

A PROPOSAL FOR ANALYSIS OF GENETIC DIVERGENCE AMONG GERMPLASM BANK ACCESSIONS

Cosme Damião Cruz¹, Antonio Vander Pereira² and Roland Vencovsky³

ABSTRACT

In view of the difficulties commonly encountered in the analysis of germplasm bank data, when large numbers of accessions are involved and several traits are evaluated, we propose a system of analysis for accession characterization in which the original group is divided into subgroups according to the traits considered to be of greatest importance by the breeder. After subdivision, divergence is evaluated within the group considered to be elite (most important for the breeder) and between this group and the others. This system of analysis permits dynamic manipulation of germplasm bank data, and the formation of groups according to the breeder's criteria leads to considerable simplification of data computation and efficient synthetization of the information available.

INTRODUCTION

Germplasm banks play an extremely important role in breeding programs as deposits of genetic variability available to the breeder. The materials can be used in a direct manner as commercial varieties or in hybridization programs for the creation of new cultivars. Thus, bank accessions should be evaluated for the more important botanical-agronomic traits in order to facilitate the work of breeders.

In germplasm banks, besides their being a large number of accessions, there is little reliable information about the degree of importance, phenotypic stability and variation in the state of each descriptor, so that generally it is necessary to evaluate a large

¹ Departamento de Biologia Geral, Universidade Federal de Viçosa, 36570 Viçosa, MG, Brasil. Send correspondence to C.D.C.

² EMBRAPA/CPAC, Caixa Postal 70.0023, 73300 Planaltina, DF, Brasil.

³ Departamento de Genética, ESALQ/USP, Caixa Postal 83, 13400 Piracicaba, SP, Brasil.

number of traits. As a consequence, a multitude of experimental data is generated which is difficult to interpret, mainly because of the scarcity of analysis methods that will permit the synthetization of this information to suit the purposes of breeding.

On the basis of the above considerations, multivariate cluster analysis for germplasm data are an important option because they permit the grouping of individuals with a high pattern of similarity while simultaneously taking into consideration the entire set of descriptors evaluated.

Two basic procedures have been used for accession clustering. The first involves accession dispersal in a coordinate system, using the first principal components or the first canonical variables as axes (Rao, 1952). The efficiency of this technique depends on the concentration of a large part of the total variation within a few components.

The second procedure uses clustering techniques based on measurements of similarity between pairs of accessions. However the dimension of the matrix can be a problem. For both procedures, an elevated number of accessions impairs analysis because of the difficulty in identifying the individuals and their groups in dispersal plots and the excessive computation time needed to process a large volume of data, a large part of which is of no interest to the breeder.

The objective of our study was to develop a system for analysis and interpretation of germplasm bank accessions that allows considerable simplification of data processing and provides efficient synthetization of available information.

METHODOLOGY

The proposed analysis system is based on the fundamental strategy of division of the original group into various subgroups, one of which is considered to be elite (i.e., of greater interest to the breeder). The information on this elite group is then maximized. The method is presented schematically in Figure 1.

The analysis is performed in two stages: a) subdivision of the original group according to criteria established by the breeder, and b) quantification of genetic divergence within the elite group and between this group and all others. Additionally, studies of the divergence relationship between the elite group and the others may be performed on the basis of genetic, economic, or other values.

a) Subdivision of the original group

At this stage the original group is subdivided on the basis of the descriptors judged by the breeder to be of major importance for his purposes (e.g., yield, precocity

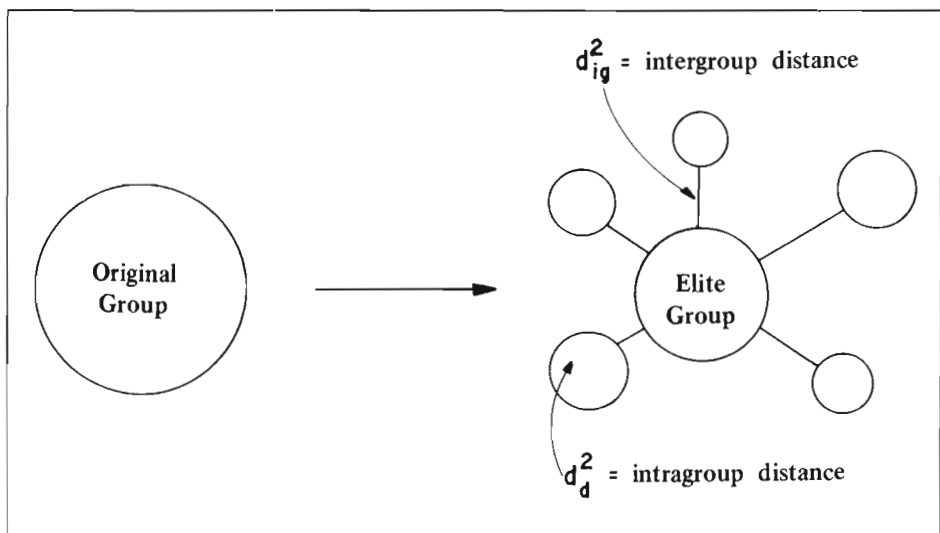


Figure 1 - Division of the original group into subgroups on the basis of criteria established by the breeder.

etc.) (see Pereira, 1989). In addition to the arbitrary choice of traits, the number of subgroups into which the original group will be partitioned is also defined *a priori*. The objective of this definition is to allocate a basic number of accessions to each subgroup and especially to the elite group, according to the working capacity and the objectives of the breeder.

Various procedures may be adopted to assign accessions to the various subgroups. An easy method which takes into consideration the economic importance of each variable is to establish an index which will be used to construct class intervals for subgroup delimitation. The index is established by combining the variables indicated by the breeder and it may be constructed according to the classical selection index theory proposed by Smith (1936) and Hazel (1943), which is based on estimates of phenotypic and genotypic covariance matrices among the traits considered and on the establishment of economic weights. Other indices which do not require the need to estimate covariance matrices may also be adopted. These estimates are often impossible, especially in trials involving a large number of germplasm bank accessions and carried out without replications. Within this context, the best indices would be the multiplicative indices proposed by Subandi *et al.* (1973), the rank sum index proposed by Mulamba and Mock (1978), and the base index proposed by Williams (1962).

b) Divergence in relation to the elite group

The degree of dissimilarity, i.e., the measurement of genetic divergence between accessions, is evaluated by the statistical parameters commonly used for this purpose, i.e., the square of the Euclidean distance or the Mahalanobis distance (Sneath and Sokal, 1973).

The aim is to evaluate the genetic divergence within the elite subgroup and between this subgroup and all others, using the following parameters:

p = number of traits evaluated

n = number of subgroups

n_k = number of individuals in group k

X_{ijk} = mean phenotypic value in relation to within-group individual replications referring to individual i ($i = 1, 2, \dots, n_k$) and to trait j ($j = 1, 2, \dots, p$) in group k ($k = 1, 2, \dots, s$)

X_{jk} = mean of the j th trait in group k

$$X_{jk} = \frac{1}{n_k} \sum_{i=1}^{n_k} X_{ijk}$$

$\sigma_{jj'k}$ = phenotypic covariance between traits j and j' within group k , or element of the positive defined T_k matrix of variances and covariances estimated for each group ($\sigma_{jj'k} = \sigma_{jk}^2$).

For the cases for which replications are available, the following items should also be considered:

$\alpha_{jj'}$ = element of the jj' order of the Φ^{-1} matrix;

Φ = positive defined residual covariance matrix common to all experimental units.

Matrix Φ is indispensable only for the calculation of generalized Mahalanobis distances. However, when it is not possible to estimate it, divergence within the elite group and between this group and all others may also be estimated by the square of the Euclidean distance, which only requires the estimate of the elements of the T_k phenotypic covariance matrix. In the expression for the calculation of these distances, the elite group is referred to as $k = 1$, as presented below. Mean divergence within the elite group is estimated by:

$$d_1^2 = \frac{1}{[n_1 (n_1 - 1)/2]} \sum_{i < i'}^{n_1} \sum^{n_1} d^2_{ii'} \quad (1)$$

or, in an analogous manner:

$$D_1^2 = \frac{1}{[n_1 (n_1 - 1)/2]} \sum_{i < i'}^{n_1} \sum^{n_1} D^2_{ii'} \quad (2)$$

where:

d_1^2 and D_1^2 , respectively refer to the mean Euclidean and Mahalanobis distance existing within the elite group.

Expressions (1) and (2) can be replaced by:

$$d_1^2 = 2 \sum_{j=1}^P \sigma_{j1}^2 \quad (3)$$

$$D_1^2 = 2 \sum_{j=1}^P \sum_{j'=1}^P \sigma_{jj'1} \alpha_{jj'} \quad (4)$$

respectively.

The intergroup distance (d_{1k}^2) can be estimated as follows:

$$d_{1k}^2 = \frac{1}{n_1 n_k} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_k} d^2_{ii'} \quad (k \neq 1)$$

or

$$d_{1k}^2 = d_{(1)(k)}^2 + \frac{n_1 - 1}{n_1} \sum_{j=1}^P \sigma_{j1}^2 + \frac{n_k - 1}{n_k} \sum_{j=1}^P \sigma_{jk}^2 \quad (5)$$

where:

$d_{(1)(k)}^2$ is the Euclidean distance between 1, considered to be elite, and group k, on the basis of the mean p variables.

Algebraically, we have:

$$d_{(1)(k)}^2 = \frac{P}{\sum_{j=1}^P} (X_{j1} - X_{jk})^2 \quad (6)$$

The $d_{(1)(k)}^2$ statistical parameter, although estimating the intergenotypic distance between group 1 (elite) and all others, does not represent a measurement that can be compared since it is influenced by the number of accessions belonging to each group and by overall group variance. It only expresses a measurement of divergence between the epicenters of the compared groups.

Expressions analogous to (5) and (6) can be deduced when the Mahalanobis distance is utilized, i.e.:

$$D_{ik}^2 = \frac{1}{n_1 n_k} \sum_{i=1}^{n_1} \sum_{i'=1}^{n_k} D_{ii'}^2$$

or

$$D_{ik}^2 = D_{(1)(k)}^2 + \frac{n_1 - 1}{n_1} \left(\sum_j^p \sum_{j'}^p \sigma_{jj'} \alpha_{jj'} \right) + \frac{n_k - 1}{n_k} \left(\sum_j^p \sum_{j'}^p \sigma_{jj'k} \alpha_{jj'} \right)$$

where:

$$D_{(1)(k)}^2 = \delta' \Phi^{-1} \delta$$

$$\delta = [l_1, l_2, \dots, l_P]$$

and:

$$l_j = X_{j1} - X_{jk} \text{ for each } j.$$

Relationship between divergence relative to the elite group and economic value of the group

Since there is information about the divergence of each group in relation to the elite group and since the economic or even the genetic value of the groups is known, it is possible to determine the interrelationships between these two variables. In this case, we recommend the use of polynomial regression analysis, in which the dependent variable is intergroup divergence and the independent variable is the economic value of the group.

The possibility of predicting genetic divergence in relation to the elite group on the basis of the economic value of the remaining groups could be of great value to breeders.

DISCUSSION

The interest of breeders in the use of measurements of genetic similarity as a parameter for the indication of parental lines to be used in crosses is based on the biometric relationship between the heterosis manifested in hybrids and the divergence in the gene frequencies of the parents (Falconer, 1981). More efforts have been devoted to the study of genetic divergence after proof was obtained for the existence of significant correlations between parental diversity and hybrid performance in different crops (Maluf *et al.*, 1983; Ghaderi *et al.*, 1984; Anand and Rawat, 1984; Shansuddin, 1985; Smith and Smith, 1987; Cruz, 1990).

The study of genetic divergence in a relatively large set of accessions available in germplasm banks, although indispensable, has been carried out in a limited manner since, in a situation of this type, traditional methodologies turn out to be inefficient because of the excessive time needed for computation of parental data, and for the processing of large volumes of observations which for the most part are of no interest to the breeder, and because they do not provide practical and concise information about the experimental results.

The analysis system presented here has great advantages both from a practical and a methodological viewpoint. It is valuable for the manipulation and interpretation of data from studied of the characterization of germplasm bank accessions in which the number of individuals evaluated is relatively large.

The methodological advantages are related to the simplicity of the calculation of the statistical parameters used, and to the effective reduction in data volume without a loss of information related to the group considered to be elite, since all attention is concentrated on this group.

The simplification afforded by the method can be better illustrated by considering, for example, an analysis of trials involving 300 accessions. In this situation, multivariate clustering techniques require the estimate of the degree of similarity between 40850 pairs of accessions for which sophisticated computational resources and a large amount of processing time would be needed. In addition, once the estimates are obtained, it would be very difficult to summarize them in such a way as to maximize the interests of the breeder.

The inconvenience involved in the analysis of a large number of accessions using traditional methodologies is apparent in a study conducted by Hussaini *et al.* (1977) on the clustering of 640 genotypic materials of *Eleusine coracana* L. These investigators

used subjective criteria and of score dispersal of the first two principal components (which involved only 55.9% of overall variation) to cluster the genotypes into 12 groups.

By applying the methodology proposed here to these 300 accessions, this data would be divided, according to the breeder's interest, into smaller subgroups such as ten subgroups each with about 30 accessions. After the groups were established, both the intra- and intergroup divergences in relation to the elite group could be evaluated simply on the basis of the variances of the traits within the subgroups (which are easy to estimate), the number of individuals per subgroup and the calculation of a single distance based on the subgroup mean.

Quantification of genetic divergence among cultivars is important for breeding programs because it involves the heterosis manifested in hybrids for those traits that exhibit dominance. Furthermore, there one expects more superior individuals in segregant generations originating from crosses between highly divergent parents.

Thus, in breeding programs involving hybridization, special care needs to be taken to choose parents which not only exhibit divergence but which also have superior traits of economic importance, since superior lines can be extracted more easily from improved rather than non-improved populations.

In view of the above considerations, the initial division of a group of genotypic materials on the basis of traits considered to be of high economic value, in addition to permitting easier data processing, is a fundamental step in maximizing the success of a breeding program and provides a more dynamic use of germplasm bank data by permitting the breeder to form accession groups.

The clustering of 280 manioc accessions of the EMBRAPA Germplasm Bank (BAGM, Cruz das Almas, State of Bahia) conducted by Pereira (1989) is an example of the application of the methodology proposed here. The use of Euclidean distances as a measure of dissimilarity and of the proposed clustering method confirmed the usefulness of the procedure and emphasized its operational ease and efficiency.

ACKNOWLEDGMENTS

Publication supported by FAPESP.

RESUMO

Considerou-se as dificuldades de análise, comumente encontradas em dados de bancos de germoplasma, onde são envolvidos um grande número de acessos e avaliados vários caracteres. Para estes casos, foi proposta uma sistemática de análise, destinada à caracterização dos acessos, na qual o grupo original é dividido em subgrupos, de acordo com aqueles caracteres julgados de maior importância pelo melhorista. Após a subdivisão, avalia-se a divergência dentro do grupo considerado elite (mais importante para o melhorista) e

deste, com relação aos demais. A sistemática de análise possibilita uma manipulação dinâmica dos dados do banco de germoplasma, sendo que a arbitrariedade do melhorista no estabelecimento dos grupos, proporciona considerável simplificação na computação dos dados e uma sintetização eficiente das informações disponíveis.

REFERENCES

- Anand, I.J. and Rawat, D.S. (1984). Genetic diversity combining ability and heterosis in brown mustard. *Indian J. Genet.* 44: 226-234.
- Cruz, C.D. (1990). Aplicação de algumas técnicas multivariadas no melhoramento de plantas. Doctoral Thesis, ESALQ-USP, Piracicaba.
- Falconer, D.S. (1981). *Introdução à Genética Quantitativa*. (Tradução de Silva, M.A. e Silva, J.C.) Viçosa, UFV. Imprensa Universitária, pp. 279.
- Ghaderi, A., Adams, M.W. and Nassib, A.M. (1984). Relationship between genetic distance and heterosis for yield and morphological traits in dry edible bean and faba bean. *Crop Sci.* 24: 24-27.
- Hazel, L.N. (1943). The genetic basis for constructing selection indices. *Genetics* 28: 476-490.
- Hussaini, S.H., Goodman, M.M. and Timothy, D.H. (1977). Multivariate analysis and geographical distribution of the world collection of finger millet. *Crop Sci.* 17: 257-263.
- Maluf, W.R., Ferreira, P.E. and Miranda, J.E.C. (1983). Genetic divergence in tomatoes and its relationship with heterosis for yield in F₁ hybrids. *Rev. Bras. Genet.* 3: 453-460.
- Mulamba, N.N. and Mock, J.J. (1978). Improvement of yield potential of Eto Blanco maize (*Zea mays* L.) population by breeding for plant traits. *Egypt J. Gen. Cytol.* 7: 40-51.
- Pereira, A.V. (1989). Utilização de análise multivariada na caracterização de germoplasma de mandioca (*Manihot esculenta* Crantz). Doctoral Thesis, ESALQ/USP, Piracicaba.
- Rao, R.C. (1952). *Advanced Statistical Methods in Biometric Research*. (John Wiley and Sons.) pp. 390.
- Shansuddin, A.K.M. (1985). genetic diversity in relation to heterosis and combining ability in spring wheat. *Theor. Appl. Genet.* 70: 306-308.
- Smith, D.F. and Smith, J.S.C. (1987). Prediction of heterosis using pedigree relationship, biochemical and morphological data. In: *23rd Annual Illinois Corn Breeders School*, Illinois, p. 1-21.
- Smith, H.F. (1936). A discriminant function for plant selection. *Ann. Eugen.* 7: 240-250.
- Sneath, P.H. and Sokal, R.R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*. W.H. Freeman and Co. pp. 573.
- Subandi, W., Compton, A. and Empig, L.T. (1973). Comparison of the efficiency of selection indices for three traits in two variety crosses of corn. *Crop Science* 13: 184-186.
- Williams, J.S. (1962). The evaluation of a selection index. *Biometrics* 15: 375-393.

(Received December 7, 1990)