

# Genetic distance and phylogenetic relationship

MARIA R. MENEZES

*National Institute of Oceanography, Dona Paula, Goa - 403004, India*

## Introduction

Biological evolution consists of two processes: anagenesis or phyletic evolution and cladogenesis or splitting (Avisé, 1976). Anagenetic change is gradual and usually results from increasing adaptation to the environment. A favourable mutation or other genetic change arising in a single individual may spread to all descendants by natural selection. Cladogenesis results in the formation of independent evolutionary lineages. Favourable genetic changes arising in one lineage cannot spread to members of other lineages. Cladogenesis is responsible for the great diversity of the biological world, allowing adaptation to a great variety of ways of life. The most decisive cladogenetic process is speciation.

The speciation process requires the development of reproductive isolation between populations, resulting in independent gene pools. For sexually reproducing organisms, isolation by geographic barriers accompanied by severe restriction of gene exchange is the usual prerequisite to genetic divergence and speciation. Geographically isolated populations accumulate genetic differences as they adapt to their different environment. They may become recognizable as races. Not all races will become species, however, because the process of geographic differentiation is reversible. If the races have not sufficiently diverged while allopatric, they may later converge or fuse through hybridization. On the other hand, allopatric populations may sometimes become sufficiently different genetically, so that if the opportunity for gene exchange ensues again, hybrids will have low fitness. Natural selection would, then, favour the completion of reproductive isolation.

In the study of evolution it is of great importance to know the number of gene differences or the number of genes identical between a pair of species. Such knowledge should provide an answer to the question of how many gene substitutions are necessary for two descendant lines to be recognised as new species (Nei, 1971). Evolutionary changes in the structure of biological macromolecules map the stream of heredity that is phylogeny (Simpson, 1945).

In recent years, several molecular techniques have been developed which allow comparison of genetic material in different populations. These techniques include nucleotide and amino acid sequencing, DNA-DNA and DNA-RNA hybridization, immunological techniques and protein electrophoresis (Avisé, 1976). These techniques permit quantification of genetic differences between species based on a sample of the genome selected without *a priori* knowledge of whether the genes sampled are more or less variable than the rest of the genetic material. For the purpose of comparing closely related species, protein electrophoresis has proved to be the most efficient technique (Avisé, 1975).

## Electrophoretic techniques

When proteins in blood or extracts of other tissues or of whole organisms are allowed to migrate in a supporting medium (for example, a slab of starch) under the influence of an electric field, differences in net electrostatic charge are reflected in different mobilities. Following electrophoresis, the position of specific enzymes, such as phosphoglucosmutase, or nonenzymatic proteins (e.g., sarcoplasmic protein) can be demonstrated by specific histochemical stains



**Figure 1.** Results of electrophoresis of the enzymes phosphoglucosmutase (a) and isocitric dehydrogenase (b) from the skeletal muscle of a sciaenid fish *N. mitsukurii*. The gene that codes for the enzyme is Pgm -1 and Idh -1 respectively. In this population, some individuals are homozygous for an allele (S) corresponding to a slow-migrating enzyme, some are homozygous for an allele (F) corresponding to a fast-migrating enzyme, and the rest are heterozygous F/S. The genotypes of the individuals are indicated beneath the enzyme bands. The enzyme, Pgm, is a monomer, so the heterozygous exhibit two enzyme bands of differing mobility. Idh is a dimeric enzyme, so the heterozygous exhibit three enzyme bands: rapidly migrating (fast + fast) and slowly migrating (slow + slow) homodimers, like those produced in the corresponding homozygotes, and a fast + slow heterodimer, which typically is found to have an intermediate electrophoretic mobility.

applied directly to the electrophoretic medium. The resulting phenotypes are bands of stain indicating regions of enzymatic activity or protein concentration. Enzyme variants (allozymes) offer many advantages for the investigation of microevolutionary processes, because they generally have a simple genetic basis (Koehn *et al.*, 1983) and their phenotypes (i.e., catalytic properties) can be objectively and accurately quantified (Fig. 1a and b).

The overall genetic differentiation between populations may be reasonably estimated by examining a number of proteins. However, potential sources of bias must be recognized. One possible source of bias is that electrophoretic methods sample only structural genes which code for soluble enzymes and proteins. Other classes of structural genes, as well as regulatory genes whose products are not translated into proteins, cannot be sampled by electrophoresis. Another difficulty is that not all changes in DNA will be registered in an altered protein mobility because of (1) the degeneracy of the genetic code, (2) the substitution of amino acids with identical net electric charge, and (3) the finite number of distinguishable band mobilities on a gel. Although we do not know how much genetic variation remains undetected by electrophoretic techniques, the results obtained provide only minimal estimates of genetic differentiation.

### Estimation of alleles frequencies, genetic distance and phylogenetic divergence

Electrophoretic data consist of allele and genotype frequencies determined by individually assaying members of populations (see Menezes, 1990). The allele frequency of a prescribed allele among a group of individuals is simply the proportion of all alleles at the locus that are of the prescribed type (Gale, 1980). The frequency of any prescribed allele in a sample is therefore equal to twice the number of homozygotes for that allele (because each homozygote carries two copies of the allele) plus the number of heterozygotes for that allele (because each heterozygote carries one copy) divided by two times the number of individuals in the sample (because each individual carries two alleles at the locus). Among the sample of 20 individuals in the gel shown in Fig. 1a for example, there are 10 F/F homozygotes, 9 F/S heterozygotes and 1 S/S homozygote. The allele frequency of F is therefore  $(2 \times 10 + 9) / (2 \times 20) = 0.725$  whereas the allele frequency of S is  $(2 \times 1 + 9) / (2 \times 20) = 0.275$ . Note that the sum of all allele frequencies, in this case those for F + S, must be 1.

Cumulative comparisons among loci between taxonomic groups can be summarized by a variety of methods, into indices of similarity (I) or conversely, genetic distance (D). Genetic distances reflect estimates of net codon differences between pairs of populations. Thus, a large value for I and a correspondingly small value for D between two populations would imply a close phenetic relationship. According to the method proposed by Nei (1972) the normalized genetic identity of genes between two populations at the j locus is defined as:

$$I_j = \frac{\sum x_i y_i}{(\sum x_i^2 + \sum y_i^2)^{1/2}}$$

where  $x_i$  and  $y_i$  represent the frequencies of the  $i$ th allele in populations x and y, respectively. For all loci in a sample, the overall genetic identity of x and y is defined as:

$$I = \frac{J_{xy}}{(J_x J_y)^{1/2}}$$

where  $J_x$ ,  $J_y$  and  $J_{xy}$  are the arithmetic means over all

loci of  $Exi^2$ ,  $Eyi^2$  and  $Exi Eyi$ , respectively. The accumulated number of codon substitutions per locus (genetic distance) since the time of divergence of two populations is estimated as:

$$D = -\log_e I$$

Another commonly employed index,  $S$ , measures the mean geometric distance between allele frequency vectors over all loci (Rogers, 1972). Estimates of  $I$  and  $S$  calculated from the same data are fairly similar, although  $S$  generally gives lower numerical values than  $I$ .

Nei (1975) proposed a formula for estimating the phylogenetic divergence time ( $t$  years) from the genetic distance ( $D$ ); that is:

$$t = \frac{D}{2cnt \lambda a},$$

where  $c$  represented the proportion of amino-acid substitution which could be detected by electrophoresis,  $nt$  the total number of amino-acids for synthesis of an average protein, and  $\lambda a$  the average rate of amino-acid substitution per site per year. He gave the values, 0.25, 400 and  $1 \times 10^{-9}$  for  $c$ ,  $nt$  and  $\lambda a$ , respectively, leading to the equation:

$$t = 5 \times 10^6 D$$

The above estimations of genetic distance and divergence time are based on the assumption that the amino-acid substitution rate is constant among different proteins.

A number of algorithms can be employed to convert genetic data into estimates of phylogenetic tree (Avice, 1976). The particular algorithm chosen depends upon the goals and assumptions of the analysis, and upon the nature of the genetic information available. If comparisons are made among many organisms (operational taxonomic units or OTU's), a distance matrix is generated. The unweighted pair - group clustering method (UPGMA; Sneath and Sokal, 1973; Nei, 1975) is by far the simplest, conceptually

and computationally. OTU's exhibiting the smallest genetic distance in a matrix are joined first to form a "cluster". Successive scans of the matrix identify next — smallest mean distances, and these form the basis for adding new OTU's to the growing tree, and for joining together clusters previously generated.

## Discussion

A substantial amount of genetic differentiation may arise either by many loci that have a moderate amount of differentiation, or by complete differentiation at only a few loci, with great similarity or identity at the other loci studied. For taxonomic purposes, these two situations are quite distinct. A locus at which complete differentiation exists between two populations can be used to diagnose the population to which an individual belong. A locus at which only partial differentiation occurs cannot be so used. Several loci at which two populations are partially different can be jointly used for diagnostic purposes, but statistical manipulations are more complex (Ayala, 1983).

Relationships between standard genetic distances observed for various taxonomic categories are about as expected. Higher categories are on the average more different than the lower ones. Nei (1975) estimates that for local races of a species,  $D$  ranges from nearly 0 to 0.05, for subspecies  $D = 0.02$  to 0.2. (The distinction between local races and subspecies is somewhat arbitrary; a race is usually defined as a group of populations of a species that are geographically separated from other such groups and that differ in allele frequencies. If races differ enough in easily observable characteristics, they are often given Latin names and formally recognized as subspecies). For full species,  $D = 0.10$  to 2.0; sibling species, which are morphologically nearly identical, generally have  $D$  values two or three times smaller than the  $D$ 's between morphologically distinct species. For genera,  $D$  is usually greater than 1. It has been proposed (Thorpe, 1979, 1982) that taxa at different levels of systematic divergence will have different probabilities of having any given value of genetic identity. Thorpe (1979) suggested that  $I$  value (Nei, 1972) between most pairs of conspecific populations would be above 0.9, whilst

between congeneric species most would be between 0.25 and 0.85. It was predicted that between species of different genera many I values would fall below 0.3.

One of the interesting findings from studies of genetic distance is that humans and chimpanzees, while taxonomically assigned to different families have  $D = 0.62$  ( $t = 3.10$  million years), a value typical of that between species and even exceeded by some sibling species in *Drosophila* (King and Wilson, 1975). Such a small value of genetic distance between humans and chimpanzees does not adequately reflect their extensive morphological divergence, but comparisons of the amino acid sequences of various proteins, as well as studies of protein differences by immunological methods, fully support the remarkable similarity of structural genes between the species. King and Wilson (1975) have therefore suggested that evolution at the level of morphology may occur by means of changes in regulatory genes rather than by changes in structural genes; indeed they suggest that major evolutionary adaptations of all types may involve regulatory genes.

The role of regulatory gene changes in adaptive evolution is difficult to assess at present, partly because appropriate techniques for the large-scale study of regulatory genes have not yet been developed and partly because the concept of "regulatory gene" is still somewhat fuzzy and not operationally defined (Hartl, 1980).

Many studies exist on genetic differentiation between vertebrate populations in early stages of evolutionary divergence (see Ayala, 1975). The sunfish genus *Lepomis* contains 11 species, all native to North America. These species are renowned for their ability to hybridize, particularly in disturbed ecological conditions or in the laboratory. F1 hybrids from 21 different combinations of two species have been found in nature. Nevertheless, the species retain their identity throughout their ranges, even where they are sympatric. Adults of all species can readily be distinguished morphologically. Avise and Smith (1974a,b) have studied electrophoretic variation in proteins encoded by 14 genetic loci in 10 *Lepomis*

species. The mean genetic distance between *Lepomis* species is  $D = 0.627 \pm 0.029$  and  $I = 0.544 \pm 0.016$ . Despite their ability to hybridize, *Lepomis* species are very distinct genetically — about 50 per cent of loci are completely distinct in allelic composition in a typical comparison between two species. This is about as much genetic differentiation as between sibling species of *Drosophila willistoni*.

*Lepomis macrochirus* contains two morphologically different subspecies in the southeastern United States. *Lepomis macrochirus macrochirus* and *L. m. purpurescens* are largely allopatric, but they do meet and hybridize in an extensive intergrade zone, which probably represents a secondary contact between allopatrically evolved population. Little genetic differentiation exists between population with a subspecies ( $D = 0.024$ ). The genetic distance between the subspecies *L. m. macrochirus* and *L. m. purpurescens* is  $D = 0.171$ , a value close to those observed between *Drosophila willistoni* subspecies or semi-species. F1 hybrids between *L. m. purpurescens* and *L. m. macrochirus* are fertile. Considerable genetic differentiation has preceded the development of reproductive isolation in these subspecies.

In Sciaenidae (Menezes and Taniguchi, 1988; Menezes *et al.*, 1990) the average genetic distance based on 23 loci (Table 1) varied between 0.0006 to 0.0018 in *Pennahia argentata* from 6 locations, 0.001 to 0.0093 in *Nibea albiflora* from 3 locations and was 0.0067 in *Nibea mitsukurii* from 2 locations. The genetic relationships between these subpopulations are shown by a dendrogram in Fig. 2. The low D values in *P. argentata* suggests the more frequent migration/gene flow between these subpopulations in the early developmental (planktonic) stage. This is supported by the fact that this species has its habitat and spawning ground more offshore than the other two and that it has a wide distribution (Taniguchi, 1982). In *N. mitsukurii* population, the genetical discontinuity between Tosa bay and Atsumi bay suggests a barrier preventing gene flow between these two localities. This result agreed with an earlier report (Taniguchi, 1980) in which it was suggested that the cold water mass occurring in the area off Cape Shiono Misaki and

the deep-sea off Kii channel may be a functional barrier against the gene flow between Tosa Bay and Kumano Basin subpopulations. The Ariake Sea subpopulation of *N. albiflora* seemed to be well isolated from the other subpopulations such as those of the East China Sea and the Inland Sea. Genetic distance is much larger in the Ariake Sea subpopulation compared to the other subpopulations (Fig. 2). The mean genetic distance ( $D = 0.0094$ ) is almost equal to the value of 0.01 that is indicative of a local race (Nei, 1972). The differentiation of Ariake Sea subpopulation of *N. albiflora* from the other subpopulations may have been accelerated by the closed nature of the sea for a long geologic time.

The average D value among species (*N. mitsukurii* and *N. albiflora*) belonging to the same genus was found to be 0.092 i.e. on average, about 9.2 electrophoretically detectable allelic substitutions for every 100 loci have occurred between these two species since their divergence from a common ancestral population. Although these two species are

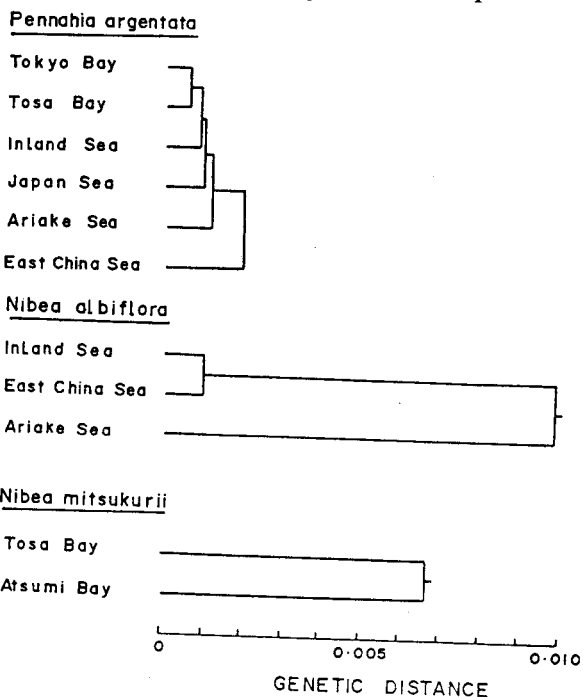


Figure 2. Dendrogram showing the relationship between *P. argentata*, *N. albiflora* and *N. mitsukurii* from different localities.

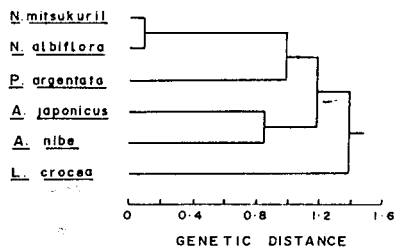


Figure 3. Dendrogram showing the relationship among the 6 sciaenid species based on values of genetic distance.

considered to be sibling species principally having allopatric distribution there seems to be introgressive hybridization in the waters around Kii channel where *N. mitsukurii* populations are faced with *N. albiflora* populations. This hypothesis is supported by findings of individuals with intermediate body color pattern between the two species.

Among genera belonging to the same subfamily average D was 1.212 (i.e. on average, about one electrophoretically detectable allelic substitution per locus has taken place), ranging from 0.861 to 1.668 (Table 2). The genetic relationship between them is given in Fig. 3. From this dendrogram, the six species were clearly divided into two distinct groups. One of the group included only *Larimichthys crocea*, while the other group, consisting of the remaining 5 species, could be further divided at lower D levels into the *Nibeia* - form (*N. mitsukurii* and *N. albiflora*), the *Pennahia* - form (*P. argentata*) and the *Argyrosomus* - form (*Argyrosomus* and *Atrobucca*). The *Pennahia* - form was found to be closer to the *Nibeia* - form, while the *Argyrosomus* - form was intermediate between the *Pennahia* - form and *Pseudosciaena* - form (*Larimichthys* - form). Evidence obtained from the biochemical genetic study suggests that the *Pseudosciaena* - form evolved independently and supports the result obtained from morphological studies.

The genetic distances thus give supplemental but important evidence pertaining to taxonomic problems on the recognition of species or subspecies which cannot be resolved by conventional methods. Building up of a systematic arrangement more consistent with the genetic relationships of organisms is expected as

**Table 1.** Genetic distance (above the diagonal) and divergence time (year) (below the diagonal) between the 3 sciaenid species.

1. *Pennahia argentata*

Locality	1	2	3	4	5	6
1. Tokyo Bay		0.0006	0.0009	0.0009	0.0009	0.0017
2. Tosa Bay	3000		0.0008	0.0010	0.0011	0.0018
3. Inland Sea	4500	4000		0.0011	0.0011	0.0026
4. Japan Sea	4500	5000	5500		0.0018	0.0015
5. Ariake Sea	4500	5500	5500	9000		0.0025
6. China Sea	8500	9000	13000	7500	12500	

2. *Nibea albiflora*

3. *Nibea mitsukurii*

Locality	1	2	3	Locality	1	2
1. Inland Sea		0.0092	0.0010	1. Tosa Bay		0.0067
2. Ariake Sea	46500		0.0091	2. Atsumi Bay	33500	
3. China Sea	5000	45500				

**Table 2.** Estimates of genetic distances (above the diagonal) and similarity (below the diagonal) of the 6 sciaenid species.

Species	1	2	3	4	5	6
1. <i>Nibea mitsukurii</i>		0.092	1.460	1.080	1.085	1.668
2. <i>Nibea albiflora</i>	0.912		1.267	0.946	1.004	1.441
3. <i>Argyrosomus japonicus</i>	0.232	0.282		1.251	0.861	1.428
4. <i>Pennahia argentata</i>	0.339	0.388	0.286		0.975	1.271
5. <i>Atrobucca nibe</i>	0.338	0.366	0.423	0.377		1.234
6. <i>Larimichthys crocea</i>	0.189	0.237	0.240	0.281	0.291	

a major result of this new approach (Bullini, 1983). Selander (1982) concludes that "the continuing study of the structure of the gene at the most fundamental level will soon tell us more about the phylogenetic relationships of organisms than we have managed to learn in all the 173 years since Lamarck".

## References

- Avise, J.C. and M.H. Smith. 1974a. Biochemical genetics of sunfish. I. Geographic variation and subspecific intergradation in the bluegill, *Lepomis macrochirus*. *Evolution*, 28: 42-56.
- Avise, J.C. 1976. Genetic differentiation during speciation. In F.J. Ayala ed., *Molecular evolution*, Sinauer Associates Sunderland, Mass., pp. 106-122.
- Ayala, F.J. 1975. Genetic differentiation during the speciation process., *Evol. Biol.*, 8:1 - 78.
- Ayala, F.J. 1976. Genetic differentiation during speciation. In F.J. Ayala ed., *Molecular evolution*, Sinauer Associates Sunderland, Mass. pp. 106-122.
- Ayala, F.J. 1975. Genetic differentiation during the speciation process. *Evol., Biol.* 8 : 1-78.
- Ayala, F.J. 1983. Enzymes as taxonomic characters. In G. S. Oxford and D. Rollinson ed., *Protein polymorphism: Adaptive and Taxonomic significance.*, Academic Press, New York, pp. 3-26.
- Bullini, L. 1983. Taxonomic and Evolutionary inferences from electrophoretic studies of various animal groups. In G.S. Oxford and D. Rollinson, ed., *Protein polymorphism: Adaptive and Taxonomic significance.*, Academic Press, New York, pp. 179-192.
- Gale, J.S. 1980. *Population genetics*. Blackie, Glasgow and London.
- Hartl, D.L. 1980. *Principles of population genetics*. Sinauer Associates, Inc., Sunderland Mass.,
- King, M.C. and A.C. Wilson. 1975. Evolution at two levels. Molecular similarities and biological differences between humans and chimpanzees, *Science*, 188 : 107-116.
- Koehn, R.K., A.J. Zera and J.C. Hall. 1983. Enzyme Polymorphism and natural selection. In M. Nei and R.K. Koehn ed., *Evolution of genes and proteins*, Sinauer Associates Inc., Sunderland Mass., pp. 115-136.
- Menezes, M.R. and N. Taniguchi 1988. Interspecific genetic divergence in Sciaenids from Japan and its adjacent waters. *Japan J. Ichthyol* , 35 : 40-46.
- Menezes, M.R. 1990. Biochemical genetic divergence in three Carangids from the Andaman Sea. *Curr. Sci.*, 59 : 209-212.
- Menezes, M.R., N. Taniguchi and S. Seki. 1990. Degree of intraspecific genetic divergence and variability in three Sciaenid species. *Japan. J. Ichthyol.*, 37 : 39-48.
- Nei, M. 1971. Interspecific gene differences and evolutionary time estimated from electrophoretic data on protein identity. *Amer. Natur.*, 105 : 385-398.
- Nei, M. 1972. Genetic distance between populations, *Amer. Natur.*, 106:283-292.
- Nei, M. 1975. *Molecular population genetics and evolution*, North - Holland, Amsterdam.
- Rogers, J.S. 1972. Measures of genetic similarity and genetic distance. *Univ. Texas Publ.*, 7213 :145-153.
- Selander, R.K. 1982. Phylogeny. In R. Milkman, ed., *Perspectives on Evolution*, Sinauer, Sunderland, Mass., pp. 32-59.
- Simpson, G.G. 1945. The principles of classification and a classification of mammals. *Bull. Am. Mus. Nat. Hist.*, 85 :1-350
- Sneath, P.H.A. and R.R. Sokal. 1973. *Numerical Taxonomy*. W.H. Freeman, San Francisco.
- Taniguchi, n. 1980. Nanseikaiiki ni okeru nibe no shuzoku bunka to kakuriyouin. *Bull. Japan Soc. Fish. Oceanogr.* 36 : 53-57 (In Japanese).
- Taniguchi, N. 1982. Nibe no seibutsugaku V. Syoki hatsuiku Dankai. *Aquabiology*, 20 : 210 - 215 (In Japanese).
- Thorpe, J.P. 1979. Enzyme variation and taxonomy : the estimation of sampling errors in measurements of interspecific genetic similarity. *Biol. J. Linn. Soc.*, 11 :369-386.
- Thorpe, J.P. 1982. The molecular clock hypothesis : Biochemical evolution, genetic differentiation and systematics. *I.A. Rev. Ecol. Syst.*, 13:139-168.

