Theor Appl Genet (1989) 78:42–48

Hierarchical cluster analysis as a tool to manage variation in germplasm collections

J.P. Peeters [1] **and J.A. Martinelli** [2]

[1] Department of Applied Biology, Cambridge University, Pembroke Street, Cambridge CB2 3DX, UK
[2] AFRC-IPSR, Department of Cereal Research, Maris Lane, Trumpington, Cambridge CB2 2JB, UK

Received January 2, 1989; Accepted February 6, 1989
Communicated by P.M.A. Tigerstedt

Summary. The potential of using hierarchical cluster analysis to classify entries from a germplasm collection according to their degree of similarity was assessed. Results suggest that similarity is generally greatest among individual entries by country of origin and that hierarchical cluster analysis could be used as a tool to classify entries from germplasm collections according to their respective gene pools, even when no passport data are available. Based on this technique, it is also shown that the segregative potential of entries can be estimated.

Key words: Cluster analysis – Germplasm collections – Variability – Relatedness

Introduction

With the increasing size of germplasm collections, methods to classify and order their variability are becoming steadily more required. Samples in such collections are generally described by a mixture of both qualitative and quantitative information which reflect, to an extent at least, their genetic characteristics. Principally because of G × E interactions, univariate approaches to exploit such data can lack precision (J. P. Peeters et al., unpublished results). Precision can be increased by replicating the estimates for the given variable, or by increasing the number of variables describing the characteristics of the sample. In the latter case, however, the treatment of the data also becomes more complex.

One obvious approach to order highly variable samples such as those found in germplasm collections is by using established computer algorithms developed in the field of multivariate statistics. Multivariate algorithms, such as hierarchical cluster analysis which determine on an entry basis the similarity of a given entry relative to other entries, appear particularly useful.

This particular statistic has several advantages. First, it allows the mixing of both qualitative and quantitative data and, therefore, all the available information on the sample can be utilized. In addition, each entry is treated as an individual entity of equal weight in the analysis, contrary to a number of other multivariate techniques which are based on the variation of groups of entries. The statistic has been used in widely different fields. For example, Lin et al. (1986) have shown the advantages of using hierarchical cluster analysis in stability analysis over a range of environments, particularly by the inclusion of reference cultivars. Murphy et al. (1986) have used cluster analysis using coefficients of parentage for red winter wheat to determine overall patterns of relationship between existing cultivars in the US. Lebeda and Jendrulek (1987) have shown that cluster analysis can be used as a method for the evaluation of genetic similarity in specific host–parasite interactions and show that it offers the possibility of predicting effective sources of resistance in the *Lactuca sativa* – *Bremia lactucae* system. Derivations of the method have also been used extensively in other types of analyses, e.g. for phylogenetic or molecular tree constructions (Tateno and Tajima 1986; Rohlf and Wooten 1988).

A difficulty in using this statistic is in the choice of the algorithm. Both Sokal (1986) and Rohlf and Wooten (1988) agree that the UPGMA clustering method (or "group-average" or "average-linkage" cluster analysis) generally yields results which are the most accurate for classification purposes, although Lebeda and Jendrulek (1987) found in their analysis (which was based on qualitative data alone) that, for six of the most commonly used procedures, results were nearly identical. Since accuracy generally increases with number of characters

treated (Rohlf and Wooten 1988) and some types of data are known to have more weight in a multivariate analysis than others (Thorpe 1985), the choice of characters represents, in addition to the algorithm, another of the variables affecting the outcome.

Hierarchical cluster analysis can readily be used to assess relatedness and distance of any type of samples characterized by any type of descriptors. Therefore, it may be used to assess genetic similarity and dissimilarity in germplasm collections, and the technique could also have applications for the selection of parental lines for which varying degrees of segregation are sought. Studies have demonstrated the potential that the knowledge of "distance" can have in relation to crossing. For example, Lefort-Buson et al. (1987) have shown that F_1 hybrids between European and Asiatic lines in *Brassica napus* L. were far superior for yield than all other tested varieties, selfed lines, and European or Asiatic F_1 hybrids crossed in all other combinations. While the potential of using other multivariate statistics such as the Mahalanobis' D^2 distance for parental selection has already been assessed (Bhatt 1970; Ariyo 1987) and analyses of data in germplasm collections based in part on hierarchical cluster analysis have already been made (Spagnoletti Zeuli and Qualset 1987; Veronesi and Falcinelli 1988), further work appears required to determine how similarity and dissimilarity in collections can be estimated.

The purpose of this study is two-fold. First, its aim is to establish the precision with which hierarchical cluster analysis can define degrees of relatedness in gene bank samples and what the best basis is to define commonness. Second, an assessment is made of the potential the similarity matrix offers to predict the degree of segregation of given samples.

Materials and methods

Calculation methods

Only the agglomerative clustering procedures of one main statistical language were investigated. The calculations were all carried out using GENSTAT IV on the Cambridge University 3084 Q IBM mainframe and variate types selected for the calculation of the similarity matrix were of the quantitative "city-block" type (Genstat 1980). For qualitative traits, if two individual ratings were different, i.e. if $x(i) \neq x(j)$, then scores of 0 were given or of 1 if they were equal. For quantitative traits, a score of $1 - \frac{|(x(i) - x(j))|}{Range}$ was assigned.

Two individual cluster analyses were performed on randomly selected groups of entries from a large data set. Both a small and a large matrix were selected. The data were that of the barley germplasm collection of the AFRC-IPSR (formerly the Plant Breeding Institute or PBI), which were composed of both qualitative and quantitative descriptors, including disease resistances. The two individual matrices were selected by means of an ALGOL program which pooled entries randomly. This was part of a much larger analysis on the same collection which is described elsewhere (Peeters 1988). The algorithms corresponding to the five clustering procedures which are available in GENSTAT IV and representing five of the most commonly used procedures are the following:

(1) Single-linkage cluster analysis, calculated according to the method of Gower and Ross (1969), where the similarity between two clusters is the greatest similarity between any two units, one in each cluster or:

$$S'(i, k) = max(S(i, k), S(j, k))$$

(2) Furthest-neighbour cluster analysis, calculated as:

$$S'(i, k) = min(S(i, k), S(j, k))$$

(3) Centroid cluster analysis, calculated as:

$$S'(i, k) = \frac{(n(i)S(i, k) + n(j)S(j, k))}{(n(i) + n(j))}$$

(4) Average-linkage cluster analysis, calculated as:

$$S'(i, k) = \frac{(S(i, k) + S(j, k))}{2}$$

(5) Median cluster analysis, calculated as:

$$S'(i, k) = \frac{(S(i, k) + S(j, k))}{\frac{2 + (1 - S(i, j))}{4}}$$

where

S = Similarity
$S'(i, k)$ = Similarity between the new group i and any group k, given the merger (i, j)
$n(i)$ and $n(j)$ = Number of units in groups i and j.

Experimental procedure

The matrix selected to extract parental lines was the larger matrix, which contained 247 accessions originating from 52 countries. The similarity matrix was searched for the 5 most highly related pairs of entries and the 5 least related pairs. Five pairs were also selected at random. The accessions were obtained from the curator of the AFRC collection and each gene bank entry from each of the pairs was grown in a glasshouse in groups of plants sown at regular intervals to ensure matching of flowering. The pairs of selected entries were then crossed. Seeds from each parental plant were saved and the F_1 plants were grown in isolation. The selected parental lines and their F_2 progenies were then tested for their variability for one important trait in barley breeding, namely resistance to powdery mildew (*Erysiphe graminis* DC f.sp. *hordei* Em. Marchal). The procedure used for this was the following.

The comparison of F_2 progenies was limited to three of the crosses. These represented a pair of most related entries, least related, and randomly selected. These pairs of accessions, together with their coefficient of similarity, their origin, and their accession number are shown in Table 1. The first successful cross for the most related entries involved two breeders' lines from Denmark (Abed 0166 and Abed 0-44). This material had entered the germplasm collection in 1962. This particular cross was labelled cross No. 48 and the similarity of the parents, as determined by the similarity matrix, was 0.98.

The first successful cross representing the least related pair of entries (cross No. 43) involved one parent from the UK and one from Tibet. The material from the UK was another breeders' line from the PBI, also from the 1960s [Hybrid HB 40/66/5 (PBI)]. The accession from Tibet was a landrace from

Table 1. Selected parental lines with their coefficients of similarity

Cross	Similarity	Countries[a]	Accessions[b]
48	0.98	DNK × DNK	03405 × 03406
43	0.51	TIB × UK	04191 × 03624
52	0.72	BGR × AFG	03679 × 04021

[a] Country codes: AFG – Afghanistan, BGR – Bulgaria, DNK – Denmark, TIB – Tibet, UK – UK
[b] Accession numbers from the AFRC-IPSR germplasm collection

Khangma, which originated form the Kew gene bank and was incorporated in the collection in 1948. The similarity of these parents was 0.51 and was the lowest of all the combinations.

Finally, for the randomly selected pair of entries, the cross (No. 52) involved material from Bulgaria (status and year of accession unrecorded), and a landrace from Afghanistan from Obeh which entered the collection in 1966. These parents had a similarity of 0.72.

Hence, a gradient of similarity of $0.98 > 0.72 > 0.51$ was expected to yield a corresponding gradient of increasingly greater segregation in the character selected for the estimate of the variability present in the germplasm.

In addition to this material, a recently collected landrace from an IBPGR collecting mission to Morocco (IBPGR 1985) was also taken and progenies from this material were screened. The collector's number for this accession, which is now kept in the Nordic Gene Bank, is 1142. Single seed descent lines were derived from this sample, and two third generation lines were selected and called NG 1142 B and NG 1142 D. This material was evaluated in another trial together with other gene bank entries in the field and had shown comparatively very little variability and no evidence of segregation (Peeters 1988). Since the material tested originated from a single collector's sample which was not segregating and the cross was made at the third generation of single seed descent from this sample, this cross was expected to represent an example of highly related germplasm from a collection.

The material was screened in the Department of Cereal Research at the Institute for Plant Science Research, and 60 F_2 plants were screened for each cross and for each isolate. Seedlings were germinated in a spore-proof glasshouse. The central part of the first leaf of each F_2 plant was cut into four equal segments of 1.5 cm. Each of the four segments were distributed randomly under spore-proof conditions into four boxes on a water agar medium containing 100 ppm benzimidazole with their abaxial sides towards the agar. These four boxes were used for the first four isolates or group of isolates which were tested. The procedure was then repeated for the second group of isolates and then for each of the four crosses. Hence, a total of $60 \times 8 \times 4$ or 1920 F_2 samples were screened.

The first group of isolates tested for each cross consisted of an old isolate which is no longer present in the field in the UK, one of the isolates currently prevalent in the field, and one isolate collected in the field in Israel. The fourth test measured the reaction to the mixture of the three isolates. The second series of four tests reproduced the first one with the exception of the isolates, which were all different.

All isolates were taken from the mildew collection at the AFRC. The first group were isolates cc1, cc64, and cc115. The second group were isolates cc13, cc135, and cc113. Isolates cc1 and cc13 are the old isolates no longer present in the field in the UK, cc64 and cc135 are the current isolates, and cc115 and cc113 are the isolates from Israel.

Inoculation was done in a sterilized settling tower of large capacity. Previous statistical tests showed that spore distribution in these conditions is completely homogeneous (J. A. Martinelli, unpublished results). One variable which could not be controlled with precision in the test was the volume of spores/inoculation, but since the tower was large enough to fit four boxes of samples, each of the individual crosses tested received the same amount of inoculum.

Following inoculation, the samples were placed in a growth chamber at 16 °C with 16 h light and were scored 10 days later. The scoring system used was based on that of Moseman et al. (1965) where 0 = Immune, 1 = Highly resistant, 2 = Moderately resistant, 3 = Moderately susceptible and 4 = Highly susceptible. In addition, ratings for each class were complemented by an estimate of number of colonies on a scale where 0 = 0% – 5% of leaf area infected, 2 = 6% – 10%, 4 = 11% – 20%, 6 = 21% – 40%, and 8 = 41% – 70%. This second scaling added an estimation of the variability present within a given class.

Results

Results showing the effects of the individual clustering procedures on the smaller data matrix are reproduced in Fig. 1. The hypothesis tested was that commonness was greater between entries within countries than between countries and that the statistic could be used to show both the inter-country and between-country relatedness of the entries. The results show "chaining" for both the single linkage and median cluster methods, i.e. poor resolution of individual groups, despite the small size of the matrix. Furthest neighbour, however, resolves the groups extremely well but the expected relationship between groups of entries by country gene pool is not confirmed. The centroid and average linkage methods, on the other hand, classify entries correctly by origin and the expected distance between countries appears correct as well. The entries from the two European countries, representing between-country groups, were expected to be closely related and this is confirmed. The third most related country to this group was expected to be the USA having, as for the European countries, a high probability of being hybrid in origin as well as being bred. This is confirmed for the average linkage or (UPGMA) method but not for the centroid method. However, in 27 other cluster analyses which were done on the data from this collection and including the analysis of the matrix containing the parental lines, both methods were found to be almost comparable and classified entries with a high degree of accuracy according to their origin (Peeters 1988).

The results of the mildew trial for the parents are shown in Table 2 and show that the parents were generally susceptible to most of the races and their mixtures. The parental lines with the highest degree of overall resistance were those from cross 48.

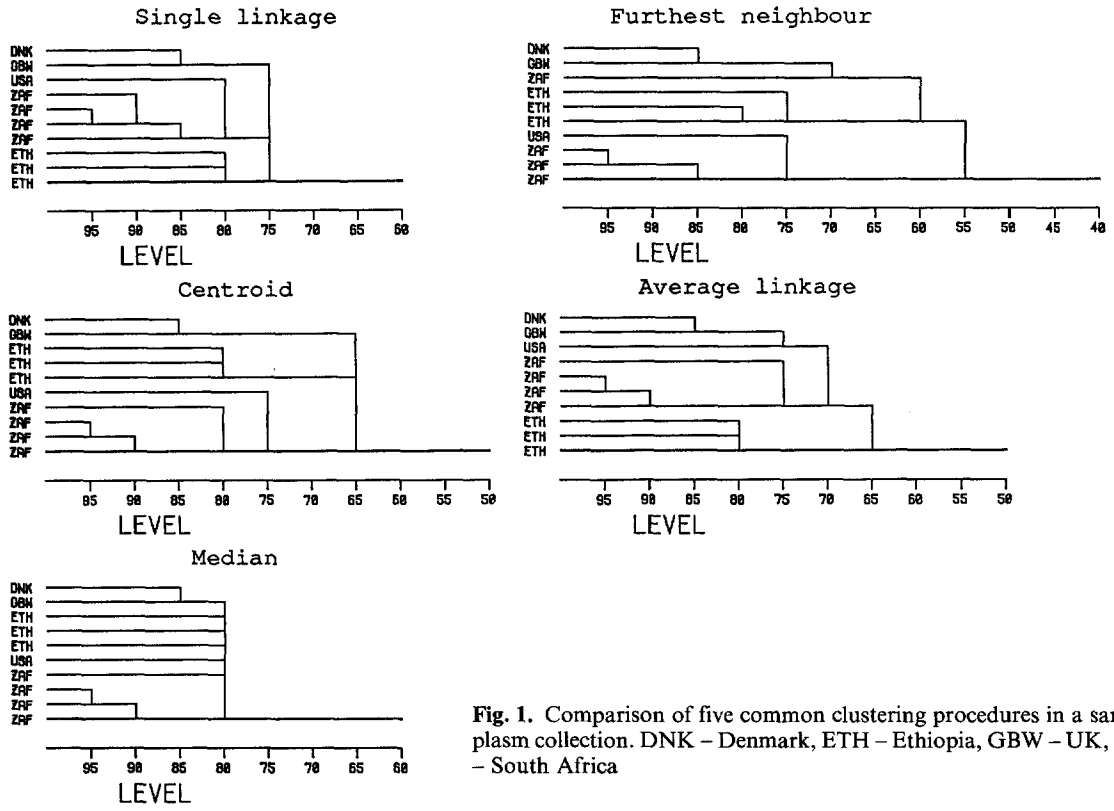

Fig. 1. Comparison of five common clustering procedures in a sample from a germplasm collection. DNK – Denmark, ETH – Ethiopia, GBW – UK, USA – USA, ZAF – South Africa

Table 2. Reactions of the parental lines to each of the powdery mildew isolates and their mixtures. Reaction types: 0 – Immune, 4 – Highly susceptible

Cross	Parent	Isolates[a]							
		cc1	cc64	cc115	Mix1	cc13	cc135	cc113	Mix2
46	NG 1142 B	4	4	4	4	4	4	4	4
	NG 1142 D	4	4	4	4	4	4	4	4
48	03405	2	4	2	3	2	2	3	3
	03406	2	4	4	4	2	2	4	3
52	03679	4	4	2	3	3	4	3	4
	04021	4	4	4	4	4	4	4	4
43	04191	2	4	4	3	4	4	4	4
	03624	4	4	4	4	4	4	4	4

[a] From the AFRC-IPSR mildew collection

The results for the F_2 progenies were summarized as frequency distributions of the reactions of the progenies to each isolate and for each cross. These are reproduced in Fig. 2. The standard deviations of the reactions to each isolate and the mean standard deviation for each cross are shown in Table 3. The mean standard deviations arranged in descending order for each cross show that cross 43 generated the most segregation and was followed by crosses 52, 46 and 48.

The more distant cross, for which both parents were generally susceptible (Table 2), nevertheless gave the best segregation. On the other hand the most related cross, as predicted from the similarity matrix, and which also had the highest proportion of resistance in the parents, gave the worst segregation overall among the three crosses selected on the basis of the multivariate statistic. The most distant cross (43) also gave by far the best segregation in comparison with all the other observed segre-

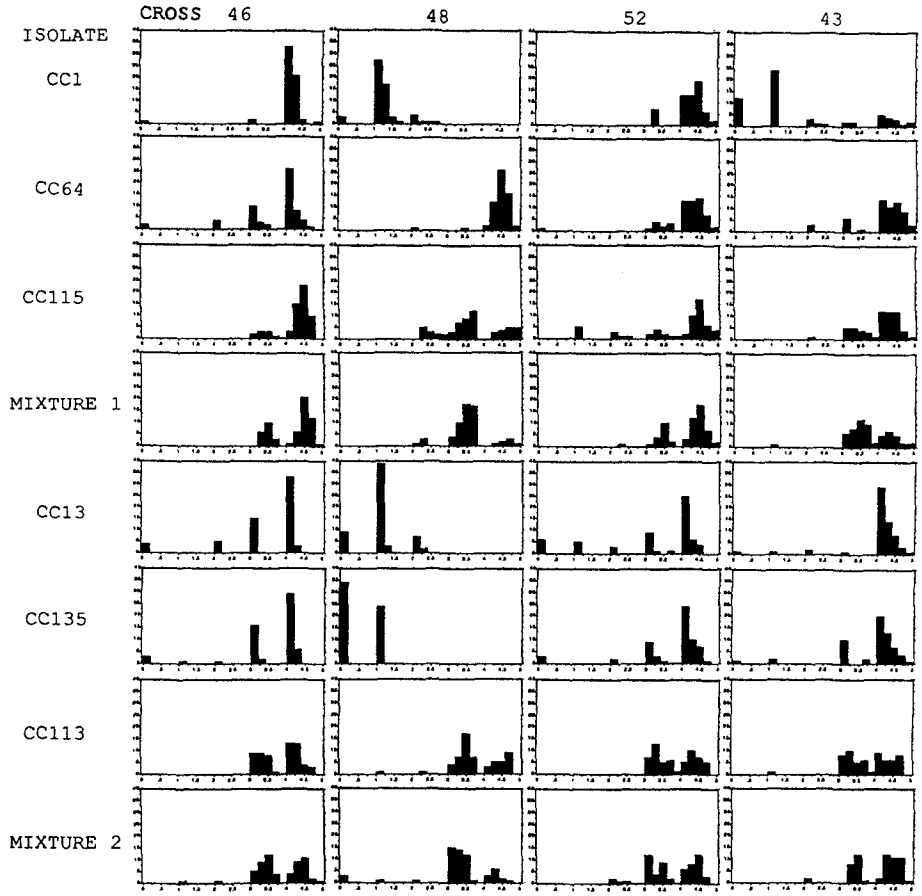

Fig. 2. Segregation patterns for powdery mildew resistance in F_2 progenies of selected germplasm samples. Sixty observations per distribution. Vertical scale from 0 to 40. Horizontal scale from 0 to 4.8

Table 3. Standard deviations and overall means of standard deviations of the reactions to the isolates and their mixtures in the F_2 progenies[a]

Cross	Isolated[b]								Means
	cc1	cc64	cc115	Mix1	cc13	cc135	cc113	Mix2	
F46	0.58	0.93	0.43	0.52	1.10	1.02	0.52	0.68	0.72250
F48	0.47	0.38	0.78	0.53	0.57	0.50	0.67	0.94	0.60500
F52	0.41	0.68	0.11	0.55	1.41	1.01	0.55	0.69	0.80000
F43	1.62	0.66	0.56	0.61	0.80	0.88	0.66	0.64	0.80375

[a] Ratings from 0 to 4.8
[b] From the AFRC-IPSR mildew collection

gations for one isolate (cc1). This particular F_2 segregation is reproduced in Fig. 3, which shows the range of variability recovered from that cross. The second best segregation was observed for the second most distant cross (52) in reaction to isolate cc13. Finally the most distant cross also gave the only observed resistant progeny to the first mixture of isolates.

A number of other points can be noted on the basis of these results. The most related cross (48), which was between two old breeders' lines, showed little variation but it was the only cross for which the progenies showed a high level of resistance to some isolates. It is interesting to note that this resistance was fully present in the parents for the old isolates. Resistance, on the other hand, was completely lost for one of the new isolates, namely cc64, but was fully retained for cc135 which is also one of the new isolates. This confirms the gene-for-gene type of resistance. However, the patterns of variation in this material for isolate cc113, of Israeli origin, are more complex and suggest either residual variation or a multigene type of resistance, which is also known to occur in barley for powdery mildew (Asher and Thomas 1987).

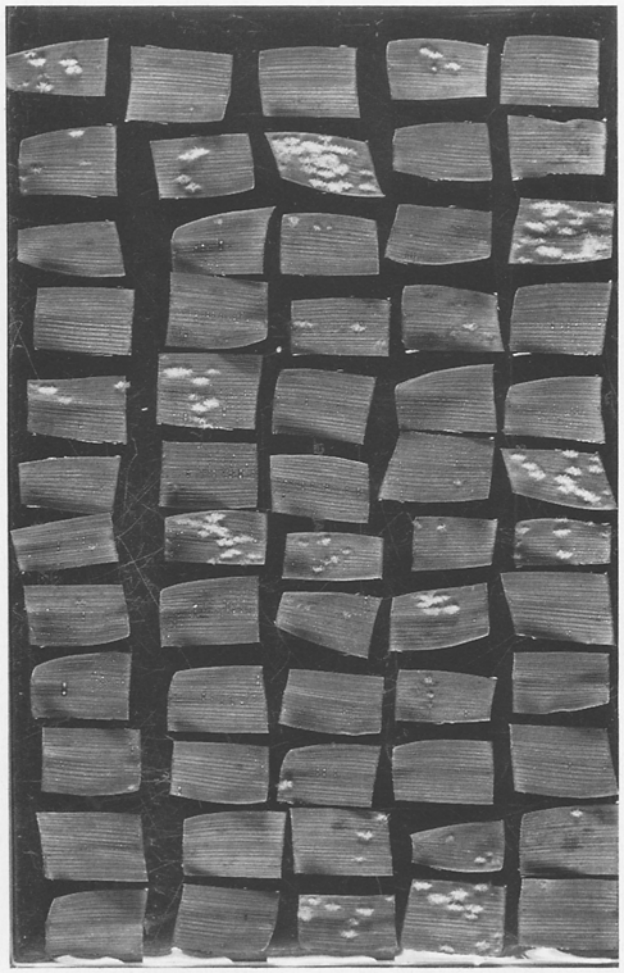

Fig. 3. Segregation pattern for resistance to a powdery mildew isolate in the F_2 progeny of two accessions selected for their dissimilarity (isolate cc1, cross 43)

The relatively high degree of variability which exists in highly related material is shown by the segregation patterns for cross 46. This cross involved recently collected material from Morocco and was performed at the third generation of single seed descent from the original sample, which did not show any sign of segregation. Uniformity was confirmed by the reactions of the parents to the different isolates (Table 2). The fact that variation was observed in the progeny of a cross between two highly related germplasm samples is particularly interesting. At least two mechanisms could explain this variability, which was either residual or may have been the result of crossing two lines which had a high degree of relatedness. It has been shown recently that variability can appear de novo in conditions of high inbreeding (Biémont et al. 1987). While the comparison of this particular cross with cross 48 is somewhat difficult to make, more light is shed on the question of variability in advanced material by the practical experience of breeders, which suggests that variability in highly bred material is very common (Duvick 1984; Peeters and Galwey 1988).

This experiment shows that resistance per se to powdery mildew is easy to obtain, at least to specific races and whatever the origin of the germplasm. For mixtures of races, patterns of resistance appear more complex than those obtained for the single isolates (Fig. 2), although a trend emerged showing that, with increased resistance to single isolates, increased resistance is also present to the mixture.

Discussion

Results from this experiment suggest that hierarchical cluster analysis could become a useful tool for the management of the variation in germplasm collections. Not only can it be used to assign entries to their respective gene pools, even when origin is unknown, but there is also clearly the potential to use this technique to estimate the segregation potential of germplasm in crosses.

The high degree of accuracy of the multivariate analysis in this study is believed to have been positively influenced by the fact that, in this collection, each entry was described by at least 20 characters and that each observation was replicated over 3 years. However, the fact that classification was highly accurate by origin confirms that origin is a simple way to partition variation in germplasm collections and that more relatedness between entries exists within a given country gene pool than between, even for two neighbouring countries. This concept is examined in detail elsewhere (Peeters 1988).

These findings offer a positive outlook to devise a global strategy for the management and more effective use of variation in germplasm collections. Particularly they offer firm support to the concept building "core" collections of variability (Frankel 1984) is not only desirable but feasible.

The results of the crossings also offer some clues on the nature of the underlying variability in the material which is kept in collections. Useful segregates were readily detected in these crosses, even in highly related material. This suggests that an over-emphasis on sampling rarity may not always necessarily be warranted, particularly when the size of the sample makes the genuinely rare accession increasingly at risk to go undetected, unused, and eventually lost (Goodman 1984). In the recent literature on strategies for collecting germplasm, there has been some debate about the need for sampling down to the "rare allele" level with perhaps too little appreciation of the level of variability normally present in samples. While it is recognized that genuinely rare samples exist and do need to be conserved, it must also be recognized that the natural variability of germplasm samples is often very high and that crossing offers a simple way to recover

large numbers of new combinations. Therefore, it is suggested that conservation efforts would benefit greatly by becoming more closely linked with programs aimed at studying and exploiting the true potential of the samples at the breeders' level.

In order to manage collections more effectively, ways to partition their variability simply, such as by country of origin, appear increasingly appealing. By using hierarchical cluster analysis, it was shown here not only that there is a basis for such groupings but that a relatively simple tool to do them is available as well. These results were confirmed by the crossings. Indeed crosses of material between country gene pools were found to yield better segregation than by using entries from within gene pools, and the more distant these pools were the greater the segregation.

With the large increases of processing power of microcomputers and corresponding software developments, classification tools such as hierarchical cluster analysis are becoming increasingly accessible (Fink 1986). Therefore, grounds are believed to exist at present both for curators and breeders to further enhance the usefulness of their collections.

Acknowledgements. R. Giles, curator of the AFRC, is thanked for his help and cooperation throughout this experiment. A. Ubaydli is thanked for his encouragements. This work was supported in part by a scholarship from the American Friends of Cambridge and by an O.R.S. Award from the Committee of Vice-Chancellors and Principals of the Universities of the United Kingdom to J.P.P. and by a scholarship from the Brazilian Council for Scientific and Technological Development (CNPq) to J.A.M.

References

Ariyo OJ (1987) Multivariate analysis and the choice of parents for hybridization in Okra (*Abelmoschus esculentus* (L.) Moench). Theor Appl Genet 74:361–363

Asher MJC, Thomas CE (1987) The inheritance of mechanisms of partial resistance to *Erysiphe graminis* in spring barley. Plant Pathol 36:66–72

Bhatt GM (1970) Multivariate analysis approach to selection of parents for hybridization aiming at yield improvement in self-pollinated crops. Aust J Agric Res 21:1–7

Biémont C, Aouar A, Arnault C (1987) Genome reshuffling of the *copia* element in an inbred line of *Drosophila melanogaster*. Nature 329:742–744

Duvick DN (1984) Genetic diversity in major farm crops on the farm and in reserve. Econ Bot 38:161–178

Fink WL (1986) Microcomputers and phylogenetic analysis. Science 234:1135–1139

Frankel OH (1984) Genetic perspectives of germplasm conservation. In: Arber W, Illmensee K, Peacock WJ, Starlinger P (eds) Genetic manipulation: impact on man and society. Cambridge University Press, Cambridge, pp 167–170

Genstat (1980) Reference manual 4.03. Lawes Agricultural Trust, Rothamsted Experimental Station, Numerical Algorithms Group, UK

Goodman MM (1984) An evaluation and critique of current germplasm programs. In: Report of the 1983 plant breeding research forum. Aug. 9–11, 1983. Conservation and utilization of exotic germplasm to improve varieties. Pioneer Hi-Bred International, Des Moines/IA, pp 195–227

Gower JC, Ross GJS (1969) Minimum spanning trees and single linkage cluster analysis. Appl Stat 18:54–64

IBPGR (1985) International Board for Plant Genetic Resources. Annual Report 1984. IBPGR Secretariat, Rome

Lebeda A, Jendrulek T (1987) Cluster analysis as a method for evaluation of genetic similarity in specific host–parasite interaction (*Lactuca sativa – Bremia lactucae*). Theor Appl Genet 75:194–199

Lefort-Buson M, Guillot-Lemoine B, Dattee Y (1987) Heterosis and genetic distance in rapeseed (*Brassica napus* L.): crosses between European and Asiatic selfed lines. Genome 29:413–418

Lin CS, Binns MR, Lefkovitch LP (1986) Stability analysis: Where do we stand? Crop Sci 26:894–900

Moseman JG, Scharen AL, Greeley LW (1965) Propagation of *Erysiphe graminis* f. sp. *tritici* on barley and *Erysiphe graminis* f. sp. *hordei* on wheat. Phytopathology 55:92–96

Murphy JP, Cox TS, Rodgers DM (1986) Cluster analysis of red winter wheat cultivars based upon coefficients of parentage. Crop Sci 26:672–676

Peeters JP (1988) Crop variability and its exploitation through germplasm collections: an assessment based on barley. PhD Thesis, Cambridge University

Peeters JP, Galwey NW (1988) Germplasm collections and breeding needs in Europe. Econ Bot (in press)

Rohlf FJ, Wooten MC (1988) Evaluation of the restricted maximum-likelihood method for estimating phylogenetic trees using simulated allele-frequency data. Evolution 42:581–595

Sokal RR (1986) Phenetic taxonomy: theory and methods. Annu Rev Ecol Sys 17:423–442

Spagnoletti Zeuli PL, Qualset CO (1987) Geographic diversity for quantitative spike characters in a world collection of durum wheat. Crop Sci 27:235–241

Tateno Y, Tajima F (1986) Statistical properties of molecular tree construction methods under the neutral mutation model. J Mol Evol 23:354–361

Thorpe RS (1985) The effect of insignificant characters on the multivariate analysis of simple patterns of geographic variation. Biol J Linn Soc 26:215–223

Veronesi F, Falcinelli M (1988) Evaluation of an Italian germplasm collection of *Festuca arundinacea* Schreb. through a multivariate analysis. Euphytica 38:211–220