Some proposed procedures for obtaining a core collection using quantitative plant characterization data

S. Hamon and M. Noirot
Laboratoire de Ressources Génétiques et Amélioration des Plantes Tropicales
ORSTOM, B.P. 5045, Montpellier, France

Summary

As well as maintaining a base collection, an optimal plant genetic resources preservation system must take into account the development and use of small collections including a large amount of diversity.

Several procedures have been suggested for obtaining this, ranging from a random sampling of 10% of the samples to sampling based on a good knowledge of population allelic frequencies.

A large part of available data for a crop usually concerns morphological and/or phenological descriptors of a quantitative nature, and in this paper we propose a statistical procedure based on the conversion of the initial variates into new independent ones chosen to represent maximum variability. Then, by a step by step search of the most dissimilar individuals, we are able to choose the number of accessions and/or the percentage of the total diversity to be conserved.

Introduction

The idea of centres of origin and domestication of plants, and the consequences for plant breeding, is mainly associated with the pioneer work of Vavilov (1935), who was also one of the first scientists to collect plant genetic resources on a scientific basis.

Several decades later, prompted by Harlan (1970), Frankel and Bennet (1974) and Frankel (1974), scientists began to study the diversity of natural populations. After early work using morphological characters, isoenzymes were used by ORSTOM to carry out studies on several species, such as *Coffea*, *Oryza*, *Panicum* and *Pennisetum* (Pèrnes, 1984).

Under the auspices of IBPGR many field collections of cultivated plants of economic importance were carried out. Progressively during the eighties, interest turned to related wild species and to minor crops.

This systematic collection resulted in a number of problems in conservation and management. Curators were faced with massive, sudden and irregular influxes of material, which was kept in cold rooms or in freezers. As underlined by Peeters and Williams (1984), documentation concerning accessions was often lacking. In addition, most collections are now large, difficult to handle and manage, and consequently are underused. Finally, curators are unable to answer specific questions and, as a result, breeders often prefer to use (or build up) a working collection composed of several tens of well-known accessions or varieties.

Frankel and Brown (1984) first highlighted the need for a reference collection containing a good representation of the available diversity, the "core collection". The main objectives are to obtain an accession with a given profile from a reduced set of accessions, or, by default, to guide the breeder within the base collection. These authors stress that the statistical sampling procedures must be correct and ensure representation and preservation of the population genetic structure. It is evident that sampling will lead to a partial loss of diversity, whatever the extent of that loss may be. The questions are what is the real level of the loss and what do we want to retain in the core collection. Brown (1980) related in detail a few examples where allozymic frequencies and the genetic structure of the populations were known. His procedure seemed to be useful because with 10% of the total accessions it was possible to conserve 80% of the allozymic diversity.

The genetic diversity of species is organized on different levels of complexity. When we study one cultivated species using a descriptor list and then consider related wild species, it is frequently the case that their mating systems, ploidy levels, distribution areas and cycle lengths are different (Hamrick *et al.*, 1979). While isozymic markers are often useful, they are not necessarily the best. While in many situations there is a good correlation between allozymic and morphologic polymorphism (Giannisi and Crawford, 1986), this is not always the case (Davis and Gilmartin, 1985). A low level of allozyme variability could be associated with a high level of morphological diversity which is of great interest for the farmer or the plant breeder. In addition, morphological divergence between populations can occur before allozyme divergence. Crawford (1985) gives examples of recent speciations where there is often no or a low level of allozyme difference.

Brown (1989) shows that when the genetic structure of a population is not known, random sampling is better than other sampling strategies and also that if the choice criteria used to select accessions are inadequate, major problems can occur. We suggest here a sampling strategy based on "passport data". The main problem would be that these data are often missing, not homogeneous or valid for all the accessions.

For most tropical crops, detailed data, especially at the biochemical level (isoenzymes, RFLP), are rarely available. Most data are botanical, morphological, agronomical or related to pests and diseases using tools such as the IBPGR descriptor lists.

In most cases, characterization of a collection in developing countries has employed less costly descriptors of this kind, so data refer to quantitative characters such as plant height, flowering period and fruit production and/or qualitative characters such as the shape and colour of different parts of the plant. Quantitative data can, under certain conditions, be analyzed using multivariate analysis such as principal component analysis (Hotelling, 1933). Qualitative data can be analyzed in a similar way by factorial analysis (Benzecri, 1973), where the data are coded by presence or absence. These two procedures are slightly different and, in this paper, we introduce the approach that can be used for quantitative data to show how, by a stepwise analysis which consists of searching for distant individuals, it is possible to assemble a core collection.

Strategy and pathway analysis

Principle

The variability of a population is defined by differences between individuals for one or more characters. To conserve maximum diversity it is necessary to retain the largest differences. We thus first search for the individual which is furthest from the centroid. To build up the working samples, we add those individuals which maximize the inertia of the sample (see below).

Prior data conversion

The initial data set is a table (individuals x variates) where individuals are the accessions and the variates the descriptors. The choice of the metric for distance calculations depends on the nature of the variates. For quantitative data we choose the Euclidian distance weighted by the standard error (to make the relative weight of each variable uniform) or the Mahalanobis distance (1930). For discrete variates, the khi-2 distance or Nei's distance (1972) are most suitable.

The value of a distance is greatly influenced by the number of differences. If these belong to strongly correlated characters (positively or negatively) the effect would be to double the weighting given to a single factor. For this reason the analysis is carried out on standardized principal component scores. This procedure clarifies the structure of the data. New factors, equal in number to the initial number of variables, are chosen in order of decreasing inertia value (% variability accounted for). All are orthogonal to each other. Each individual is then characterized by its factor axis scores. Thus, we have a new data set where variables are independent and where a given individual has a value (positive or negative) relative to the centroid. It may then be possible to calculate distances weighted by the standard deviation of the factor, i.e. the square root of the eigenvalue.

This procedure is very useful for eliminating data redundancy due to the correlation that exists between variates. One remaining drawback is that all factors taken into account are given the same weight. This could lead to bias due to random variation or errors in the initial data set. For this reason we decided to consider only the axes for which the eigenvalue (lambda) is greater than 1. Another possibility consists of introducing a dummy variable in the original set and to conserve only the axes for which the eigenvalue is equal to or larger than that for the dummy.

Making up the core collection

For each individual, characterized by a set of axis scores, we calculate the sum (P) of the squares of the standardized coordinates for the k factors selected: $P = \text{Sigma } x_{ij}^2$. The ratio P/N (where N is the total number of individuals) is the inertia of a given individual. The sum of the partial inertias of the N points is the total inertia (100%). The relative contribution of one individual is the ratio P/(N*K). In the same way, for a subset of individuals, the contribution is equal to Sigma P_i(N*K). This is the selected diversity (SD).

For this selection procedure, we first search for the individual which makes the maximum contribution to total inertia. Then, among the rest, we search for the individual which with the former gives the maximum SD value, and so on. At each step, SD is estimated, so it is possible to stop at any level between a few % and 100% of total inertia. A selection could also be made on a preselected number of individuals.

An example with okra using principal components analysis

1. *Main factorial axis*

The best data set for this purpose is certainly one that contains a maximum number of individuals studied in similar conditions, but for our purpose we have chosen the data for the actual core collection, arbitrarily limited to 152 individuals, and have removed non-quantitative data.

The coded names of the variables are as follows: first flowering day (FFD), plant height (PLHT), number of nodes on the main stem (NNS), stem diameter at the stem base (DIAM), first fruiting node (N1FR), flowering amplitude (FLAM), pod length (LGPD), pod width (WIPD), number of ridges (NBRI), 1000 seed weight (TSW), and seed production (SDPR). The sign (-) just in front of a variable indicates that its contribution to a principal component is opposite to that of the others.

If we retain only axes with a lambda value equal to or greater than 1, axes 1 to 4 are valid, if the critical lambda value is 0.5, axes 1 to 6 are valid. For axis number 1, major contributions are associated with FFD, N1FR and NNS, followed by DIAM and (-) TSW. For axis 2, major contributions are associated with the following variables: (-) WIPD, LGPD, (-) NBRI; and for axis 3, PLHT, (-) PRDG. Axis 4 is mainly associated with (-) FLAM but also with the remaining part of PLHT and (-) SDPR not accounted for by axis 3.

All 11 original variables are taken into account in the system defined by axes 1 to 4. Several variables are correlated in the same direction for a number of axes. This indicates that some descriptors are less informative than others and that they could be removed. Without going into the details of the analysis, we can see that axis 1 reflects plant precocity, the initial vigour and the ability to produce nodes and light seeds. Axis 2 is mainly associated with pod characteristics, with pod width and ridge number inversely correlated with pod length. On axis 3, plant height and seed production are inversely correlated.

2. *Increase of the selected variability*

At the beginning, the data set constituted 152 individuals. The progression of SD is shown in Fig. 1. Thus we can see that with 15% of the individuals we have selected 30% of the inertia. We reach a level of 50% of total inertia with 30% of the individuals. As a result of the decrease of the slope of the curve, the further addition of new elements becomes progressively less useful.

3. *Position, on the axis, of the selected individuals*

By definition, the axis which corresponds to the higher level of inertia is horizontal and the next vertical. Each such factorial plan is divided in four sectors called, by convention, A (-+), B (++), C (+-) and D (--). Signs in brackets correspond to negative or positive values on the axis, ordered by their decreasing value. For example A (-+) means a negative value for axis 1 and a positive one for axis 2. The distribution of the samples is shown with the total number of individuals by quarter and the relative percentage in the sample.

a) Plan (1 X 2) - A (47, 42.5%); B (31, 41.9%); C (41, 41.4%), D (33, 27.2%).

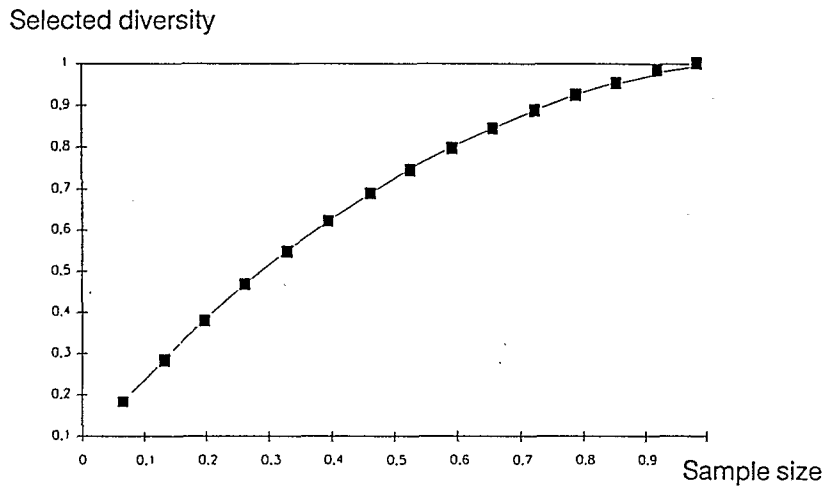

Fig. 1. The relationship between selected diversity and sample size

The sampling result in each quarter, except for D, is well balanced. In Fig. 2, we can see that the sampling is very good in the peripheral area but there is an important zone next to the centre without selected individuals.

b) Plan (2 X 3) - A (45, 33.3%), B (39, 43.5%), C (38, 42.1%), D (30, 36.6%).

Here we also observe a low density of selected individuals near the centre and a small disequilibrium in the quarters, where the limits are 33% and 43%.

4. *The picture given by clustering analysis*

The step by step analysis of each factorial plan gives an interesting but restricted view. A more general appreciation is obtained by a clustering analysis made on the first 4 factorial axis. The clustering procedure is the variance criteria on weighted Euclidian distances.

The dendrogram is characterized by a division into 3 main clusters. The following symbols are used: C = cluster number, S = number of sub-cluster in a given cluster, I = number of individual in a cluster and P = the sampling percentage in a cluster. The dendrogram, when the total number of selected individuals is limited to 60, can be summarized as follows: (C1-S7-I36, P 30.5%); (C2-S20-I44, P 52.3%); (C3-S14-I72, P 36.1%). Thus, we observe that when a cluster becomes complex (S7 < S14 < S20) the sampling percentage increases (31, 36, 52). If the total number of individuals is limited to 30, then the breakdown is C1 (20%), C2 (56%), C3 (24%) and again the choice is made according to the internal structure of the cluster.

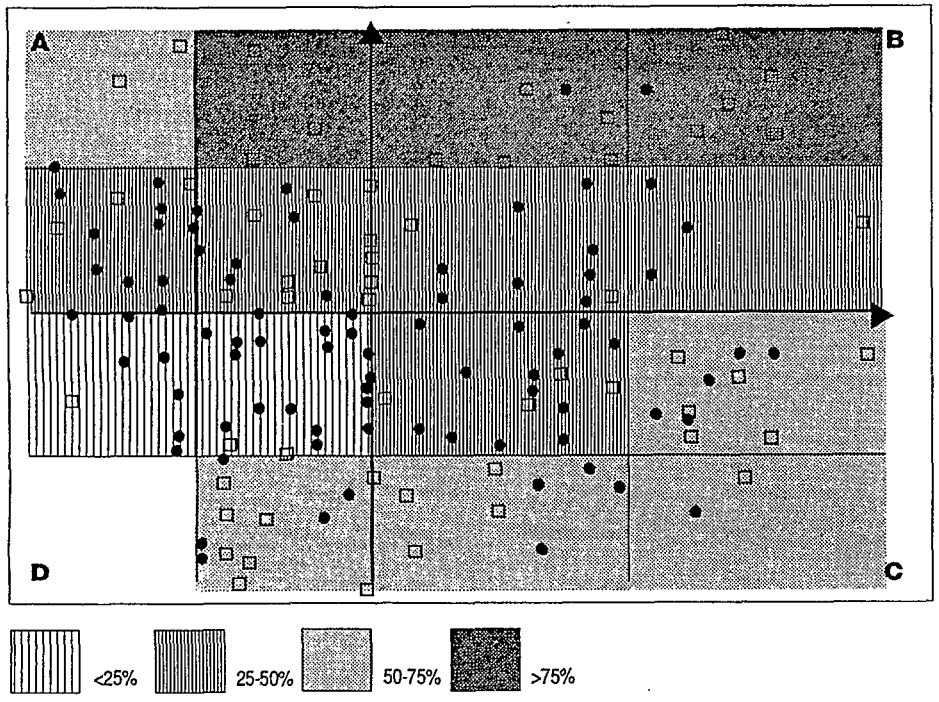

Fig. 2. PCA scatter plot showing position of individuals with reference to principal components I and II. Squares represent individuals selected by the inertia procedure

Discussion

A method has been described which permits statistical selection of a sample of a given set of individuals, based on selecting maximum variability using quantitative data.

First, the initial data set (accessions x descriptors) containing more or less correlated variables is transformed to an independent axis system where individuals are characterized by their scores defined on each axis. We then identify a sample which contains the maximum existing variability by a step by step aggregative procedure. Choice criteria allow us to fix the number of individuals or the percentage of the total variability retained in the sample.

In the example studied, this method gives a sample which is equally distributed in the sectors defined by the factorial plans. However, the percentage of selection is greater in areas further from the factorial axes. The clustering analysis, which gives a more synthetic view, shows that more individuals are selected from the more complex clusters.

This procedure can be criticized on the following points:

- contrary to random sampling, it does not take into account the central area of diversity. The clustering analysis gives a first partial answer. However, it seems likely that it will be easier to regenerate the diversity of the central part using peripheral individuals than the converse;
- it is also possible to argue that the heritability of the variables used is not well-known. Such criticism is open to experimentation but does not affect the selection procedure *per se*. However, when the level of heritability of characters used is known, it is better to use those with high heritability.

The method described provides a possible way of improving the way the choice of a core collection is made. To obtain a really efficient method qualitative data should also be included in the analysis.

November 1991

INTERNATIONAL BOARD FOR PLANT GENETIC RESOURCES

REPORT
OF AN INTERNATIONAL WORKSHOP ON OKRA GENETIC RESOURCES

held at

National Bureau for Plant Genetic Resources (NBPGR)
New Delhi, India

8-12 October 1990

IBPGR
Rome, 1991

The International Board for Plant Genetic Resources (IBPGR) is an autonomous international scientific organization under the aegis of the Consultative Group on International Research (CGIAR). The basic function of IBPGR is to foster the collecting, conservation, documentation, evaluation and use of plant germplasm and thereby contribute to raising the standard of living and welfare of people throughout the world. Financial support for the core programme is provided by the Governments of Australia, Austria, Belgium, Canada, China, Denmark, France, Germany, India, Italy, Japan, the Netherlands, Norway, Spain, Sweden, Switzerland, the UK, the USA and the World Bank

Citation:
IBPGR. 1991. International Crop Network Series. 5. Report of an International Workshop on Okra Genetic Resources. International Board for Plant Genetic Resources, Rome

ISBN 92-9043-210-1

IBPGR Headquarters
Via delle Sette Chiese 142
00145 Rome
Italy