral equilibrium subject to the "drift" caused by the sampling errors of reproduction.

This hypothesis of neutrality leads to predictions that are amenable to empirical tests. Some predictions have been examined elsewhere (6–12). Here we shall examine a prediction concerning the distribution of allelic frequencies in different species.

Consider a diploid population of effective size $2N$, which at time $t = 0$ is divided equally into two isolated populations that remain constant, over time, at size N. At a given locus with k alleles (or k classes of electrophoretic "alleles"), denote the frequency of the ith allele by X_i, such that

$$\sum_{i=1}^{k} X_i = 1, X_i \geq 0. \quad [1]$$

Denote the initial state of the locus by the k element vector, $\bar{\mathbf{X}}$. If there is no selection and no mutation, the state of the locus in either population will "drift" as a result of the sampling error that occurs between generations. The probability distribution for any genetic state $\mathbf{X}$ of either population at a future time t, denoted by the distribution function $\Phi(\mathbf{X}, \bar{\mathbf{X}}, t)$, can be obtained by solving the Fokker–Planck diffusion equation with the initial condition

$$\Phi(\mathbf{X}, \bar{\mathbf{X}}, 0) = \delta(\mathbf{X} - \bar{\mathbf{X}}), \quad [2]$$

where δ is the Dirac delta function. It is impossible to obtain this function in closed form, but for most applications using the first few terms in the power series solution is sufficient.

$\Phi(\mathbf{X}, \bar{\mathbf{X}}, t)$ is the same for both populations, but the actual state of either population depends on chance, and will almost certainly be different for the two populations. Assume that the distance between the states of the two populations, $\mathbf{X}_1$ and $\mathbf{X}_2$, is measured by some function $D(\mathbf{X}_1\ \mathbf{X}_2)$. The argument to follow is quite insensitive to the exact form of this function, so we shall define D as the normalized length of the difference of these two vectors:

$$D(\mathbf{X}_1, \mathbf{X}_2) = ||\mathbf{X}_1 - \mathbf{X}_2||/\sqrt{2} = \sqrt{\sum_{i=1}^{k} (X_{1i} - X_{2i})^2/2}, \quad [3]$$

where X_i represents the ith allelic component of the genetic vector $\mathbf{X}$, and where $\sqrt{2}$ is a normalization constant. Nei's measure of genetic similarity would work equally well (15).

The expected distribution of the distance, d $(0 \leq d \leq 1)$, between the two populations can be formally expressed as

$$f(d, \bar{\mathbf{X}}, t) = \int \Phi(\mathbf{X}_1, \bar{\mathbf{X}}, t)\delta[D(\mathbf{X}_1, \mathbf{X}_2) - d] \times \Phi(\mathbf{X}_2, \bar{\mathbf{X}}, t)d\mathbf{X}_1 d\mathbf{X}_2. \quad [4]$$

This integral can only be evaluated numerically. A typical solution is represented in Fig. 1. For a three-allele locus whose initial state $\bar{\mathbf{X}}$ was assumed to be (1/3, 1/3, 1/3), $f(d, \bar{\mathbf{X}}, t)$ is drawn for $t = 0.1N$, $0.2N$, $0.5N$, and N generations. (N is the number of breeding individuals in each population.) Fixations do not occur within these times, but as t advances beyond N generations, the distribution function for states, Φ, starts to "flatten out" and first one, then two, alleles are lost, so that

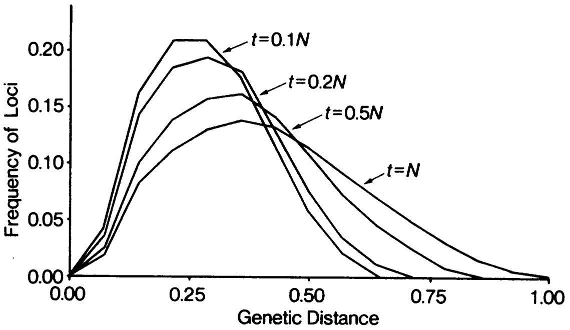

FIG. 1. Distribution of genetic distance between taxa according to the neutrality hypothesis. Using a digital computer, all permutations of the ordered pair ($\mathbf{X}_1$, $\mathbf{X}_2$), where $\mathbf{X}_1$ and $\mathbf{X}_2$ are possible genetic states at a common locus for two different populations, are generated. For each permutation, the distance between the elements of the pair, $D(\mathbf{X}_1, \mathbf{X}_2)$ from Eq. **3**, is calculated. This distance is then weighted by the product of the probabilities of occurrence of $\mathbf{X}_1$ and $\mathbf{X}_2$, which are given by the solution of the diffusion equation, $\Phi(\mathbf{X}_1, \bar{\mathbf{X}}, t)$. This weight is then added into the distance interval into which $D(\mathbf{X}_1, \mathbf{X}_2)$ falls. Since the probability of occurrence of $\mathbf{X}_1$ and $\mathbf{X}_2$ depends on time, the distribution function of the distances, $f(d, \bar{\mathbf{X}}, t)$, also depends on time. This distribution is drawn for $t = 0.1N$, $0.2N$, $0.5N$, and N generations, where N is the effective number of breeding individuals per taxon.

the populations tend to become fixed with equal probability at one of the states (1, 0, 0), (0, 1, 0), or (0, 0, 1). The distance distribution then consists of one-third of the cases in the class $d = 0$ and two-thirds in the class $d = 1$. More generally for a symmetrical $\bar{\mathbf{X}}$ with k alleles, as $t \rightarrow \infty$ the distribution function of distances would consist of $1/k$ zeros and $(k - 1)/k$ ones. Mutation makes boundary states reflecting and greatly reduces the $d = 0$ class for large values of t.

It should be noted in Fig. 1 that the probability of $d = 0$ is vanishingly small, and that the mode of the distribution occurs between 0 and 1. This is a quite general result; and it can be seen intuitively. Consider a region in the genetic state space away from any boundaries and where Φ is relatively flat; that is, where all combinations of allelic frequencies have about equal probability. The dimensionality of this space is $k - 1$. The number of states at a distance between d and $d + \Delta d$ increases with d as the surface area of a $k - 1$ dimensional hypersphere increases with its radius, i.e., as d^{k-2}. Since all available states will tend to be occupied with equal probability, there will be many more cases of intermediate distance than of identity.

The foregoing is subject to the criticism that population sizes do not remain constant. In particular, at the time of separation ($t = 0$), one of the two populations is likely to be small, which produces the "Founder Effect" (16). Or at some later time either one of the populations could become small, producing a "bottleneck." As long as population sizes do not become exceedingly small, i.e., on the order of a hundred individuals for the whole species, this will cause no serious problems. The diffusion processes will change speeds, and it will be impossible to relate time to the population size. Nevertheless, the states through which the distance distribution function passes will remain unchanged. That is, starting from a peak at $d = 0$, the distance distribution will move through a succession of bell-shaped forms becoming lower and wider, eventually peaking at $d = 1$.

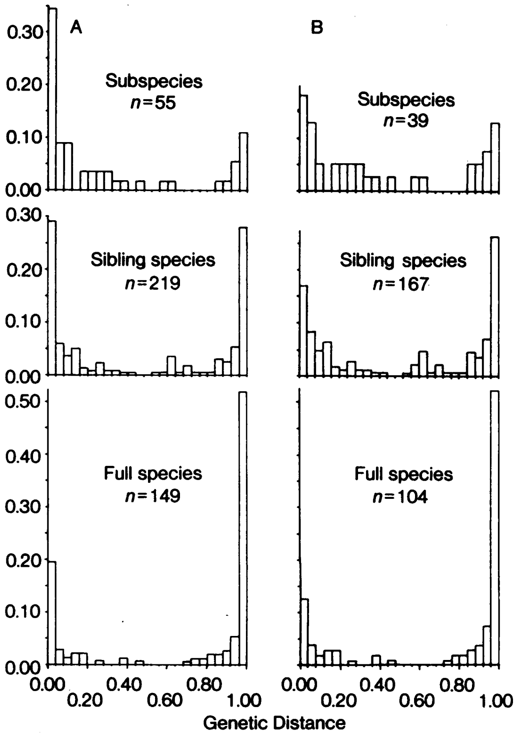

FIG. 2. (A) Distribution of loci relative to genetic distance between subspecies (upper), sibling species (middle), and nonsibling species (lower) of the *Drosophila willistoni* group. n is the number of pairwise comparisons. (B) The same distributions as in (A) after removing all nearly monomorphic loci, i.e., all loci with the frequency of the most common allele ≥ 0.95 in the two taxa being compared.

In order to test the predictions derived from the neutrality hypothesis, we shall utilize data from our electrophoretic study of genetic variability in several species in the *Drosophila willistoni* group (6–12). We shall be concerned here with three sibling (morphologically very similar) species: *D. willistoni*, *D. equinoxialis*, and *D. tropicalis*; one nonsibling species: *D. nebulosa;* and two pairs of subspecies: *D.w. willistoni–D.w. quechua* and *D.e. equinoxialis–D.e. caribbensis*. On the average, more than 30 natural populations, and more than 2000 wild genomes, have been sampled per species. We assayed 36 gene loci, 30 of them common to all species. The mean number of electrophoretic alleles per locus found in the whole group is 10.2 ± 0.6; per species, this statistic is 5.7 ± 0.2. Since little genetic differentiation exists among local populations of a given taxon (6–12), the genetic distance between taxa has been measured using the allelic frequencies observed in the whole taxon.

Fig. 2A gives the distribution of genetic distance (calculated by Eq. **3**) for all pairwise comparisons between taxa at a given level. The distribution is U-shaped in all cases. That is, more than half of the distances are either 0 or 1, and the remainder are spread rather evenly in the middle states between 0.04 and 0.96. The only difference between the three cases is that the full species have the greatest number of distances in the $d = 1$ class, the subspecies the least. Thus, there is a positive correlation between the degree of phylogenetic divergence and the probability of the $d = 1$ class. Comparison of the distribution of d among the three taxonomic levels indicates that loci move from the class $d = 0$ to the class $d = 1$ relatively fast, rather than slowly as predicted by the neutrality hypothesis.

The contrast between the observed U-shaped distribution and the predictions from the neutrality hypothesis is striking. This is not due to averaging the within-taxa genetic frequencies, since the distributions of d are much the same when comparisons are made between *local* populations of different taxa. Furthermore, the distributions of d for *each* pairwise comparison within a given taxonomic level have the same form as the *average* distributions shown in Fig. 2A, although different sets of loci fall in the two extreme classes of d when different taxa are compared (11, 12).

The frequency of the $d = 0$ class is perhaps the clearest deviation from the neutrality hypothesis, which offers only two possible explanations for it. The class $d = 0$ would be frequent if the number of generations since genetic isolation were very much less than the effective population size, but this leaves unexplained the high frequency of the class $d = 1$.

The alternative explanation is that the number of generations since genetic isolation is very much greater than the effective size, in which case all taxa would have degenerated to monomorphic states. If the number of alleles were small, some populations would arrive at the same monomorphic state, accounting for the $d = 0$ class, while the majority would arrive at alternative states, accounting for the $d = 1$ class. But the *Drosophila* populations studied are very polymorphic, not monomorphic. The average proportion of polymorphic loci per population for these species is 50.9% (a locus is considered polymorphic when the frequency of the most common allele is no greater than 0.95); about 17.7% of the loci are heterozygous in an average individual; and the mean number of electrophoretic alleles per locus is about 10 (6–12). In any case, Fig. 2B shows the distribution of d for each taxonomic level after the "monomorphic" loci (i.e., those whose most common allele has a frequency of 0.95 or higher in the two taxa compared) have been removed. The distribution is little changed and is still U-shaped.

The neutrality hypothesis, which contends that genetically controlled molecular differences between species (and other taxa) are the result of random processes, fails to explain the empirical results reported here. A considerably more consistent explanation is possible with natural selection. It appears that some form of normalizing selection maintains quasi-stable polymorphic equilibria at many of the loci that code for enzymes, and that changes in genetic background or ecological niche can trigger quite rapid changes to quite different polymorphic equilibria.

To visualize this, consider a modified form of Sewall Wright's metaphor of an adaptive landscape. For a locus, construct a k dimensional space such that each axis measures the frequency of one of the distinct alleles that occurs in one or more of the taxa under study. The vector $\mathbf{X}_j$ describes the genetic state of the jth taxon in this space; its tip lies on the surface of the $k - 1$ dimensional hyperplane defined by Eq. **1**. Our data show that state vectors for different taxa tend to cluster at only a few points on this rather "wide" surface. These points may be thought of as adaptive foci. The states of taxa spend a relatively long time on these foci, accounting in part for the $d = 0$ class, and then move relatively rapidly to different and "orthogonally" located foci, accounting for the $d = 1$ class. Different species may arrive at the same focus independently. This can be seen since a pair of taxa will have loci with $d = 0$ and with $d = 1$, but the loci in these classes are different when different pairs of taxa are compared.

In conclusion, our data support the hypothesis that natural selection determines the frequencies of electrophoretic alleles, and thus the distribution of the genetic distances between taxa. As pointed out above, the arguments presented in this paper do not assume that alleles coding for enzymes with similar electrophoretic mobilities are identical. The observed distributions of genetic distances are incompatible with the neutrality hypothesis, even if it is assumed that each electrophoretic "allele" comprises several alleles which may or may not be identical in different species.

We thank Profs. Th. Dobzhansky and W. W. Anderson for criticism of the manuscript. This work was supported by grants from the U.S. National Science Foundation and the U.S. Public Health Service.

1. Lewontin, R. C. & Hubby, J. L. (1966) *Genetics* **54**, 595–609.
2. Harris, H. (1966) *Proc. Roy. Soc. Ser. B* **164**, 298–310.
3. Lakovaara, S. & Saura, A. (1971) *Genetics* **69**, 377–384.
4. Richmond, R. (1972) *Genetics* **70**, 87–112.
5. Selander, R. K. & Kaufman, D. W. (1973) *Proc. Nat. Acad. Sci. USA* **70**, 1875–1877.
6. Ayala, F. J., Powell, J. R. & Dobzhansky, Th. (1971) *Proc. Nat. Acad. Sci. USA* **68**, 2480–2483.
7. Ayala, F. J., Powell, J. R., Tracey, M. L., Mourao, C. A. & Pérez-Salas, S. (1972) *Genetics* **70**, 113–139.
8. Ayala, F. J. (1972) *Proc. Sixth Berkeley Symp. Math. Stat. Prob.* **V**, 211–236.
9. Ayala, F. J., Powell, J. R. & Tracey, M. L. (1972) *Genet. Res.* **20**, 19–42.
10. Ayala, F. J. & Tracey, M. L. (1973) *J. Heredity* **64**, 120–124.
11. Ayala, F. J. & Tracey, M. L. (1974) *Proc. Nat. Acad. Sci. USA* **71**, 999–1003.
12. Ayala, F. J., Tracey, M. L., Barr, L. G., McDonald, J. F. & Pérez-Salas, S. (1974) *Genetics* **77**, 343–384.
13. King, J. L. & Jukes, T. H. (1969) *Science* **164**, 788–798.
14. Kimura, M. & Ohta, T. (1971) *Nature* **229**, 467–469.
15. Nei, M. (1972) *Amer. Natur.* **106**, 283–292.
16. Mayr, E. (1963) *Animal Species and Evolution* (Belknap Press, Cambridge, Mass.).