Arunachalam, V.
Genetic Distance in Plant Breeding.
Indian Journal of Genetics and Plant Breeding,
41 (1981): 226-236.

Indian J. Genet., 41(1981): 226–236

GENETIC DISTANCE IN PLANT BREEDING

V. ARUNACHALAM

Indian Agricultural Research Institute, Regional Station, Rajendranagar, Hyderabad - 500 030

ABSTRACT

As exposition of the theoretical concepts behind genetic distance has been made. Problems relating to choice of characters, non-repeatable classification configuration, heterosis as explained by genetic divergence and methods of classifying a large number of entries are dealt with with special reference to plant breeding. The role of genetic divergence in plant breeding has been stressed with examples.

THE concept of 'genetic distance' has been of vital utility in many contexts and more so in differentiating well-defined populations. Several measures of distance have been proposed over the past two decades to suit various objectives (for a comprehensive review, see Jacquard, 1970), of which Mahalanobis' generalised distance (Mahalanobis, 1930, 1936; Rao, 1952) occupied a unique place in plant breeding. Yet, as it happens in biology, several problems, under the influence of random, unpredictable changes due to environment, evade the direct grip of concepts well-proven in more exact fields like mathematical statistics, physics and others where environmental influence is not a major component to deal with. It then becomes essential not only to delineate the limits within which these concepts do work but to comprehend the conditions basic to their application. A number of investigations concerned with assessment of the genetic diversity in a number of diverse food crops has been published in the past decade and a half. However, clear answers to the following questions could not easily be obtained:

(a) In so far as the genetic distance is based on the quantitative characters defining the genotypes, what are the criteria behind an apt choice of characters?

(b) What are the methods available to take into account the non-repeatability of the configuration of clusters including the intra-and inter-cluster distances?

(c) What can unambiguously be said of the relationship between genetic divergence and realisable heterosis?

(d) How can the need for classifying a large number of genotypes using Mahalanobis' generalised distance be met?

(e) Keeping the above questions in view, what can one say on the role of genetic distance in plant breeding?

We attempt here a brief exposition of the theoretical concepts behind genetic distance in as simple (even at a nominal cost of theoretical rigour) a

manner as possible and set the experimental results that are relevant to our questions in a clear perspective.

Genetic Distance: If two entities can be totally characterised by measurement on a single character, then it is easily seen that the difference between them in the value of the character is the best measure of distance. Let us consider that a complete characterisation of a genotype G can only be made by the values of two *independent* characters. Let G_1, G_2 be two genotypes and let them be characterised by measurements, G_1 (x_{11}, x_{12}) and G_2 (x_{21}, x_{22}) on the two independent characters X_1 and X_2; we adopt the convention that x_{ij} denotes the value of the genotype i for character j. If the two characters are taken to represent two axes of reference, the genotypes can be represented as in Fig. 1. We note that the axes are mutually perpendicular as the characters are statistically independent. Then by elements of analytical geometry, we obtain the distance, d between the genotypes G_1 and G_2 as $d^2=(x_{21}-x_{11})^2+(x_{22}-x_{12})^2$.

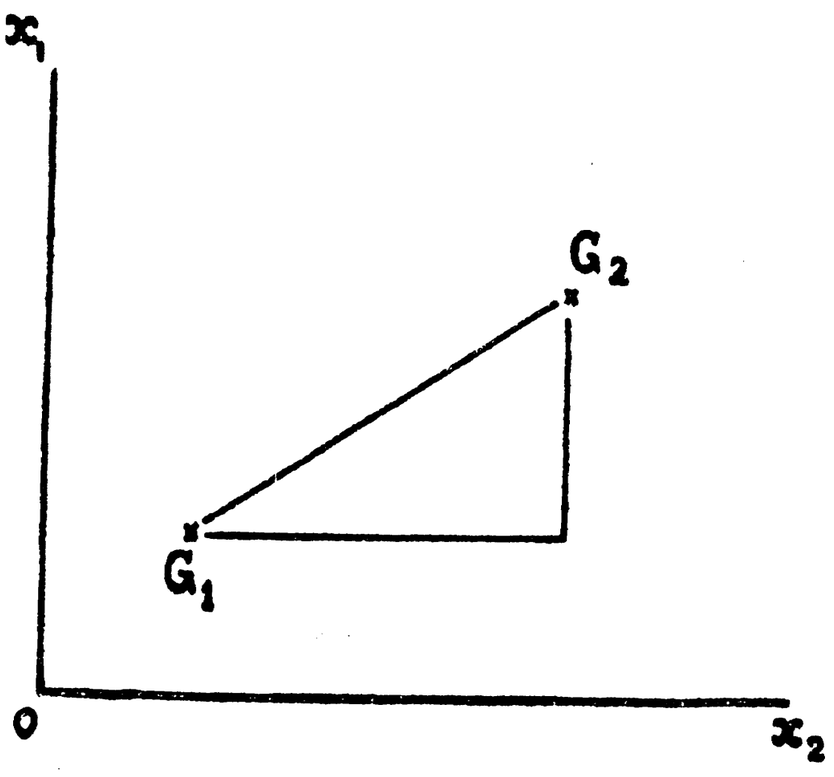

Fig. 1

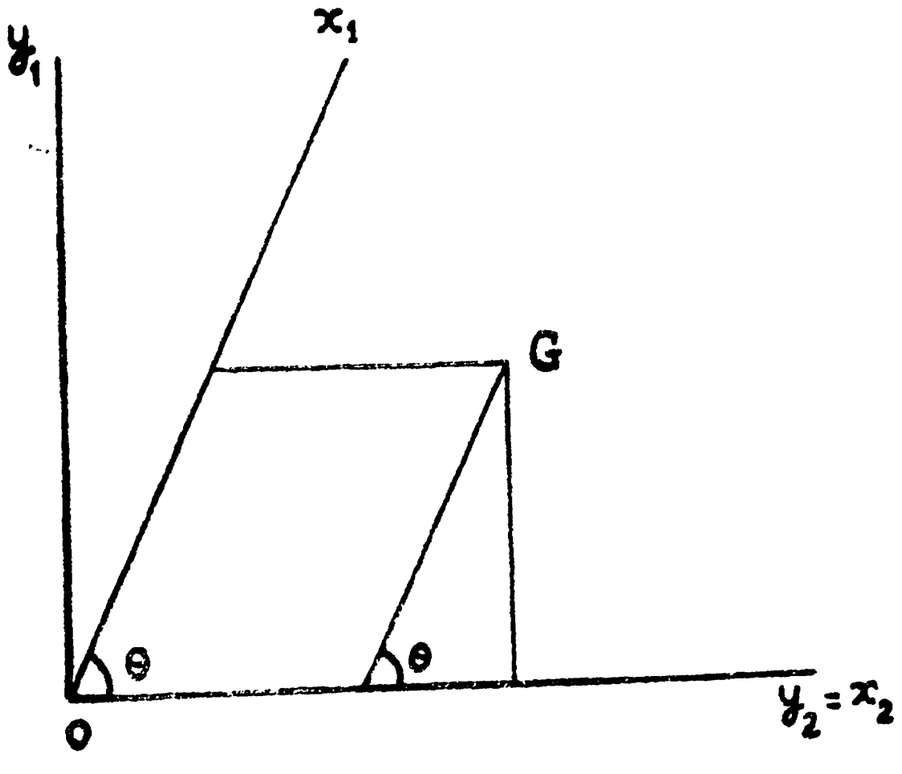

Fig. 2

If now, we denote the differences in the value of characters of the genotypes G_1 and G_2 by d_1 and d_2, we have

$$d_1=x_{21}-x_{11}$$

$d_2=x_{22}-x_{12}$ so that

$$d^2=d_1^2+d_2^2 \qquad \ldots\ldots(A)$$

If, on the other hand, the characters X_1 and X_2 are not independent, the axes become oblique and inclined at an angle θ (Fig. 2). It is then easy to transform the coordinates to conform to rectangular axes, Y_1 and Y_2. Since Y_2 axis can be made to coincide with X_2 axis, the transformation is given by—

$$\left.\begin{aligned} Y_1 &= X_1 \sin\theta \quad \ldots\ldots\ (B1) \\ Y_2 &= X_1 \cos\theta + X_2 \quad \ldots\ldots\ (B2) \end{aligned}\right\} \quad \ldots\ldots\ (B)$$

The transformed a values are then given by

$$G_1\ (y_{11}, y_{12}) : (x_{11} \sin\theta,\ x_{11} \cos\theta + x_{12})$$

$$G_2\ (y_{21}, y_{22}) : (x_{21} \sin\theta,\ x_{21} \cos\theta + x_{22})$$

$$d^2 = (y_{21}-y_{11})^2 + (y_{22}-y_{12})^2$$

$$= d_1^2 + d_2^2 + 2\, d_1 d_2 \cos\theta, \text{ using the equations (B)}$$

We note that $\cos\theta$ is the correlation coefficient between the variabies X_1 and X_2. When $\theta = 90°$, $\cos\theta = O$; the variables are uncorrelated and $d^2 = d_1^2 + d_2^2$, as expected.

We observe that a linear transformation of the form,

$$Y_1 = a_1 X_1$$

$$Y_2 = a_2 X_1 + b_2 X_2 \quad \ldots\ldots\ (\text{cf..(B)})$$

transforms the coordinates of G_1 and G_2 to rectangular ones so that the distance is given in the simplest form (A).

The above formulation is based on the assumption that the value x_{ij} of the character j of the genotype G_i is the genotypic value and as such should be invariant under all environments. This is, however, hypothetical and as we know, we can only measure the phenotypic value in practice which is the result of genotype interacting with environment. In general, the environmental effect cannot be estimated, though by an appropriate use of field designs, an estimate of environmental variance can usually be obtained. Thus, x_{ij}, the phenotypic value of the genotype G_i for the character j, can, in general, have a mean and a variance when we consider a range of environments in which it is expressed. Alternately, the variable x_{ij} can have a mean and a variance when we consider it in any particular environment, over a number of genotypes (we wish to study). The latter situation is of relevance in genetic divergence studies providing a dispersion matrix of the variables included in the divergence analysis.

Recalling the formula (A) for d^2, we observe that the character variables X_1, X_2 are independent. When the variables are dependent we derived a formula (B) which involve $\cos\theta$ which is a measure of the degree of dependence. But we did not consider the dispersion matrix of the variables.

Let us now consider D, the dispersion matrix of the variables X_1 and X_2,

$$((D)) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

where $\sigma_i^2 = \text{Var}(X_i)$ and $\sigma_{ij} = \text{Cov}(X_i, X_j)$. The dispersion matrix of transformed variables should be a unit matrix, if formula (A) is to hold (Rao, 1952, Chap. 9; Jacquard, 1970, Chap. 14). This would mean that we need to normalise the equation (B1) by dividing throughout by the standard deviation of Y_1 and similarly for the equation (B2). Further, Cov $(Y_1, Y_2)=O$.

$$\text{Now, Var } (Y_1) = \sigma_1^2 \sin^2\theta$$

$$\text{Var } (Y_2) = \sigma_1^2 \cos^2\theta + {}^2\sigma_{12} \cos\theta + \sigma_2^2$$

$$\text{Cov } (Y_1, Y_2) = 0 \text{ gives}$$

$$\sigma_{12} \sin\theta + \sigma_1^2 \sin\theta \cos\theta = 0$$

$$\text{i. e. } \cos\theta = -\sigma_{12} / \sigma_1^2 \qquad \ldots\ldots (1)$$

Normalising the equations (B), we get

$$Z_1 = Y_1 / \sqrt{\text{Var}}\,(Y_1) = X_1 / \sigma_1$$

$$Z_2 = Y_2 / \sqrt{\text{Var}}\,(Y_2)$$

$$= (X_1 \cos\theta + X_2) / \sqrt{(\sigma_1^2 \cos^2\theta + 2\sigma_{12} \cos\theta + \sigma_2^2)}$$

$$\text{Using (1), } Z_2 = (1/t\sigma_1) \quad (-X_1 \sigma_{12} + X_2 \sigma_1^2) \text{ where } t = \sqrt{\sigma_1^2 \sigma_2^2 - \sigma_{12}^2}$$

If now, we rename Z_1 and Z_2 as Y_1 and Y_2 for convenience, we get the transformation as

$$Y_1 = X_1 / \sigma_1$$

$$Y_2 = (1 / t\sigma_1)\ (-X_1 \sigma_{12} + X_2 \sigma_1^2) \qquad \ldots\ldots (C)$$

which are identical to those obtained (See Appendix I) from the matrix ((D)) by pivotal condensation method described by Rao (1952) (Chap. 8 Appendix B2 and Chap. 9 § 9b. 1; for an alternate presentation of (C), refer Thomas, Grafius and Hahn, 1971).

Using (C), we get,

$$d^2 = \frac{d_1^2 \sigma_2^2}{t^2} + \frac{d_2^2 \sigma_1^2}{t^2} - \frac{2 d_1 d_2 \sigma_{12}}{t^2}$$

The transformations (C) can also be expressed in terms of cos θ, the correlation coefficient of the variables X_1 and X_2.

$$\cos\ \theta\ =\ \sigma_{12}\ /\ \sigma_1\ \sigma_2 \text{ so that}$$

$$\cot^2\ \theta\ =\ \frac{\sigma_{12}^2}{\sigma_1^2\ \sigma_2^2\ -\ \sigma_{12}^2}$$

$$\text{We then get}\quad Y_1\ =\ \frac{X_1}{\sigma_1}$$

$$Y_2\ =\ \frac{-\ X_1\ \cot\ \theta}{\sigma_1}\ +\ \frac{X_2\ \operatorname{cosec}\ \theta}{\sigma_2}$$

$$\text{so that}\quad d^2\ =\ \left(\frac{1}{\sin^2\ \theta}\right)\left(\frac{d_1^2}{\sigma_1^2}+\frac{d_1^2}{\sigma_1^2}-\frac{2\ d_1\ d_2\ \cos\ \theta}{\sigma_1\ \sigma_2}\right)\ \ldots.\ \text{(E)}$$

which is an alternate form of (C). This makes it clear that in deriving (E) we are essentially adjusting for the correlation between the variables X_1 and X_2.

When the genetic distances between a number of genotypes are to be measured on the basis of a number of characters, it is usual to grow the genotypes in an appropriately replicated field design. The error dispersion matrix common to all the genotypes is then the appropriate matrix (*as error mean squares are* distributed as a multivariate normal distribution) for obtaining the transformation of X to uncorrelated Y. The rest of the process of obtaining D^2 is the same as described above (with an example of two variables X_1 and X_2).

We now attempt to answer the questions posed above in the light of published results on a variety of probelms relevant to plant breeding.

The major criteria for an efficient classification using D^2, as given by Rao (1952), (Chap. 9, § 9a·2 and 9a·3) are: (i) The distance must not decrease when additional characters are considered. (ii) The increase in distance by the addition of some characters to a suitably chosen set must be relatively small so that the group constellations arrived at on the basis of the chosen set are not distorted when additional characters are considered. (iii) Mahalanobis' generalised D^2 is applicable only when the variables (or measurements) are normally distributed.

While it may be difficult to pinpoint the characters that are to be necessarily included for an efficient classification of plant genotypes, it can be said from earlier experience that component characters that are important to fitness and natural selection provide usually a good choice (Murty and Arunachalam, 1966, 1967; Chandrasekariah, Murty and Arunachalam, 1969; Murty, Arunachalam and Jain, 1970). By logical arguments, it can be observed that functions of direct yield components may not add more scope to efficient classification when compared to themselves. A useful method to decide the set of characters to be

included, in any particular case, is to compare the average per cent contribution of D^2 added by each component character to the total D^2, when all possible D^2 among the genotypes are taken into consideration. Since only quantitative characters that can *a priori* be taken to follow a normal distribution are relevant to the D^2 function, one must be cautious before including discrete variables like intensity of pigment, presence or absence of an attribute like awning, glume covering etc., or internode position denoted by the sequential order from base (say from 1 to 10 etc.), arbitrary scores for disease or insect resistance, quality of grain etc., for classification. In some cases, appropriate transformations may restore the distribution to normality. But such cases should carefully be scrutinised before inclusion.

When characters are chosen to satisfy the criteria suggested above and when environments (defined in space or time) maintain the relative expression of characters with regard to the genotypes, to a great extent, it may be possible to obtain largely identical clustering pattern (as for example, in linseed, (Murty, Arunachalam and Anand, 1973)). But it is a safe strategy to expect non-repeatability and take adequate corrective steps by working out the divergence pattern afresh before chalking out useful breeding programmes. A compromise may, however, be possible, if, for instance, hybridisation between varieties belonging to clusters having the maximum inter-cluster divergence is only to be attempted, since the top and bottom clusters (and most of the varieties included in them) are likely to be largely repeatable unless environment causes a major change in the trend of D^2 values. Such events will, however, be uncommon.

When land races or appropriate sub-species, whose past history of evolution is recorded, are used, the clustering pattern based on D^2 along with the magnitudes of intra and inter-cluster distances can help, to some extent, in answering some questions related to their evolutionary pattern.

In general, divergence analysis is attempted to identify suitable parents for realising heterosis and recombination in breeding programmes. Vast literature is available relating genetic divergence with realised heterosis in crop plants, though definite conclusions could not emerge from them. It is a logical conclusion that most certainly, the parents are genetically divergent when heterosis is realised. But it does not follow that heterosis will result when parents are divergent. Cress (1966) has provided a theoretical analysis of heterosis in monogenic multi-allelic systems and Arunachalam (1977) in monogenic systems with inbreeding and digenic systems with epistasis. These studies underlined the fact that parents chosen to be genetically divergent through D^2 analysis, can fail, at times, to show high gene frequency differences, a major reason why heterosis fails to result on hybridisation. The inappropriate choice of the number and nature of characters, inadequate accounting of the environmental modification of character expression, inadequate field experimentation and sample sizes for recording character values, are a few of the common reasons for the above results.

A recent study on relating the degree of heterosis with genetic divergence (Srivastava and Arunachalam, 1977) in triticale has brought out some interesting results. In a 10 × 10 full diallel set of F_1 crosses, the divergence among the parents was classified into four classes based on the maximum (H = 14·3), and minimum (L = 3·5) distance, when the intra-and inter-cluster values of only those clusters which contained the 10 parents were considered, their mean (M=8), the mean of D-values between M and H (MH = 10) and that of M upto L (ML = 5·2). The five points delineated four distinct classes of genetic divergence (Table 1). The range of divergence among the clusters containing the parents was considered in preference to that among the parents themselves, since a breeder's decision to select parents will be based only on the distances among the clusters (and that is the essence of grouping the varieties into clusters also). The difference of the mean of the hybrid from that of the superior parent was tested for its statistical significance. Heterosis over the superior parent was calculated only when this difference was significant and in the desirable direction. Otherwise it was supposed to be absent. Heterosis was claisified into four classes by defining five points of division (based on the range of values of heterosis)

TABLE 1

Heterosis as related to genetic divergence in triticale (Adapted from Srivastava and Arunachalam, 1977)

		DC	NT		GW		YD		OV
			h	h+	h	h+	h	h+	h
L	—— 3.5								
	I	a	0	0	6	0	5	2	0
		b	6	0	2	0	0	0	0
ML	—— 5.2								
	II	a	17	4	25	8	7	0	13
		b	23	6	16	0	7	0	0
M	—— 8								
	III	a	62	16	46	23	59	31	54
		b	47	18	64	31	54	21	70
MH	—— 10								
	IV	a	21	12	23	9	29	15	33
		b	23	6	18	9	39	11	30
H	——14.3								

a=F_1 (year 1); b=F_1 (year 2); (for other symbols, see text).

following a method similar to that of classifying parental divergence. The percentage of heterotic crosses (h) falling in the 4 × 4 divergence - heterosis classes was then computed. The above analysis was done independently for the three characters, number of effective tillers (NT), 100-grain weight (GW) and single plant yield (YD). Further, the percentage of crosses heterotic for all the three characters, NT, GW and YD (abbreviated as OV in Table 1) was also scored for each divergence class (DC), in addition to the percentage of heterotic crosses (h+) which recorded heterosis above the mean heterosis value. In fact, they were the relatively superior heterotic crosses, providing prospective breeding material. We may call them 'super-heterotic' for convenience.

The following resulls of practical value emerge: (1) The divergence class I defined by the boundaries L and ML had negligible proportion of heterotic crosses and contained no super-heterotic ones.

(2) The maximum proportion of heterotic and super-heterotic crosses was found in the divergence class III bounded by M and MH.

(3) Though the divergence class IV defined by MH and H did contain heterotic and super-heterotic crosses, they were next only to the class III.

Despite the need for more confirmatory evidence, we can infer that too high a divergence may not produce the highest frequency of heterotic crosses. Nevertheless, the probability of realising a large number of fruittul heterotic crosses is high if we confine our hybridisation to parents with a divergence range above M (and particularly to class III).

The above observations are in tune with the relevant theoretical concepts of heterosis. In a single gene di-allelic system (Falconer, 1964), heterosis over mid-parent is a direct function of the square of gene frequency difference between populations from which the parents are drawn and in addition, dominance effect. The maximum gene frequency difference of unity, implies a cross of the type, aa × AA where A, a are the alleles, which is a super-heterotic cross, as expected. When, for example, a two-gene system is considered (Arunachalam, 1977) with unity gene frequency difference for each gene, H, the difference of the hybrid mean over mid-parent value = $h_1 + h_2 - i$ where h_1, h_2 are the dominance effects of the two genes and i, additive × additive interaction effect. With varying gene frequency differences of the two genes, all other epistatic effects may operate making the functional relationship between heterosis and the various genetic effects complex. With multi-genic systems more relevant to plant breeding, this relationship is too complex to analyse or interpret. This implies that realised heterosis can be modified by epistasis so that the frequency of super-heterotic crosses is reduced. This can be the case with respect to the divergence class IV in triticale, referred to above. Nevertheless, divergence classes IV and III are relevant to hybridisation programmes aimed at breeding for superior attributes.

With the spread of a number of high-yielding varieties, the task of breeding varieties outyielding the existing ones in yield, disease-pest resistance, quality and other attributes has become arduous. Further, large germplasm collections have

been built up and are available to the breeder. In the bid to generate genotypes possessing those attributes, the breeder would like to choose genetically distant parents for hybridisation. Even with preliminary scanning, the breeder, at times, has to choose those divergent parents from a fairly large number of entries (from germplasm collection or otherwise). The process of classification using D^2 can then go beyond practical reach.

A number of procedures for classification in such cases was tried (Murty, Arunachalam and Saxena, 1967) and a procedure combining principal component and D^2 analysis (Vairavan, Siddiq, Arunachalam and Swaminathan, 1973) was found appropriate. The procedure consists in a preliminary grouping based on the means of the first two canonical vectors (if they account for at least 70% of the total variation or more) and then finding the D^2 between all pairs of preliminary groups. The classification process using D^2 then becomes viable since the number of preliminary groups cannot usually be large. When they are large, the only alternative is to sub-divide them and treat each sub-division separately. Even when the first two canonical vectors account for a high proportion of variation, the simple two-dimensional representation of the multi-dimensional disposition of varieties cannot be as exact as Tocher's method of grouping (based on all possible inter-varietal D^2) which scans the full multi-diminsional space. The above procedure was found to work well for 194 collections of rice (Vairavan *et al.* 1973) and 160 world germplasm entries of peanut (unpublished). There is no ground so far to doubt the utility of this procedure for comparable situations in other crop plants.

It would thus be seen that genetic distance has a definite role to play in an efficient choice of parents of hybridisation programmes. With methods of classification (based on genetic distance) of a large number of genotypes becoming available, genetic divergence concepts should encompass a larger horizon of plant breeding in future.

REFERENCES

Arunachalam, V. (1977). Heterosis for characters governed by two genes. *J. Genet.*, **63:** 15–24.

Chandrasekariah, S. R., B. R. Murty and V. Arunachalam. (1969). Multivariate analysis of divergence in the genus Eu-Sorghum. *Proc. Nat. Inst. Sci. India. B.* **35:** 172–95.

Cress, C. E. (1966). Heterosis of the hybrid related to genefrequency differences between two populations. *Genetics*, **53:** 269–74.

Falconer, D. S. (1964). *An Introduction to Quantitative Genetics* (3rd Imp.) Oliver and Boyd, Edinburgh and London.

Jacquard, A. (1974). *The Genetic Structure of Populations.* Springer-Verlag Berlin, Heidelberg and New York.

Mahalanobis, P. C. (1930). On tests and measures of group divergence. *J. and Proc. Asiat. Soc. Bengal*, **26:** 541–88.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proc. Nat. Inst. Sci. India, B*, **2:** 49–55.

Murty, B. R. and V. Arunachalam. (1966). The nature of divergence in relation to breeding systems in some crop plants. *Indian J. Genet.*, **26A:** 188–98.

Murty, B. R. and V. Arunachalam. (1967). Factor analysis of genetic diversity in the genus Sorghum. *Indian J. Genet.*, **27:** 123–35.

Murty, B. R., V. Arunachalam and M. B. L. Saxena. (1967). Cataloguing and classifying a world collection of genetic stocks of sorghum. *Indian J. Genet.*, **27A:** 1–312.

been built up and are available to the breeder. In the bid to generate genotypes possessing those attributes, the breeder would like to choose genetically distant parents for hybridisation. Even with preliminary scanning, the breeder, at times, has to choose those divergent parents from a fairly large number of entries (from germplasm collection or otherwise). The process of classification using D^2 can then go beyond practical reach.

A number of procedures for classification in such cases was tried (Murty, Arunachalam and Saxena, 1967) and a procedure combining principal component and D^2 analysis (Vairavan, Siddiq, Arunachalam and Swaminathan, 1973) was found appropriate. The procedure consists in a preliminary grouping based on the means of the first two canonical vectors (if they account for at least 70% of the total variation or more) and then finding the D^2 between all pairs of preliminary groups. The classification process using D^2 then becomes viable since the number of preliminary groups cannot usually be large. When they are large, the only alternative is to sub-divide them and treat each sub-division separately. Even when the first two canonical vectors account for a high proportion of variation, the simple two-dimensional representation of the multi-dimensional disposition of varieties cannot be as exact as Tocher's method of grouping (based on all possible inter-varietal D^2) which scans the full multi-diminsional space. The above procedure was found to work well for 194 collections of rice (Vairavan *et al.* 1973) and 160 world germplasm entries of peanut (unpublished). There is no ground so far to doubt the utility of this procedure for comparable situations in other crop plants.

It would thus be seen that genetic distance has a definite role to play in an efficient choice of parents of hybridisation programmes. With methods of classification (based on genetic distance) of a large number of genotypes becoming available, genetic divergence concepts should encompass a larger horizon of plant breeding in future.

REFERENCES

Arunachalam, V. (1977). Heterosis for characters governed by two genes. *J. Genet.*, **63:** 15–24.

Chandrasekariah, S. R., B. R. Murty and V. Arunachalam. (1969). Multivariate analysis of divergence in the genus Eu-Sorghum. *Proc. Nat. Inst. Sci. India. B.* **35:** 172–95.

Cress, C. E. (1966). Heterosis of the hybrid related to genefrequency differences between two populations. *Genetics*, **53:** 269–74.

Falconer, D. S. (1964). *An Introduction to Quantitative Genetics* (3rd Imp.) Oliver and Boyd, Edinburgh and London.

Jacquard, A. (1974). *The Genetic Structure of Populations.* Springer-Verlag Berlin, Heidelberg and New York.

Mahalanobis, P. C. (1930). On tests and measures of group divergence. *J. and Proc. Asiat. Soc. Bengal,* **26:** 541–88.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proc. Nat. Inst. Sci. India, B,* **2:** 49–55.

Murty, B. R. and V. Arunachalam. (1966). The nature of divergence in relation to breeding systems in some crop plants. *Indian J. Genet.*, **26A:** 188–98.

Murty, B. R. and V. Arunachalam. (1967). Factor analysis of genetic diversity in the genus Sorghum. *Indian J. Genet.*, **27:** 123–35.

Murty, B. R., V. Arunachalam and M. B. L. Saxena. (1967). Cataloguing and classifying a world collection of genetic stocks of sorghum. *Indian J. Genet.*, **27A:** 1–312.

Murty, B. R., V. Arunachalam and O. P. Jain. (1970). Factor analysis in relation to breeding system. *Genetica*, **41:** 179–89.

Murty, B. R., V. Arunachalam and I. J. Anand. (1973). Effect of environment on the genetic divergence among some populations of linseed. *Indian J. Genet.*, **33:** 305–15.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometic Research.* John Wiley & Sons, New York.

Srivastava, P. S. L. and V. Arunachalam. (1977). Heterosis as a function of genetic divergence in triticale. *Z. Pflanzenzuchtg.*, **78:** 269–75.

Thomas, R. L., J. E. Grafius and S. K. Hahn. (1971). Transformation of sequential quantitative characters. *Heredity*, **26:** 189–93.

Vairavan, S ., E. A Siddiq, V. Arunachalam and M. S. Swaminathan. (1973). A study on the nature of genetic divergence in rice from Assam and North East Himalayas. *Theor. Appl. Geuet.* **43:** 213-21.

Appendix I

Let $\mathbf{X}=(X_1, X_2, \ldots . Xp)$ be p dependent characters based on which D^2 is computed. Let S be the variance-covariance matrix of **X**. To use the simple formula (A) for D^2, we transorm **X** to **Y**, a linear function in **X**. In fact, we find a matrix A so that $\mathbf{Y}=A\mathbf{X}$ has a unit variance-covariance matrix and $Y_1, Y_2, \ldots .$ Yp are mutually uncorrelated.

Now, variance-covariance matrix of $\mathbf{Y}=A\mathbf{X}$ is ASA′, where A′ is the transpose of the matrix A.

So, ASA′=I, a unit matrix. Multiplying on the left by A–1, and on the right by (A′)–1, we get

$S=A{-}1\ (A')=(A'\ A){-}1$ or $A^1A=S{-}1 \ldots .(i)$

Let be the vector of difference in the value of **X** between genotypes i and j,

$\mathbf{d}\approx X_i-X_j$

Let **e** be the corresponding vector for y

$\mathbf{e}=Y_i-Y_j$; so that $\mathbf{e}=A\mathbf{d}$

Then the distance between the genotypes i and j is given by

$D^2=\mathbf{e}'\mathbf{e}$ ....(ii)

$=d'A'A\mathbf{d}=d'S{-}1\mathbf{d}$ by (i)

We use the formula (ii) for computing D^2

Let us take the example, considered in the text, with the variables X_1 and X_2.

$$S=\begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$$

The following are the steps in the pivotal condensation method to get uncorrelated linear functions, Y.

S-matrix	Unit matrix		
$\sigma_1^2 \quad \sigma_{12}$	1	0	(1)
$\sigma_{12} \quad \sigma_2^2$	0	1	(2)
$1 \quad \frac{\sigma_{12}}{\sigma_1^2}$	$\frac{1}{\sigma_1^2}$	0	(3) = (1) ÷ σ_1^2 (Pivotal rows)
$0 \quad \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}$	$-\frac{\sigma_{12}}{\sigma_1^2}$	1	(4) = (2) — [(3) × σ_{12}]

$$\text{Var } (Y_1) = \sigma_1^2$$

$$\text{Var } (Y_2) = \sigma_2^2 - \frac{\sigma_{12}^2}{\sigma_1^2}$$

The uncorrelated y's are given by

$$Y_1 = \frac{1}{\sqrt{\text{Var } (Y_1)}} X_1 \quad \text{from the right side of (1)}$$

$$Y_2 = \frac{1}{\sqrt{\text{Var } (Y_2)}} \left(\frac{-\sigma_{12}}{\sigma_1^2} \right) X_1 + \frac{X_2}{\sqrt{\text{Var } (Y_2)}} \quad \text{from the right side of (4)}$$

On rearranging the terms, we get equation (C).