Offprint from

PROCEEDINGS

OF THE

IUFRO JOINT MEETING OF WORKING PARTIES

ON

POPULATION AND ECOLOGICAL GENETICS,

BREEDING THEORY

AND

PROGENY TESTING

STOCKHOLM 1974

Published by

the Department of Forest Genetics,
the Royal College of Forestry,
S-104 05 Stockholm, Sweden

On the concept of genetic distance between populations based on gene frequencies

von Dr. Hans-Rolf Gregorius
Lehrstuhl für Forstgenetik und Forstpflanzenzüchtung
der Universität Göttingen

ABSTRACT

A brief discussion of some measures of genetic distance in use, led to the conclusion that in most cases there exist open points as well as insufficiencies in definiton. We particularly pointed out that the interpretation of genetic distance between distributions (sample distributions) provides no satisfactory answer to the proper question about the 'true' genetic distance between populations. For this reason we tried to establish a perspective on the concept of genetic distance based on several fundamental requirements which are implied by the statement of the problem. Within this scope we specified a few actual measures of distance from which one has been distinguished because of its good accordance with intuitive ideas.

INTRODUCTION

The term 'genetic distance' is frequently used with entirely different consequences and by some authors defined in such a way that it is aplicable only to very specific situations. Nei (1971) for example, conceived genetic distance as the number of gene differences of a pair of species while some time later (1972) he used the same term for the measurement of differences between the genetic compositions of two populations. An extensive discussion of several genetic distance measures in use (including his own) has been given by Latter (1973). He compared the results which he obtained from application of the different measures on some empirical data, as well as tried to investigate the relationship of these measures to the coefficient of kinship.

To my knowledge all authors which dealt with genetic distance between populations chose as a representation of a populations genetic composition for one locus the corresponding vector of gene frequencies and regarded this (stochastic) vector as a point in an euclidian space of specified dimension. The case of multiple loci is treated by taking the cartesian product of all the single spaces each containing the representations of genetic compositions of one locus. This construction considers each locus as a single entity and makes no contribution to the compoundedness of genes located on one chromosome or gamete. Therefore, in some cases, it seems to reasonable to use gamete frequencies instead of gene frequencies by which the same conditions are obtained as with one locus.

After having specified the representation of the population space the next step should be the construction of a meaningful distance measure defined for each pair of populations to be compared. But at this point, there is no agreement about what principal properties a genetic distance measure should have. Furthermore, it particularly doesn't make sense to develop estimates of distances which are not properly defined; examples of which we can find a great deal in the technical literature. Above all one should bear in mind that with estimation problems the 'true' genetic distance is a functional (and not a distributional) parameter of the distribution of the sample variables which,in most cases, is multinomial. Last, but noch least, it should be pointed out that genetic distance between populations in the above sense cannot be interpreted as distance between sample or any other distributions, a problem which has been treated by Mahalanobis (1936) for normal distributions Bhattacharyya (1946) for multinomial distributions.

Because of all these ambiguities it seems to be suitable to attempt a concept of genetic distance between populations.

THE CONCEPT

We start with genetic compositions of populations at one locus. Each population is represented by a stochastic vector, the components of which are the relative frequencies of the alleles in case the population is of finite size and the probabilities of the alleles in case the population is of infinite size. Let us assume that all populations considered possess at most n different alleles. Therefore, all populations are to be found in the n-dimensional simplex Θ_n which is a subspace of the n-dimensional euclidian space:

$$\Theta_n := \left\{ x \mid x = (x_1, \ldots, x_n);\ x_i \geq 0 \text{ for } i=1, \ldots, n;\ \sum_{i=1}^{n} x_i = 1 \right\}$$

1) First of all a distance should be a function which assigns to each pair of populations (that is to each pair of elements from Θ_n) a non-negative value, this function may be called d(x,y).
2) There should be no directional effect when we measure the distance from one population to another, i.e. d(x,y) = d(y,x).
3) A distance is 0 if and only if the two populations have identical genetic compositions which means if x=y.
4) The distances from one population to two others should be comparable in the sense that the distance between these two other populations should not exceed the sum of their distances to the first population, i.e. for x,y,z from Θ_n the triangle inequality

$$d(x,z) + d(z,y) \geq d(x,y)$$

should be valid.

A function which satisfies these four conditions is called a metric and meets our intuitive ideas about the term 'distance' best. Up to that point no property has been mentioned which describes a special genetic aspect. But at least one genetically suggestive condition is indispensable:

5) if and only if two populations have no alleles in common they always should have the same maximum and finite distance. The geometrical interpretation of the statement 'two populations have no alleles in common' reads: two vectors are orthogonal, which is equivalent to the fact that the scalar product of vectors x and y defined as

$$(x|y) := \sum_{i=1}^{n} x_i y_i$$

is 0.

It must be emphasized that a great number of additional conditions may be reasonable for more specific situations, but that the above conditions 1) to 5) are principal.

The following examples shall demonstrate the existence of metrics obeying condition 5). For convenience we use the euclidian length $\|x\|$ of a vector x defined by $\|x\|^2 = (x,x)$.

$$d_0(x,y) := \sum_{i=1}^{n} |x_i - y_i|$$

$$d_1(x,y) := \left\| \frac{x}{\|x\|} - \frac{y}{\|y\|} \right\| = \sqrt{2\left(1 - \frac{(x|y)}{\|x\| \cdot \|y\|}\right)}$$

$$d_2(x,y) := \arccos \frac{(x|y)}{\|x\| \cdot \|y\|}$$

$$d_3(x,y) := \|T(x) - T(y)\| = \sqrt{2(1 - (T(x)|T(y)))}$$

$$d_4(x,y) := \arccos (T(x)|T(y));$$

where $T(x_1,\ldots,x_n) = (\sqrt{x_1},\ldots,\sqrt{x_n})$ is the angular transformation. All x, y being elements of Θ_n.

d_1 is the euclidian distance between the normalized vectors; while d_2 is the arc of the corresponding angle between them. The same interpretations apply to d_3 and d_4 taking the transformed vectors as the basis. The validity of condition 4) for d_2 and d_3 has been proved e.g. by Rinow (1961, pp. 4-5), all the other proofs being trivial. It should be mentioned that

Edwards and Cavalli-Sforza (1964) already used d_3 and d_4 for measuring genetic distance but started with a different object in view. Bhattacharyya (1946) applied d_4 to the measurement of divergence between multinomial distributions. The maximal values of d_0 to d_4 are 2, $\sqrt{2}$, $\pi/2$, $\sqrt{2}$, $\pi/2$, so that it is possible to normalize these metrics.

In its most elementary and therefore clearest form the intuitive concept of distance refers to measurement of length, i.e. the difference between real numbers which didn't undergo any transformation. Because d_0 is constructed on this basis it should be distinguished. Another justification for a preferation of d_0 is given by the fact that the ratio of two lengths measured with help of d_0 may be converted by application of d_3 e.g. (see appendix).

An extension to the multiple locus case has to deal with two points of view chiefly:

a) Gametes and not genes are considered as units of genetic information; the distance between two populations is to be measured relative to these units
b) All the single loci are considered to contribute separately from each other but with possibly different weights to the over all distance.

In a) the gene is substituted by the gamete which indicates the equivalence to the one locus case. The existence of at least one locus at which two populations have no alleles in common implies maximal distance between these two populations. The reversal does not apply in all cases.

b) requires definition of distances for all the single loci and of a function of these distances which reflects different weights given to the loci. Let us assume that for the i-th locus the genetic composition of a population may be represented as an element of the n_i-dimensional simplex Θ_{n_i}, so that for k loci the population may be regarded as an element of the cartesian product

$$\Theta = \bigtimes_{i=1}^{k} \Theta_{n_i}.$$

For each Θ_{n_i} a distance d_i satisfying conditions 1) to 5) shall be given. The over all genetic distance being defined on Θ as a function of the d_i again should satisfy conditions 1) to 5), at which condition 5) now reads: if and only if two populations at each of the k loci have no alleles in common, they always should have the same maximum and finite distance. There are at least two such functions:

$$d(x,y) := \sum_{i=1}^{k} a_i \cdot d_i(x_i,y_i)$$

$$d'(x,y) := \sqrt{\sum_{i=1}^{k} a_i \cdot d_i(x_i,y_i)^2}.$$

$a_1,\ldots,a_k$ are positive real numbers summing up to 1.

Examination of some measures of genetic distance in use with help of the concept

In what follows explicit presentation of formulae is omitted in favour of general discription. If a measure is defined for sample values only, we considere its stochastic limit (in case it exists) because we are interested in 'true' distances solely. The following measures are treated:

G_s by Sanghvi (1952, 1953); E by Edwards (1971) and Edwards and Cavalli-Sforza (1972); B by Balakrishnan and Sanghvi (1968); D bei Nei (1972).

G_s: As has been shown by Kurczynski (1970), G_s is equivalent to another measure of genetic distance D_k suggested by Steinberg et al. (1967). It is built up on the χ^2 statistic and therefore does not satisfy conditions 4) and 5).

E : E satisfies conditions 1) to 4) because it is the euclidian distance between two times transformed genetic compositions. These transformations are the angular transformation and the stereographic projection. On the other hand, condition 5) is not met because the maximum value of E depends on the number of different alleles and is not assumed for populations only which possess no alleles in common.

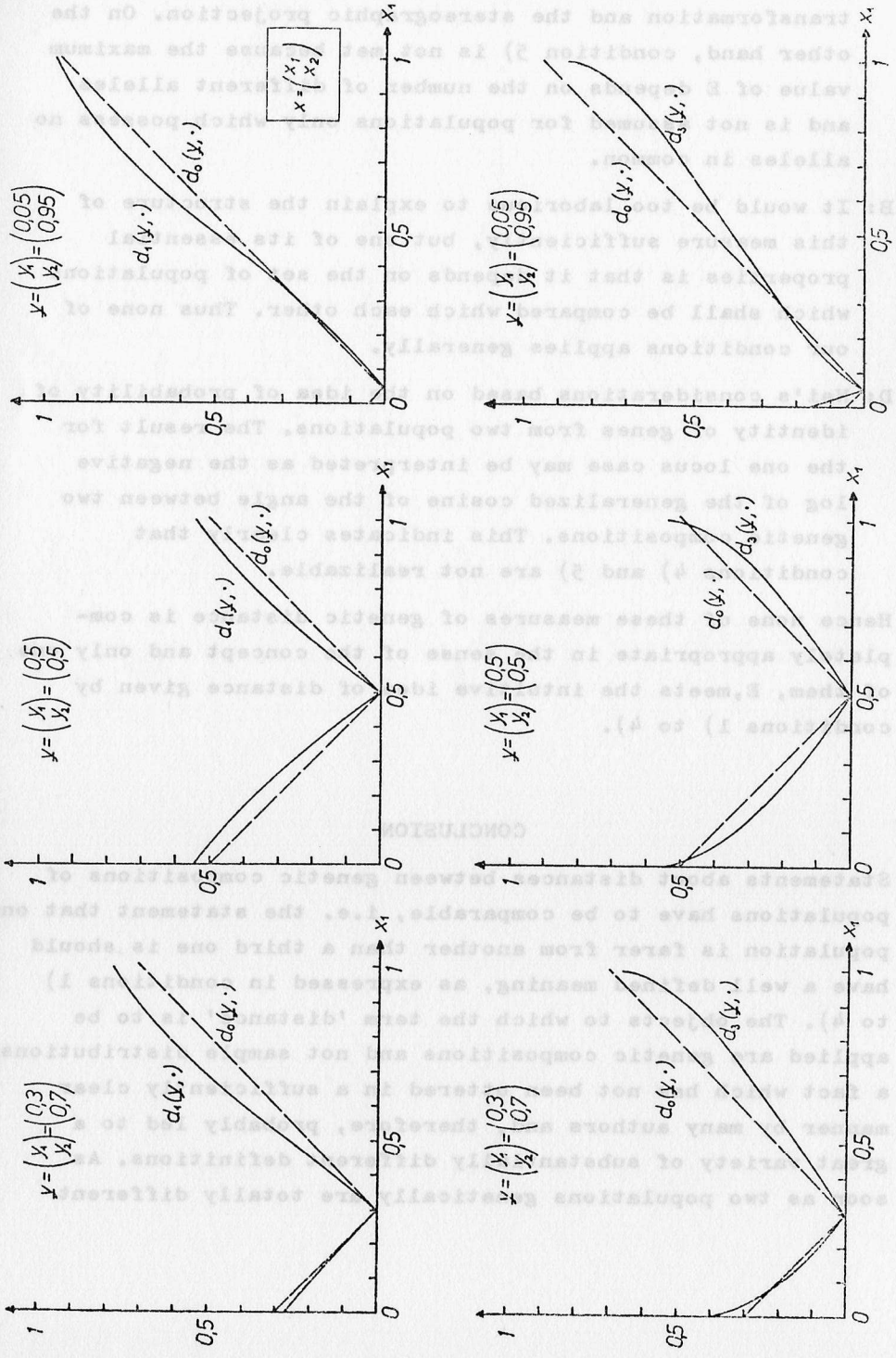
B: It would be too laborious to explain the structure of this measure sufficiently, but one of its essential properties is that it depends on the set of populations which shall be compared which each other. Thus none of our conditions applies generally.

D: Nei's considerations based on the idea of probability of identity of genes from two populations. The result for the one locus case may be interpreted as the negative log of the generalized cosine of the angle between two genetic compositions. This indicates clearly that conditions 4) and 5) are not realizable.

Hence none of these measures of genetic distance is completely appropriate in the sense of the concept and only one of them, E,meets the intuitive idea of distance given by conditions 1) to 4).

CONCLUSION

Statements about distances between genetic compositions of populations have to be comparable, i.e. the statement that one population is farer from another than a third one is should have a well defined meaning, as expressed in conditions 1) to 4). The objects to which the term 'distance' is to be applied are genetic compositions and not sample distributions, a fact which has not been uttered in a sufficiently clear manner by many authors and, therefore, probably led to a great variety of substantially different definitions. As soon as two populations genetically are totally different



(for the loci considered) their special genetic compositions cannot increase or decrease this dissimilarity further on, because there is no suggestive intensification of disjunctiveness thinkable. Thus in just that case and no other the distance should attain a maximum pointed out in condition 5). But all these requirements do not suffice to confine the number of all possible distances to only one as shown above. However, because of its close relationship to simple measurement of length d_o may be distinguished.

APPENDIX

The following figures shall examplarily demonstrate the relationship of the metrics d_1 and d_3 to d_o. A metric d will be called invariant if for any genetic compositions u,v,x,y with $d(u,v) \leq d(x,y)$ the implication $d_o(u,v) \leq d_o(x,y)$ is true. This is an important property because if it does not apply, conclusions about the ratio of two distances drawn from d_o may be inverted by d.

We consider one locus with two alleles and choose the genetic composition $\underline{y}$ of a fixed population which makes it possible to regard the distance $d(\underline{y},x)$ as a function of the genetic composition x of a variable population. This is done for three different values of $\underline{y}$.

(figures)

It may be taken from the figures that essentially d_1 assumes larger and d_3 smaller values than d_o. The condition of invariance in some cases is not met:

$\underline{y} = \begin{pmatrix}0.3\\0.7\end{pmatrix}$, $z = \begin{pmatrix}0\\1\end{pmatrix}$, $x = \begin{pmatrix}0.7\\0.3\end{pmatrix}$:

$d_o(\underline{y},z) = 0.3 < d_o(\underline{y},x) = 0.4$ but

$d_3(\underline{y},z) = 0.404 > d_3(\underline{y},x) = 0.289$

and taking $x = \begin{pmatrix}0.55\\0.45\end{pmatrix}$:

$d_o(\underline{y},z) = 0.3 > d_o(\underline{y},x) = 0.25$ but

$d_1(\underline{y},z) = 0.284 < d_1(\underline{y},x) = 0.336$

These statements of course are not transferable in a direct manner to the general case and do not exclude the existence of further metrics which are invariant, as an example given by Rinow (1961, p. 70) shows.

LITERATURE CITED

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

Bhattacharyya, A., 1946: On a measure of divergence between two multinomial populations. Sankhya 7, 401-6

Edwards, A.W.F., 1971: Distances between populations on the basis of gene frequencies. Biometrics 27, 873-81

Edwards, A.W.F., L.L. Cavalli-Sforza, 1964: Reconstruction of evolutionary trees. In Phenetic and Phylogenetic Classification, Publ. No. 6, 67-76 Systematics Association, London

Edwards, A.W.F., L.L.Cavalli-Sforza, L.L., 1972: Affinity as revealed by differences in gene frequencies. In The Assessment of Population Affinities in Man. Clarendom Press, Oxford

Kurczynski, T.W., 1970: Generalized distance and discrete variables. Biometrics 26, 525-34

Latter, B.D.H., 1973: The estimation of genetic divergence between populations based on gene frequency data. Amer.J.Hum.Genet. 25, 247-261

Mahalanobis, P.C., 1936: On the generalized distance in statistics. Proc.Nat.Inst.Sci. India 2, 49-55

Nei, M., 1972: Genetic distance between populations. The American Naturalist, Vol.106,No.949,283-92

Rinow, W., 1961: Die innere Geometrie der metrischen Räume. Springer Verlag, Berlin-Göttingen-Heidelberg

Steinberg, A.G., Bleibtreu, H.K., Kurczynski, T.W., Martin, A.O., Kurczynski, E.M., 1967: Genetic studies on an inbred human isolate. Proc.Third Int.Congress Hum.Genetics. Eds. J.F. Crow and J.V. Neel. John Hopkins Press, Baltimore

Session I, report by chairman: A. NANSON

Topic: CRITICAL REEVALUATION OF BASIC CONCEPTS OF QUANTITATIVE GENETICS WHEN APPLIED TO FOREST TREES

Conclusions and Recommendations

1) There are important differences in some situations encountered in forest tree breeding as compared to animal breeding for which classical quantitative genetics has been developed. In these cases procedures and even concepts must be reviewed and adapted.

2) At the provenance or "variety" level, there is a gap not covered by classical quantitative genetics, which should be filled.

3) The introduction of concepts such as genotypic element, macrosite and microsite, genotypic heritability and gains, developmental stages, and chiefly early test theory with its possible generalization could contribute to cover the whole range of forest tree breeding with a unified and often more realistic theory.

4) Narrow sense heritability does not seem to have the general meaning it has in animal breeding. Particular caution must be made for its interpretation. In many cases, it should be replaced by the concept of correlation, which is more general. Likewise gains should be expressed as far as possible in terms of correlated gains at rotation end.

5) Many problems still need investigation for developing optimum breeding strategies, for example: inbreeding and outbreeding, crosses between individuals from different populations, addition of gains and information from relatives in different macrosites and developmental stages, use of incomplete and non-orthogonal sources of information, the scale problem, correlated gains from early indeces, selected populations, epistasis, linkage, common environment, etc.