Comparison of Different Methods for Classification of Gene Bank Accessions

Zrinka KNEZOVIĆ [1]
Jerko GUNJAČA [2*]
Zlatko ŠATOVIĆ [2]
Ivan KOLAK [2]

SUMMARY

The objective of this study was to group common bean (*Phaseolus vulgaris* L.) accessions originating from different collection sites according to their morphological similarity. Classification based on genetic diversity should improve the maintenance of the collection of bean accessions and enhance its use as a valuable source of desirable traits in plant breeding. Materials used in this study are accessions of common bean, collected at various locations in Croatia, and Bosnia and Herzegovina, and stored in the Croatian Bank of Plant Genes (HBBG). In order to use a number of qualitative and quantitative morphological traits, scored according to IPGRI (International Plant Genetic Resources Institute, Rome), and HBBG descriptor lists, we followed the method of classification described by Franco and Crossa (2003). Results obtained from this study reveal some weaknesses as well as advantages of methods used for classification of gene bank accessions.

KEY WORDS

distance measure, cluster analysis, genetic diversity, modified location model

[1] University of Mostar, Faculty of Agriculture
Kralja Zvonimira 14, BIH-88000 Mostar, BIH
[2] University of Zagreb, Faculty of Agriculture
Svetošimunska 25, HR-10000 Zagreb, Croatia
* E-mail: jgunjaca@agr.hr

Received: May 30, 2005

INTRODUCTION

With the constant income of new accessions, germplasm collections of plant gene banks can substantially increase in their size, which can decrease the efficiency of maintenance and utilization. A solution to this problem can be obtained by using proper statistical methods for classification and categorization of collected accessions. The Croatian Bank of Plant Genes (HBBG) was founded in 1991 as a project initiated by the Ministry of Science and Technology in co-ordination with the Ministry of Agriculture and Forestry of the Republic of Croatia. Among others, HBBG stores the collection of common bean (*Phaseolus vulgaris* L.) accessions, gathered at several collecting missions (Orlandini et al., 1996). On several previous occasions, we have used this common bean collection in search of the adequate and most efficient statistical methods for the classification of gene bank accessions (Gunjača et al., 1998; Gunjača et al., 2000).

Common bean accessions have been scored for a number of morphological traits. Scores were used to calculate a matrix of dissimilarities between accessions, which is then subjected to cluster analysis in order to classify accessions. Integrating information from both quantitative and qualitative traits (hence continuous and categorical variables) requires the use of generalized type of distance proposed by Gower (1971) for the dissimilarity matrix calculation. Basically, this distance is the average distance across all variables, and it can take any value between 0 and 1. For the quantitative traits, this imposes the range standardization. Meanwhile, in a number of studies, the most appealing approach for the treatment of qualitative traits was to convert them to binary variables. Using this approach, each qualitative trait would be transformed into binary variables whose number is equal to the number of different states recorded for that variable. This approach is appealing, because the distance between two accessions will immediately be 0 or 1; but also questionable, because variables will be unequally weighted. There are two aspects to this problem. In a multistate situation, more weight will be given to the variables with more states, while in a single state situation mismatches will have (twice) more weight than matches do. Cole-Rodgers et al. (1997) proposed a solution for the first aspect, which employs the division by the number of states. Regarding the second aspect, Peeters and Martinelli (1989) proposed to preserve the qualitative variables in their original state and treat them as asymmetric nominal variables when calculating Gower's distance. In order to estimate the efficiency of these two proposed modifications, in our previous study (Gunjača et al., 2000) we have compared them to a straightforward approach without any modification. Although all three approaches yielded optimal five-cluster solution after using Ward's cluster algorithm, there was a notable discrepancy in the size and shape of the clusters. Nevertheless, evaluation of the results using criteria proposed by Franco et al. (1997) showed that all methods have similar efficiency, on the basis of number of influential variables criteria. In their subsequent work, Franco et al. (1998) developed a nonhierarchical clustering method for classification using both continuous and categorical variables, called the Modified Location model (MLM). Using the sequential "Ward after Gower" – MLM clustering strategy, they concluded that posterior use of MLM can improve the composition of the clusters obtained by Ward's method and produce compact and well-separated groups. Detailed review of these methods and criteria is given in Crossa and Franco (2004).

The aim of this study is to investigate possible improvement of previously obtained classifications (Gunjača et al., 2000) of common bean accessions, by subsequent use of the MLM after three original clustering methods. Furthermore, efficiency of different methods will be further evaluated by comparison of their improved cluster solutions.

MATERIALS AND METHODS

The original data set was comprised of 123 accessions of common bean (*Phaseolus vulgaris* L.) landraces collected at various locations in Croatia, and Bosnia and Herzegovina, scored for 9 continuous and 22 categorical traits, according to the IPGRI (International Plant Genetic Resources Institute) and HBBG descriptor list (Henneberg, 1990). Due to certain limitations of MLM, we had to restrict this original data set to include only those traits and accessions which did not have any missing data or multistate records. The optimal solution to this problem, with a minimum loss of information, was to exclude 4 accessions and 5 categorical traits. Therefore, ten plants per plot of each of 119 accessions used in the present study were scored for 9 continuous (plant height, leaflet length, leaflet width, number of flower buds per inflorescence, stem thickness, plant height at maturity, ordinal number of lowest flower-bearing node, number of nodes on main stem at maturity, number of primary branches per plant at maturity) and 17 categorical traits (Table 1).

For the classification of accessions we used a two-stage clustering strategy proposed by Franco et al. (1998). The first stage includes a hierarchical clustering method, in order to obtain clusters, which are then used in the second stage as initial groups for the MLM. In order to comply with the aim of this study, we used the initial groups obtained from the previous study (Gunjača et al., 2000), although it caused a slight discrepancy. Data sets used in present and previous studies differed because of the exclusion of 4 accessions, due to the missing data occurrence. However, it should not be unreasonable to assume that excluding 4 out of 123 accessions will

Table 1. Qualitative traits and their states

Trait	States
Type of growth	(1) Determinate; (2) Indeterminate
Anthocyanin in stem	(1) Present; (2) Absent
Angle between leaf petiole and stem	(1) < 25°; (2) 25-45°; (3) > 45°
Branching pattern	(1) From basal nodes; (2) From higher nodes
Flower ground colour	(1) White; (2) Pink; (3) Violet; (4) Red
Increased colour intensity on the tip of the banner	(1) Present; (2) Absent
Colour of pattern on pod	(1) Present; (2) Absent
Pod ground colour	(1) Green; (2) Pale green; (3) Yellow
Pod shape	(1) Straight; (2) Curved
Pod cross-section	(1) Eliptic; (2) Round; (3) Narrow, oblong
Threads in pod sutures	(1) Present; (2) Absent
Mature pod ground colour	(1) White; (2) Straw-coloured; (3) Sulphureous; (4) Buff; (5) Variegated
Middle leaflet shape	(1) Rhomboidal-lengthened; (2) Rhomboidal-oval; (3) Heart-shaped; (4) Triangular-oval; (5) Triangular-lengthened
Leaf tip shape	(1) Obtuse; (2) Acute; (3) Lengthened
Anthocyanin in leaf	(1) Present; (2) Absent
Node production on main stem after floweringb	(1) Type I; (2) Type II; (3) Type III; (4) Type IV
Branching habit	(1) Sparsely branched; (2) Densely branched

not produce a substantial difference in classification, and furthermore, during the optimization procedure in the second stage of analysis these initial clusters will be reshuffled, anyway. Therefore, three initial groupings (with five clusters each) were those obtained using methods of "straightforward" (G), Peeters and Martinelli (PM), and Cole-Rodgers (CR) method for Gower's distance calculation (Gower, 1971; Peeters and Martinelli, 1989; Cole-Rodgers et al., 1997).

The second stage feature, MLM, operates by transforming all categorical variables into a unique multinomial variable, thus forming a vector of continuous variables, which also includes this single categorical variable (values of which correspond to the observed combinations of categorical variables). Assuming the homogeneity of variance-covariance matrices within subpopulations, this model uses the likelihood function of the whole sample data as an objective function for the maximization process. Estimation of the parameters and calculation of the probability of belonging to a subpopulation is done using the expectation-maximization (EM) algorithm.

Statistical analysis is done by using the SAS software (SAS Institute Inc., 1989), and applying the programming code given by Franco and Crossa (2003). This code also includes canonical discriminant analysis, for the purpose of evaluation of classification based on the continuous variables.

RESULTS

Homogenization of the initial groups by MLM required 96 iterations for CR method, 102 iterations for PM method, and 199 iterations for G method, until EM algorithm reached convergence. Furthermore, number of accessions transferred to different clusters from their initial groupings is 35 for PM, 43 for G, and 44 for CR. There is only one observation classified with less than 75% probability, detected in CR solution.

Correspondence between initial and improved groups is shown in Tables 2, 3 and 4. What can be generally observed for all three methods is slight deviation from the initially almost clear distinction between accessions with determinate (clusters 3, 4, and 5) and indeterminate types of growth (clusters 1 and 2), as well as the dissipation of the initial group 4, which fails to retain more than 50% observations in any of the final groups.

There are also several specific features for each of the methods. The initial grouping by G and PM puts one determinate type in cluster 1, and one indeterminate type in cluster 4. Improved solutions included 4 and 5 observations moved from cluster 1 to clusters with predominantly determinate types. In the same time, initial solution for CR included 4 indeterminate types in cluster 4, while final solution moved only one accession from cluster 1 to cluster 3. PM achieved the best overall agreement between initial and final solution, while CR, besides already described heterogeneity of initial cluster 4, suffers from the similar problem in cluster 3, with half of its original members reassigned to cluster 5.

First canonical variable explained 91% and 98% of the variability between G and PM groups, respectively, but only 66% of the variability between CR groups. However, inclusion of the second canonical variable raises this value up to 92%, thus making graphical representation of first two canonical variables a useful tool for the visualization of the relationships between accessions within and between the groups. Plots of the first two canonical variables are shown on Figures 1-3.

Table 2. Frequencies (upper) and percentages (lower values) of the row totals for the observations in the initial (G) and final (MLM) groups

Initial group	Final group					Total
	1	2	3	4	5	
1	31	1	1	3		36
	86.11	2.78	2.78	8.33		
2	2	11				13
	15.38	84.62				
3			12	7	5	24
			50.00	29.17	20.83	
4			11	10	5	26
			42.31	38.46	19.23	
5			1	7	12	20
			5.00	35.00	60.00	
Total	33	12	25	27	22	119

Table 3. Frequencies (upper) and percentages (lower values) of the row totals for the observations in the initial (PM) and final (MLM) groups

Initial group	Final group					Total
	1	2	3	4	5	
1	27	5		5		37
	72.97	13.51		13.51		
2	1	11				12
	8.33	91.67				
3			29	4	6	39
			74.36	10.26	15.38	
4			6	6	2	14
			42.86	42.86	14.29	
5			2	4	11	17
			11.76	23.53	64.71	
Total	26	16	37	19	19	119

Table 4. Frequencies (upper) and percentages (lower values) of the row totals for the observations in the initial (CR) and final (MLM) groups

Initial group	Final group					Total
	1	2	3	4	5	
1	22	8	1			31
	70.97	25.81	3.23			
2	3	11				14
	21.43	78.57				
3			16		16	32
			50.00		50.00	
4			3	9	7	19
			15.79	47.37	36.84	
5			4	2	17	23
			17.39	8.70	73.91	
Total	25	19	24	11	40	119

G plot (Figure 1.) reveals clear separation between determinate (clusters 3-5, negative values) and indeterminate types (clusters 1-2, positive values) along the first canonical variable. Hence, it is not surprising that first canonical variable is highly correlated (r = 0.73) to the plant height. Furthermore, it can also be noted that indeterminate types are divided into cluster 2, with tall accessions, and cluster 1, with relatively shorter accessions. Second canonical variable enables the distinction between cluster 3, with positive values, and cluster 4, with predominantly negative values for this variable, which shows the highest correlation with leaflet length (-0.68) and width (-0.55). Finally, although it is possible to observe more or less clear distinction between cluster 5 and cluster 3 and 4, it is not easy to find a straightforward explanation for it.

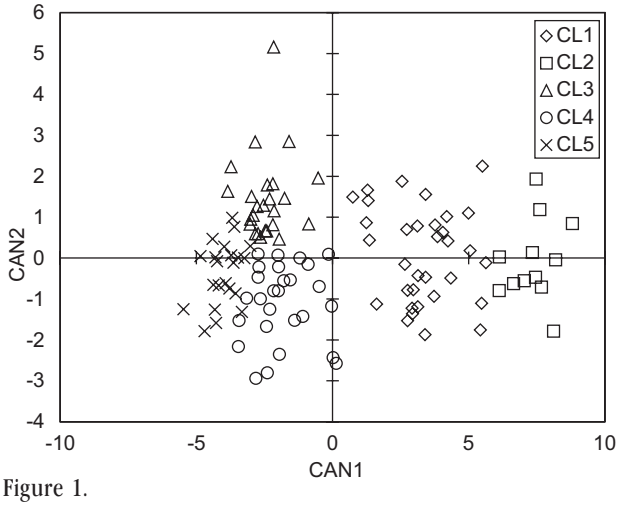

Figure 1.
Plot of the first two canonical variables for the G groups

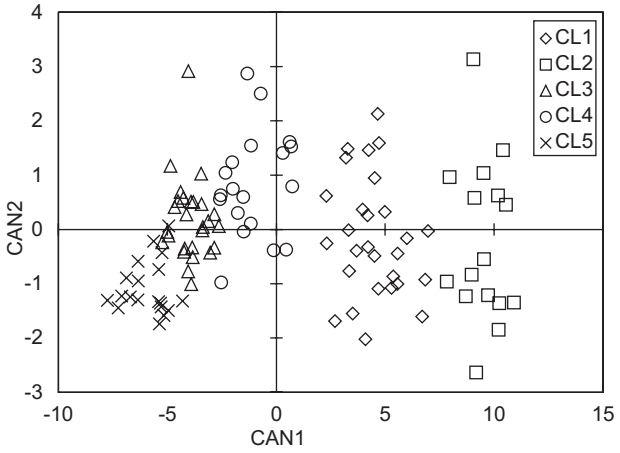

Figure 2.
Plot of the first two canonical variables for the PM groups

PM plot (Figure 2.) again shows the dominant role of the plant height in the grouping of the accessions, this time highly correlated with both second (-0.57) and first (0.55) canonical variable. All other conclusions are similar to those for G groupings, except for the low influence of leaflet shape on the classification by this method. Slightly different conclusions can be drawn from the inspection of CR plot (Figure 3.). Although it follows the general rule for the distinction between determinate and indeterminate types, with

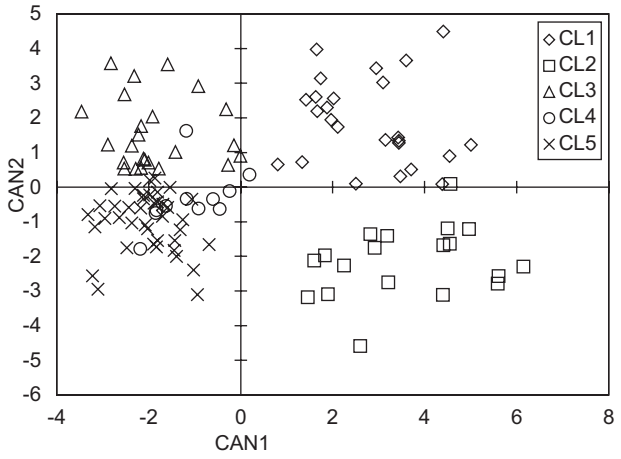

Figure 3.
Plot of the first two canonical variables for the CR groups

highest correlation between first canonical variable and plant height among methods (0.83), there is a substantial difference in further classification. Clusters 1 and 2 have opposite values for second canonical variable, correlated to the number of lateral branches on main stem (0.44), which is higher for the accessions in cluster 1. Second canonical variable is also responsible for distinction of cluster 3 with the higher number of lateral branches from the other clusters of the determinate type. Finally, using only continuous variables it is not possible to distinguish between clusters 4 and 5 obtained by CR method, due to substantial overlapping.

DISCUSSION

MLM homogenization of the initial groups obtained by clustering based on three different approaches to Gower's distance estimation (G, PM, and CR) revealed some weaknesses as well as advantages of these three methods. G required twice as many iterations to reach convergence of the EM algorithm, as the two remaining methods; PM required reassignment of less accessions than others; only CR classified one accession with less than 75% probability. Insight of the distribution patterns between the initial and improved groupings revealed good overall agreement between these two groupings for the PM method, while CR being the opposite with two completely reshaped clusters. This is in agreement with the results from a previous study (Gunjača et al., 2000), in which we detected the slight inferiority of CR method, using three different criteria for the comparison of methods. Furthermore, detecting only one observation classified with very low probability seems reasonable for the given number of accessions and groups. Franco et al. (1998) observed a small number of misclassifications (1 and 9) while analyzing data sets of similar size to ours; however, when they increased the number of accessions and groups, these figures got highly inflated.

Canonical analysis showed satisfying separation of the clusters on the basis of only continuous variables for the G and PM methods, while CR needed categorical variables to distinguish between clusters 4 and 5. Similarly, although comparing different clustering methods, Franco et al. (1997) have also been able to identify methods that could not clearly distinguish the groups. On the other hand, for all methods the dominant role of type of growth was still preserved, although canonical analysis used only quantitative variables. This could be explained by correlation between type of growth and plant height, which lead to the clear distinction between clusters 1-2 and 3-5.

ACKNOWLEDGEMENT
The authors wish to thank Dr. D.W. Smith who kindly provided us his SAS code for calculating distance according to aforementioned CR method

REFERENCES

Cole-Rodgers P., Smith D.W., Bosland P.W. (1997). A novel statistical approach to analyze genetic resource evaluations using *Capsicum* as an example. Crop Sci. 37: 1000-1002

Crossa J., Franco J. (2004). Statistical methods for classifying genotypes. Euphytica 137: 19-37

Franco J., Crossa J., Villasenor J., Taba S., Eberhart S.A. (1997). Classifying Mexican maize accessions using hierarchical and density search methods. Crop Sci. 37: 972-980

Franco J., Crossa J., Villasenor J., Taba S., Eberhart S.A. (1998). Classifying genetic resources by categorical and continuous variables. Crop Sci. 38: 1688-1696

Franco J., Crossa J. (2003). A method for classifying observations using categorical and continuous variables. In: Kang MS (ed) Handbook of Formulas and Software for Plant Geneticists and Breeders, The Haworth Press Inc., Binghamton, NY, pp 153-169

Gower J.C. (1971). A general coefficient of similarity and some of its properties. Biometrics 27: 857-874

Gunjača J., Šatović Z., Kolak I. (1998). Genetic diversity among Croatian common bean accessions. In: Kalpić D, Hljuz-Dobrić V (eds) Proceedings of the 20th International Conference on Information Technology Interfaces, Pula, Croatia, pp 247-250

Gunjača J., Šatović Z., Kolak I. (2000). Combining qualitative and quantitative trait data in classification of gene bank accessions. In: Kalpić D, Hljuz-Dobrić V (eds) Proceedings of the 22nd International Conference on Information Technology Interfaces, Pula, Croatia, 311-315

Henneberg R. (1990). Descriptor list for *Phaseolus vulgaris*, *Ph. coccineus*, and *Ph. lunatus*. Faculty of Agriculture, Zagreb

Orlandini S., Kolak I., Šatović Z., Rukavina H. (1996). Collecting local populations of grain legume species in Croatia. Sjemenarstvo 13: 399-416

Peeters J.P., Martinelli J.A. (1989). Hierarchical cluster analysis as a tool to manage variation in germplasm collections. Theor. Appl. Genet. 78: 42-48

SAS Institute Inc. (1989). SAS/STAT User's Guide, Version 6, 4th ed., SAS Institute Inc., Cary, NC