ON SOME THEORETICAL STUDIES ON GENE DIFFERENTIATION IN NATURAL POPULATIONS *

RANAJIT CHAKRABORTY

Center for Demographic and Population Genetics,
University of Texas Health Science Center, Houston, Texas, USA

Different mathematical approaches to study the extent of genetic variation of natural populations are reviewed. The modern understanding of the gene structure permits new interpretations of existing concepts like fixation or inbreeding. A more recent measure of genic divergence, which at molecular level is designed to measure net codon differences is also seen to be related with gene diversity in a substructed population. It is argued that such variations are produced and preserved possibly by simultaneous action of migration, mutation, selection, and random genetic drift. At the present moment it is very difficult to isolate out the effect of each factor because of varying degrees of variation at the different gene sites and between different sets of populations.

From the evolutionary point of view one can distinguish two kinds of biological variation: *individual variation*, referring to differences among individuals of a single population, and *group variation*, referring to divergences between populations. It is the second aspect of variation which will be dealt with in this presentation. But it is by now well established that these two variations are interlinked, one leading to the other. When the differences among spatially segregated populations are considered, we call such group variation *geographic variation.* The adjective « geographic » stems from the fact that the phenomenon was first noticed when geographically far off populations were compared. Numerous theoretical and observational studies now reveal that even neighboring populations differ from each other; indeed in sexually reproducing organisms of finite population size, two demes hardly can ever be identical. This type of local differentiation presents valuable information to the evolutionary biologists and is studied usually under the heading of microdifferentiation yielding what is presently known as microevolution.

Biological variation can be measured with respect to traits which can be broadly classified as: morphological (external or internal), physiological, behavioral, immunological, and biochemical. Traits belonging to only the last two categories show discontinuous variation — that is, characters whose different expressions can be attributed to the presence of different alleles, segregation of which in families conform to simple Mendelian expectations.

* Based on the author's presentation in a panel discussion on " Genetic and evolutionary factors in human differentiation " during the Ninth International Congress of Anthropological and Ethnological Sciences held at Chicago, September 1973. Research supported by Grant GM 19513-02 from the U.S. National Institutes of Health.

Acta Genet. Med. Gemellol. (1975), 24: 111-117

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press

This is so because the immunological or biochemical expressions depend directly on the protein structures which in turn are determined by genetic codes. There is, thus, little room for other genetic differences or environmental effects to obscure the picture of such characters. These are the traits which are kept in mind when simple theories are formulated and mathematics developed to study the genetic causes of geographic variation.

In this article, we shall first review the existing methods of analyzing the genetic variation between populations through the different distance indices. One of these measures is seen to be useful in extending the study of differentiation in substructed population initiated by Wright (1931, 1943, 1946). The statistical properties of such a distance measure are used to see how large or small is the extent of such genetic differentiation. It is further discussed how this differentiation changes with time reflecting the nature of the action of the evolutionary forces. The study of such dynamics now enables us to isolate the factors which force the variability to increase. Nevertheless, there do exist eroding evolutionary factors which retard or hinder the increasing genetic variability in a changing world.

GENETIC DISTANCE AS A MEASURE OF GENIC DIFFERENTIATION

If genic differentiations among populations are to be interpreted as differences in gene frequencies among a set of populations, there exist many such measures in the literature. An early effort towards this problem, as applied to this field of science directly, was made by Sanghvi (1952, 1953). This measure, in its original form, or a modified form of this (Balakrishnan and Sanghvi 1968) is based on chi-square statistic and is analogous to the generalized distance statistic — D^2 (Mahalanobis 1936) based on metric characters. However, it is not clear enough what biological unit is being measured with such a measure. Furthermore, it is a difficult task to find the significance of a distance value computed by this statistic because of its unknown statistical properties.

Among other attempts of defining a measure of genetic distance, Edwards and Cavalli-Sforza's (1964) devise of considering angular transforms of gene frequencies had a definite geometric motivation. Their distance statistic, thus, turned out to be proportional to the direction cosine between two population vectors representing the position of two populations relative to each other on a trasformed scale. In a series of subsequent papers these authors discussed several genetic and geometric properties of this statistic (see Edwards and Cavalli-Sforza 1972, for a summary of this aspect and further references).

The assessment of such affinities as revealed by similarities or distances are also studied by Steinberg et al. (1966), Hedrick (1971), Morton et al. (1971), and Rogers (1972). Morton's measure is seen to provide a good estimate of the coefficient of kinship if the number of populations (among which the variation is being measured) is infinitely large. However, there had been arguments indicating drawbacks of the use of such estimates for measuring genetic differences between populations (Nei 1973*a*) in view of the fact that the number of natural populations is never infinitely large.

Nei recently proposed a new measure of distance which is expressed in terms of the accumulated number of gene substitutions per locus or the number of codon differences (Nei 1971, 1972). This measure is also shown to be related with the concepts of inbreeding and kinship coefficient when these are interpreted in terms of the molecular view of gene structure.

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press

INBREEDING COEFFICIENT AND ITS NEW INTERPRETATION

Inbreeding coefficient as viewed by Wright (1931, 1943, 1946) is shown to be a useful measure of the degree of genetic differentiation among a group of populations. This interpretation of the coefficient of inbreeding as a measure of deviation from Hardy-Weinberg equilibrium indirectly assumes infinite size of population, fixed number of alleles and the constancy of mean genotype frequency by a balance between reversible mutation, migration and selection. Such a model is described as fixed-allele model by Nei (1973*a*).

In such a model if the deviation from Hardy-Weinberg expectation occurs due to subdivision alone, and furthermore if in the subsequent generations gene migration does not take place between subpopulations, the inbreeding coefficient in t^{th} generation is shown to be

$$F_t = 1 - e^{-t/2\mathrm{N}} \tag{1}$$

where N is the size of each subpopulation and the subpopulations are infinite in number.

Cavalli-Sforza's distance statistic (Cavalli-Sforza 1969), f_θ, is shown to be related with the inbreeding coefficient when viewed from this angle. For a locus with two alleles, he showed that $f_\theta \simeq \sigma_x^2/\bar{x}\ (1 - \bar{x})$ where $\bar{x}$ and σ_x^2 are the mean and variance of an allele frequency among the subpopulations. Thus, Cavalli-Sforza related his distance measure with the time of divergence. At this point it must be mentioned that Nei (1973*b*) has recently shown that such a relationship is unwarranted since equation (1) holds only for infinite number of subpopulations. Another limitation of Cavalli-Sforza's analysis is that he has ignored all new mutations as well as the alleles which have been lost or fixed by genetic drift after the sexual isolation. This last supposition, as shown by Nei (1973*b*) would lead to complete loss of genetic variability within populations.

Malecot took a leading role after Wright's initial contribution in the interpretation of inbreeding coefficient. Malecot (1948, 1967) developed the probabilistic interpretation of inbreeding and coined the concept of coefficient of kinship to study the identity by descent of two genes, chosen at random, one from each of the two populations to be compared. The mathematical model considered in this aspect is concerned with the steady-state value of the kinship coefficient with little attention on evolutionary change. Morton's distance measure (Morton 1973) measures this quantity, which, according to Malecot's theory, is related asymptotically with the euclidian distance of the two populations by

$$\Phi(k) = ae^{-\mathrm{bk}}/k^{\mathrm{c}} \tag{2}$$

where $\Phi(k)$ is the kinship coefficient between two populations k units apart from one another (k not very small) and a, b, c, are constants. But the assumptions of steady gene-frequency distribution is quite questionable since many polymorphic loci are in a transient state and in the process of undergoing gene substitution. Furthermore, if the value of the inbreeding coefficient in a subpopulation has to be evaluated relative to the total population, then Wright's method of *F*-statistics would be better than Malecot's method.

The discovery of the molecular structure of genes brought a considerable change in the above two interpretations of inbreeding coefficient and associated measures. The classical concept of recurrent mutations is found to be inaccurate since at molecular level most of the

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press

new mutations occurring in a population are found to be different from the alleles preexisting in the population. This lead to a new concept, called effective number of alleles (Kimura and Crow 1964) which is the inverse of the inbreeding coefficient. Under this model, inbreeding coefficient is treated as the probability of structural identity of two randomly chosen genomes from the population. Once this molecular basis of mutation is accepted, the gene frequency equilibrium is difficult to perceive unless there is strong stabilizing selection or all the deleterious alleles at a locus are treated collectively. Furthermore, expected number of alleles, thus, can vary from time to time, and existing forms of alleles may be different in two different evolutionary time points. The difference of the alleles at the hemoglobin α-chain locus in man and gorilla is a classical example of such variable-allele models.

GENE IDENTITY AND RELATED MEASURES

Keeping this modern view of gene structure in mind, Nei (1971) has defined the concept of gene identity which is the straightforward extension of coefficient of kinship. The expectation of such a measure (suitably normalized), I, is used by Nei (1971) and many others to study the divergence time under mutation-drift-migration balance. Essentially the same theory is used by Nei (1972, 1973*a*) where he defined the three measures of genetic distance designed to measure the net codon difference per locus between two populations. All these quantities are estimable from the gene frequency data only. For example, consider two populations, X and Y, in which multiple alleles segregate at a locus. Let x_i and y_i be the frequencies of the i^{th} allele in X and Y, respectively. The identity probabilities in X and Y are given by $j_X = \Sigma x_i^2$ and $j_Y = \Sigma y_i^2$ and the analogue of kinship; the gene identity between X and Y is $j_{XY} = \Sigma x_i y_i$. This estimation requires no assumption about selection, mutation, and migration. Now, the arithmetic means of j_X, j_Y and j_{XY} over all loci (including the monomorphic loci) give an idea about the entire genome. Denoting them by J_X, J_Y and J_{XY} respectively, the genetic distance (minimum) is computed as

$$D_m = (J_X + J_Y)/2 - J_{XY} \qquad (3)$$

At this point, it is worthwhile to note that the probabilities of non-identity (structural, as per molecular theory) of genes are computed by $1 - J_X$, $1 - J_Y$, etc. If the populations are random mating, $1 - J_X$ represents the heterozygosity (average) of the population X. The general terminology for such expressions, as suggested by Nei (1973*c*), is gene diversity and are denoted by $H (= 1 - J)$. For studying the gene differentiation due to population subdivision, Nei also decomposed the total gene diversity in the whole population (H_T) into the inter- and intra-subpopulational components (D_{ST} and H_S, respectively). The coefficient of gene diversity, G_{ST}, is then defined as

$$G_{ST} = D_{ST}/H_S \qquad (4)$$

In fact D_{ST} is the pooled minimum genetic distance [as estimated by equation (3)] averaged over all possible comparisons of the subpopulations. This measure of describing gene frequency variation due to substructure of a population leads to the equation

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press

$$(1 - G_{ST})(1 - J_T) = (1 - J_S) \tag{5}$$

which is parallel to the Wrightian equation

$$(1 - F_{IT}) = (1 - F_{IS})(1 - F_{ST})$$

where F's have the correlation interpretations. Though F_{IS}, F_{IT} may be negative numbers, F_{ST} is always positive. On the other hand all the quantities in equation (5) are positive since they represent identity probabilities. Furthermore, F-statistics are only applicable if there are two alleles at a locus, or if there is random gene differentiation in the presence of multiple alleles without any selective forces. However, G_{ST} and F_{ST} are identical in a two-allelic case and even in a multi-allelic case there is a relationship between them.

The relationship between F_{ST} and evolutionary time as described by equation (1) is also extended in the case with finite number of subpopulations by considering G_{ST}. However, for estimating the evolutionary time a better parameter is the normalized gene identity defined by

$$I = J_{XY}/\sqrt{J_X J_Y}$$

where J_X, J_Y and J_{XY} are as defined earlier. The population dynamics of gene differentiation in completely or partially isolated populations are studied through this parameter by Nei and Feldman (1972) and Chakraborty and Nei (1974).

One important appeal of this particular approach of measuring the genetic distance or genetic variation is that the sampling variance of such measures allow us to see how large or small the observed value of the parameter is.

ACTION OF EVOLUTIONARY FACTORS

The controlling factors of genetic variation can broadly be classified in three groups: (*a*) the input of new genetic material through mutation and immigration, (*b*) the erosion of this variation by directional selection and genetic drift due to random sampling process in finite population, and (*c*) the protection of the stored variability by cytophysiological devices and ecological factors. The action of all these factors being simultaneous, the actual process is much more complicated than the mathematical simplifications provided by the theoreticians. The genetic variation present in the different populations, thus, cannot be plainly ascribed to any particular cause of variation. Furthermore, it is also noticed that various characters of a species may have variations of different order. Human races give a very good example of this. While the three major races (Caucasoid, Mongoloid and Negroid) have large non-overlaps with respect to morphologic traits like pigmentation, hair texture, body build and facial features, serologic and biochemical traits held the three races only minutely apart from one another. In fact about 92% of such variations in these "general" genes can be ascribed to within-race variation (Lewontin 1972, Nei and Roychoudhury 1972). A possible reason for such differential variation is the action of selection. The "general" genes, on the other hand, are not subjected to such a great extent of selection. However, there are a lot

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press

of disputes over such factors in the case of allozymic variations. Crow-Kimura's (Kimura and Crow 1964) neutrality theory is the one which advocates a minor role of selection so far as the majority of loci are involved. Since the inception of this theory there have been various reports supporting the theory and criticizing it. But, one can always find either approximation or inaccuracy in such arguments.

REFERENCES

Balakrishnan V., Sanghvi L.D. 1968. Distance between populations on the basis of attribute data. Biometrics, 24: 859-65.

Cavalli-Sforza L.L. 1969. Human diversity. Proc. 12th Int. Congr. Genet., Tokyo, 3: 405-16.

Chakraborty R., Nei M. 1974. Dynamics of gene differentiation between incompletely isolated populations of unequal sizes. Theor. Pop. Biol., 5: 460-469.

Edwards A.W.F., Cavalli-Sforza L.L. 1964. Reconstruction of evolutionary trees. Publs. Syst. Ass., 6: 67-76.

Edwards A.W.F., Cavalli-Sforza L.L. 1972. Affinity as revealed by differences in gene frequencies. In J.S. Weiner and J. Huizinga (eds.): The Assessment of Population Affinities in Man (pp. 37-47). Oxford: Clarendon Press.

Hedrick P.W. 1971. A new approach to measuring genetic similarity. Evolution, 25: 276-280.

Kimura M., Crow J.F. 1964. The numbers of alleles that can be maintained in a finite population. Genetics, 49: 725-38.

Lewontin R.C. 1972. The apportionment of human diversity. In T. Dobzhansky, M.K. Hecht, W.C. Steere (eds.): Evolutionary Biology (Vol. 6, pp. 381-398). New York: Appleton-Century-Crofts.

Mahalanobis P.C. 1936. On the generalized distance in statistics. Proc. Natl. Inst. Sci., India, 12: 49-55.

Malecot G. 1948. Les Mathematiques de l'Hérédité. Paris: Masson et Cie.

Malecot G. 1967. Identical loci and relationship. Proc. 5th Berkeley Symp. Math. Stat. Prob., Vol. 4, pp. 317-32.

Morton N.E. 1969. Human population structure. Ann. Rev. Genet., 3: 53-74.

Morton N.E., Yee S., Harris D.E., Lew R. 1971. Bioassay of kinship. Theor. Pop. Biol., 2: 5-7-24.

Morton N.E. 1973. Kinship and population structure. In N.E. Morton (ed.): The Genetics of Population Structure. Honolulu: University of Hawaii Press.

Nei M. 1971. Identity of genes and genetic distance between populations. Genetics, 68: s47.

Nei M. 1972. Genetic distance between populations. Amer. Nat., 106: 283-92.

Nei M., Feldman M. 1972. Identity of genes by descent within and between populations under mutation and migration. Theor. Pop. Biol., 3: 460-65.

Nei M., Roychoudhury A.K. 1972. Gene differences between Caucasian, Negro, and Japanese populations. Science, 177: 434-36.

Nei M. 1973*a*. The theory and estimation of genetic distance. In N.E. Morton (ed.): The Genetics of Population Structure. Honolulu: University of Hawaii Press.

Nei M. 1973*b*. Dynamics of gene differentiation among a finite number of populations. (Unpublished).

Nei M. 1973*c*. Analysis of gene diversity in subdivided populations. Proc. US Nat. Acad. Sci, 70: 3321-3323.

Rogers J.S. 1972. Measures of genetic similarity and genetic distance. In: Studies in Genetics, VII. Austin: University of Texas Publ.

Sanghvi L.D. 1952. Biological studies on some endogamous groups in Bombay, India. Ph.D. thesis, Columbia University, New York.

Sanghvi L.D. 1953. Comparison of genetical and morphological methods for a study of biological differences. Amer. J. Phys. Anthrop., 11: 385-404.

Steinberg A.G., Bleibtrue H.K., Kurczynski T.W., Martin A.O., Kurczynski E.M. 1966. Genetic studies on an inbred human isolate. Proc. 3rd Int. Congr. Hum. Genet., Chicago, pp. 267-289.

Wright S. 1931. Evolution in Mendelian populations. Genetics, 16: 97-159.

Wright S. 1943. Isolation by distance. Genetics, 23: 114-38.

Wright S. 1946. Isolation by distance under diverse systems of mating. Genetics, 31: 39-59.

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press

RIASSUNTO

Alcuni Studi Teorici sulla Differenziazione Genica nelle Popolazioni Naturali

Vengono esaminati diversi apporti matematici allo studio dell'entità della variazione genetica delle popolazioni naturali. Alla luce della moderna visione della struttura del gene, vengono cercate nuove interpretazioni di concetti quali fissazione o endogamia. Una piú recente misura della divergenza genica, che a livello molecolare vuole misurare la differenza netta di codon, viene anche vista in relazione alla diversità genica in una popolazione substrutturata. Si sostiene che tali variazioni siano verosimilmente prodotte e preservate dall'azione simultanea di migrazione, mutazione, selezione e deriva genetica casuale. Al momento è molto difficile isolare l'effetto di ciascun fattore a causa del diverso grado di variazione sui diversi siti genici e fra popolazioni diverse.

RÉSUMÉ

Quelques Etudes Théoriques sur la Différentiation Génétique chez les Populations Naturelles

Différents apports mathématiques à l'étude de l'entité de la variation génétique des populations naturelles sont examinées. A la lumière de la vision moderne de la structure du gène, l'on cherche de nouvelles interprétations de concepts tels que fixation ou endogamie. Une mesure plus·récente de la divergence génique, qui au niveau moléculaire veut mesurer la différence nette de codon, est aussi observée en relation à la diversité génique dans une population substructurée. On soutient que ces variations sont vraisemblablement produites et préservées par l'action simultanée de migration, mutation, sélection et dérive génétique casuelle. Sur le moment, il est très difficile d'isoler l'effet de chaque facteur à cause du différent degré de variation sur les divers sites géniques et entre populations diverses.

ZUSAMMENFASSUNG

Einige theoretische Untersuchungen über die Erbdifferenzierung bei den natürlichen Bevölkerungen

Es werden verschiedene mathematische Beiträge zum Studium des Umfangs der Erbvariation bei den natürlichen Bevölkerungen untersucht. Der modernen Anschauung über die Genstruktur folgend suchten die Verf. neue Deutungen für Begriffe wie Genfixierung oder Endogamie. Ein neuerer Masstab für die Gendivergenz, der die genaue Codondifferenz messen soll, wird auch bezüglich der Erbverschiedenheit in einer substrukturierten Bevölkerung gesehen. Man denkt, dass diese Variationen wahrscheinlich durch das Zusammenwirken von Migration, Mutation, Selektion und zufälliger Drift bedingt und auch erhalten worden sind. Zur Zeit ist es sehr schwer, die Wirkung eines jeden dieser Faktoren zu isolieren, weil sich die Variation zu unterschiedlich auf die verschiedenen Genstellen und unter den einzelnen Populationen auswirkt.

Dr. R. Chakraborty, Center for Demographic and Population Genetics, University of Texas Health Science Center, Houston, Texas 77025, USA.

https://doi.org/10.1017/S1120962300021958 Published online by Cambridge University Press