Genetic diversity analysis of peppers: a comparison of discarding variable methods

Elizanilda R. do Rego*[1]; Mailson M. Do Rêgo[1]; Cosme D. Cruz[2]; Paulo R. Cecon[3]; Dany S.S.L. Amaral[4] and Fernando L. Finger[4]

[1]Departamento de Fitotecnia, Centro de Ciências Agrárias, Universidade Federal de Roraima, Campus do Cauamé, BR174 Km 12, Monte Cristo, CEP 69310-270, Boa Vista, RR, Brazil; [2]Departamento de Biologia, UFV, Viçosa, MG, Brazil; [3]Departamento de Informática, UFV, Viçosa-MG; [4]Departamento de Fitotecnia, UFV, CEP 36571-000, Viçosa, MG, Brazil. (* Corresponding Author. E-mail: elizanilda@aol.com)

ABSTRACT

There are a lot of variables in genetic diversity studies, and it is necessary to know whether or not they are all important and which ones can be discarded. There are often little changes in clustering patterns if a subset of these variables is used, because the discarded variables are redundant or of little contribution to the variability. This study aimed at comparing two discards of variables methods – the Singh method and the principal components method – as well as evaluating the effect of the discards on the cluster analysis. In this analysis data of six ripe fruits traits were used. Other characters with previously known variability or collinearity were added to the analysis. The method considered being the most efficient was the one, which indicated variables that did not alter the initial clustering pattern when discarded. The Singh method did not detect variation differences when standardized data were used. When the distance was obtained by the non-standardized data, the pericarp thickness (0.018%), total soluble solids (0.1668%) and minimum width (2.99%) had the lowest contribution to the divergence. The principal components pointed out that the characteristics fruit length, total soluble solids and seeds yield/fruit were considered as dispensable variables. There were changes in the initial clustering pattern when the variable pericarp thickness was discarded, and the Singh method was not efficient in detecting the importance of this variable. There were no changes in the initial clustering pattern when fruit length was discarded. The data showed that the two compared methods differed, since Singh's and principal component methods showed different variables to be discarded. The Singh method was not efficient in detecting multicollinearity among variables. The principal component method was more efficient in pointing out the variables that can be discarded. It is advisable that the genetic divergence is calculated based on the scores of the principal components. In future studies, when there is no replicated data, the genetic divergence and the pinpoint of characters should be calculated based on the principal component scores to avoid discarding some important variables when determining divergence. However, if the variable values differ independently, the Singh method based on Euclidean distance is appropriate.

KEY WORDS: Multivariate analysis, *Capsicum,* hot peppers, biodiversity.

INTRODUCTION

The study of genetic divergence is a useful and effective tool for screening accessions in germplasm banks, studies of organism evolution and identification of superior parents in breeding programs. The importance of genetic diversity for plant selection and breeding has been emphasized in previous works (Jolliffe, 1972, 1973; Arunachalam, 1981). In these kinds of studies, several multivariate analyses can be applied, including the principal component, the canonical and the cluster analysis. The last method differ from the former two due to the dependence on previous measurements of the genetic distance, which is done by the Euclidean distance or Mahalanobis´ generalized distance (Jolliffe, 1972).

In a multivariate analysis, when a large number of correlated characters are available, the results are no hardly changed if only a subset of the total data is used (Jolliffe, 1972, 1973). The remaining variables are usually redundant and therefore can be discarded. In addition to that, time and money are saved if some variables are discarded. Likewise, computing spending time is reduced, since in further analyses fewer variables will be necessary, which facilitates the interpretation of the collected data.

In most of multivariate analyses, more variables are presented than those actually needed. The question whether they are all necessary arises and, if they are not, which variables should be discarded. Also, if distance is affected when one or more variables are added or retrieved (Arunachalam, 1981; Beale et al., 1967).

This study aimed at the evaluation of the effectiveness of two variable discarding methods: the Singh method (Singh, 1981), which compares the relative contribution of each character to the total distance, and the principal component method, which allows the elimination of variables with the largest coefficient (eigenvectors) for the last components (eigenvalues) and evaluates the effect of discarding on the cluster analyses (Jolliffe, 1972). Real data obtained from several pepper accesses were used for this analysis. Other characters with previously known variability or collinearity were added to the analysis.

MATERIAL AND METHODS

Thirty six accessions of pepper (*Capsicum baccatum* and *C. annuum*) from the Germplasm Bank of Universidade Federal de Viçosa/UFV, Viçosa, MG, Brasil were evaluated for six ripe fruit traits: minimum and maximum widths (wmin and wmax), pericarp thickness (PT), seed yields/fruit (SY), total soluble solids (TSS) and fruit length (FL).

The data were subjected to the following multivariate analyses: 1) The square of Euclidean distance was employed to determine the degree of divergence among accessions, using standardized and non-standardized data; the groups were formed following Tocher's method (Rao, 1952); 2) the analysis of the relative importance of each character, by the Singh method, with standardized and non-standardized data (Singh, 1981); and 3) the divergence analysis and relative importance of the characters using principal components.

After these analyses, the discarding of variables was done and a new distance and grouping analysis was made to evaluate the influence of the discarded characters in the initial grouping. The most efficient method was that with the least important characteristic, which did not influence the initial clustering pattern after elimination.

With the objective of comparing the relative efficiency of the two techniques for discarding variables, additional data with previously known variability and colinearity of the variables was also used. The technique that identified the variable with the least importance for diversity, either showing lowest variance or high correlation with the others was accepted as most consistent. The following strategies were adopted:

a) Incorporation of Multicollinearity - To evaluate the most effective method for detection of redundant variables, two new variables were added to the group of six original variables: G = wmin + wmax and H = TSS + FL. The analysis of the new data was carried out as previously described. The technique that showed the G and H variables as susceptible for discarding was considered the most efficient, since these are linear functions of the original variables.

b) Analysis of independent characters with different degrees of variability – a case where the characters are linearly independent, but different in variability was simulated. For such task, the scores of principal components, which are independent and retain maximal information of the total variation present in the collected data, were used. The technique that pointed out the variables with less variability, as susceptible to discard, was considered the most efficient.

All statistical and genetic analyses were carried out by the software Genes (Cruz, 1997), developed by the Department of Biology of the Universidade Federal de Viçosa/UFV, Brasil.

RESULTS AND DISCUSSION

The 36 accessions were grouped into three different clusters by the Tocher's method (Table 1) when standardized data were used. Cluster II joined two accessions, while cluster III was composed by only one accession. The remaining accessions belonged to cluster I (Table 1). For non-standardized data, the accessions were grouped into two clusters (Table 1): cluster I, which grouped the accessions 6 and 24, and a second cluster that included the remaining accessions.

A two dimensional representation of the relative positions of each accession can be seen in Figure 1. The first two principal components accounted for about 75% of the total variability among the accessions. Similar results were observed between this method and the Tocher´s method with standardized data (Figure 1 and Table 1).

The relative contribution of the fruit characteristics for the Singh´s method is presented in Table 2, and the variance (eigenvalue) associated to the principal components and their respective eigenvectors are shown in Table 3. The technique described by Singh did not detect variation differences when standardized data were used and the clustering pattern was identical to the principal component (Table 2). When the distance was obtained by non-standardized data, the pericarp thickness (0.018%), total soluble solids (0.1668%) and minimum width (2.99%) had the lowest contribution to the divergence among the accessions (Table 2). The principal component method pointed out that the fruit length, total soluble solids and seeds yield/fruit characteristics were considered as dispensable variables (Table 3). Data showed non-agreement between the two methods, since the characteristic fruit length was the one that contributed most to the divergence by the Singh´s method, 77.52% (Table 2).

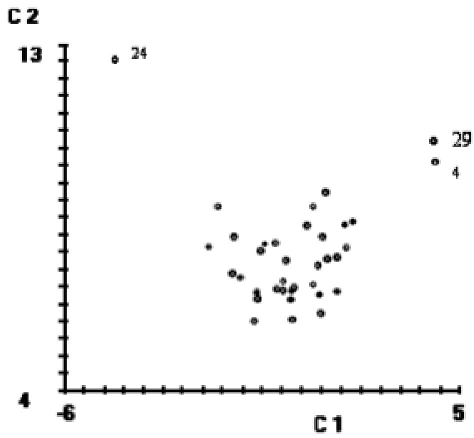

Figure 1. Relative distribution of pepper accesses of the first two principal components (C1 and C2).

Discarding variables

After discarding fruit length, which was the least important variable by the principal component analysis and following the rearrangement of the genotypes by the Tocher's method, no changes from the initial grouping pattern was observed (Table 4). Identical data were obtained when the discarded variable was total soluble solids (Table 4).

When the total soluble solid and fruit length were eliminated at the same time, there were alterations in the initial grouping pattern (Table 4). These characteristics presented a high correlation (0.78) (data not shown), indicating that only one of them should be discarded. The morphologic characters that are easier to be measured, in this particular case the fruit length, should be maintained.

When the variable pericarp thickness (PT) was discarded, there were changes in the grouping pattern for the standardized data (Table 4). The variable PT was adequate for discarding by the Singh method (Table 2). This method was not efficient in detecting the importance of this variable for the genetic divergence. It is important to point out that the Euclidean distance is influenced by the measurement scale and by the degree of correlation among

Table 1. Cluster composition of 36 accessions of pepper.

Data	Cluster / Accesses number		
	I	II	III
Standardized	Remaining	4 and 29	24
Non-standardized	Remaining	6 and 24	-

Table 2. Percentage contribution for 6 characters in pepper based on standardized and non-standardized data, by the Singh method.

Characters	Percentage Contribution	
	Non-standardized data	Standardized data
Maximum width	5.1978	16.6667
Minimum width	2.9935	16.6667
Pericarp thickness	0.0182	16.6667
Seed yield/fruit	14.096	16.6667
Total soluble solids	0.1668	16.6667
Fruit length	7 7.5276	16.6667

Table 3. Eigenvalues and eigenvectors of 6 characters in pepper[1/].

Principal Component	Eigenvalue	Variance (%)	Variance accumulated (%)	wmax	wmin	PT	SY	TSS	FL
PC1	2.64	44.12	44.12	0.4324	0.4503	0.4487	-0.0926	-0.4689	-0.4249
PC2	1.92	32.04	76.16	0.3790	0.297	0.2592	0.6089	0.3481	0.4572
PC3	0.64	10.74	86.90	-0.3434	-0.5177	0.5754	0.430	-0.3106	-0.0414
PC4	0.36	6.15	93.05	-0.4323	0.4359	-0.3832	0.5685	-0.1220	-0.3718
PC5	0.27	4.65	97.70	-0.4892	0.3398	0.5019	-0.2746	0.5440	-0.1483
PC6	0.13	2.30	100.0	-0.3540	0.3682	0.0417	-0.1931	-0.5017	0.6697

[1/]Minimum and maximum widths (wmin and wmax), pericarp thickness (PT), seed yields/fruit (SY), total soluble solids (TSS) and fruit length (FL).

Table 4. Influence of the discarding characters in the initial grouping, by the Singh (with standardized and non-standardized data) and principal component (PC) methods.

Rejection method / Discarded character	Cluster and accesses				
	I	II	III	IV	V
PC and Singh´s method (Standardized data)					
With all characters	Remaining	4 e 29	24	-	-
Without fruit length (FL)	Remaining	4 e 29	24	-	-
Without total soluble solids (TSS)	Remaining	4 e 29	24	-	-
Without TSS and FL	Remaining	4 e 29	17, 22, 19, 36 e 18	24	-
Without pericarp thickness	Remaining	29	24	-	-
SINGH (non-standardized data)					
With all characters	Remaining	6 e 24	-	-	-
Without fruit length (FL)	Remaining	4 e 29	18, 36, 22 e 17	24	3
Without total soluble solids (TSS)	Remaining	6 e 24	-	-	-
Without TSS and FL	Remaining	4 e 29	18, 36, 22 e 17	3 e 14	24
Without pericarp thickness	Remaining	6 e 24	-	-	-

characters. To overcome the former problem, usually standardized data were used.

Multivariate analyses with multicollinearity

The increment of variables correlated to the original data did not affect the grouping of the studied genotypes, which was the same obtained with the original or standardized data, as previously presented in the Table 1. This was expected since the Euclidean distance does not take into account the correlations among characters. The Singh method was not efficient in detecting the correlated characteristics as the least important (Table 5), while the principal component method indicated the two correlated variables should be the first to be discarded (Table 6).

The correlation between G and wmax, G and wmin, H and TSS and H and FL were 0.9454, 0.9023, 0.8062 and 0.996, respectively. In eucalyptus, there was no agreement among Singh's and canonical method between the selected variables and any of the rejected variables for pulp quality variables (Garcia, 1998), and Rêgo (2001) showed the same results to fruit quality traits in peppers using the Singh method and canonical analysis.

Analysis of independent characters with different degrees of variability

When independent characters with differentiated degrees of variability was used to obtain the genetic divergence based on principal component scores, there were not alterations in the grouping pattern, using the original or standardized data. The Singh method was as efficient as the principal component method in detecting variables with the largest contribution to the divergence (Tables 7 and 8), if no correlation among the variables was present.

Table 5. Percentage contribution for 6 characters in pepper based on standardized and non-standardized data, by the Singh method, with correlated characters.

Characters	Percentage Contribution	
	Non-Standardized data	Standardized data
Maximum width (wmax)	2.6252	12.50
Minimum width (wmin)	1.5007	12.50
Pericarp thickness	0.0091	12.50
Seed yield/fruit	7.0663	12.50
Total soluble solids (TSS)	0.0836	12.50
Fruit length (FL)	39.4052	12.50
G = wmax +wmin	6.9542	12.50
H = TSS + FL	42.3557	12.50

Table 6. Variance estimates (eigenvalues) of principal components and their vectors associated (eigenvectors) of 6 characters in pepper, with correlated characters[1/].

Principal component	Ev	Var. (%)	Vac. (%)	wmax	wmin	PT	SY	TSS	FL	G	H
PC1	3.77	47.14	47.14	0.3606	0.38613	0.3425	-0.0976	-0.3863	-0.375	0.4009	-0.3788
PC2	2.64	32.93	80.07	0.3663	0.2841	0.2037	0.4708	0.2881	0.3952	0.357	0.394
PC3	0.73	9.16	89.23	-0.2024	-0.3275	0.6541	0.5165	-0.2788	-0.0327	-0.2765	-0.0439
PC4	0.41	5.10	94.33	-0.3734	0.4343	-0.3743	0.6063	0.1088	-0.2827	-0.0277	-0.2678
PC5	0.28	3.55	97.88	-0.3653	0.3255	0.5215	-0.3206	0.6062	-0.0942	-0.0732	-0.064
PC6	0.17	2.12	100.00	0.4301	-0.4826	-0.0349	0.178	0.5565	-0.3631	0.0401	-0.3255
PC7	0.00	0.00	100.00	-0.4868	-0.368	0.0000	0.0000	0.0000	0.0000	0.7922	0.0000
PC8	0.00	0.00	100.00	0.0000	0.0000	0.0000	0.0000	-0.032	-0.6939	0.0000	0.7194

[1/]Eigenvalue (Ev), Variance (Var.), Variance accumulated (Vac.), Minimum and maximum widths (wmin and wmax), pericarp thickness (PT), seeds yields/fruit (SY), total soluble solids (TSS), fruit length (FL), G = wmax + wmin and H = TSS + FL.

Table 7. Percentage contribution for 6 characters in pepper based on standardized and non standardized data, by the Singh method, with independent characters and differentiated degrees of variability.

Characters	Percentage Contribution	
	Non-Standardized data	Standardized data
Maximum width (wmax)	44.12	16. 67
Minimum width (wmin)	32.04	16.6667
Pericarp thickness	10.74	16.6667
Seed yield/fruit	6.15	16.6667
Total soluble solids (TSS)	4.65	16.6667
Fruit length (FL)	2.30	16.6667

Table 8. Variance estimates (eigenvalues) of principal components and their vectors associated (eigenvectors) of 6 characters in pepper, with independent characters and differentiated degrees of variability[1/].

Principal Component	Eigenvalue	Variance (%)	Variance Accumulated (%)	wmax	wmin	PT	SY	TSS	FL
PC1	1.00	16.6667	16.6667	-0.3408	0.0975	0.2966	0.2156	0.6897	-0.5141
PC2	1.00	16.6667	33.3334	-0.1192	0.4117	-0.7334	0.3831	-0.163	-0.324
PC3	1.00	16.6667	50.0001	-0.1051	0.675	0.0943	-0.7105	-0.0603	-0.1267
PC4	1.00	16.6667	66.6668	0.8719	0.2687	0.2283	0.2025	0.0291	-0.2714
PC5	1.00	16.6667	83.3335	-0.0416	0.506	0.0818	0.3426	0.3056	0.7244
PC6	1.00	16.6667	100.000	0.3109	-0.1927	-0.5536	-0.379	0.6324	0.1275

[1/]Minimum and maximum widths (wmin and wmax), pericarp thickness (PT), seeds yields/fruit (SY), total soluble solids (TSS) and fruit length (FL).

Arunachalam (1981) considers a good procedure to join the distance and principal component analysis, if the first two components accumulate at least 70% of the total variation, which was showed by the data presented in this study (Table 3).

Considering the presented data, based on the discarded variables and the multicollinearity, we can conclude that the principal component method was more efficient than the Singh method in pointing out the variables that can be discarded, without causing alterations in the original clustering pattern.

When the data were collected in an appropriately replicate field design, and was based on the Mahalanobis' generalized distance and then compared with the canonical method, there were no differences among the variables pointed out by both methods (data not shown). This fact was observed because the Mahalanobis distance use uncorrelated transformed variables. The problem is that, usually, the accession numbers that will be measured are large, and a field design is not possible. In this case, the Euclidean distance is frequently used.

In future studies, when there is no replicated data, the genetic divergence and the pinpoint of characters should be calculated based on the principal component scores to avoid discarding some important variables when determining divergence. However, if the variable values vary independently, the Singh method based on Euclidean distance is appropriate.

ACKNOWLEDGEMENTS

This study was supported by CAPES/PICDT – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior and CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico.

RESUMO

Análise da diversidade genética de pimentas: uma comparação de métodos de descarte de variáveis

Em estudos de diversidade genética, quando é utilizado grande número de variáveis correlacionadas ocorrem poucas mudanças nos resultados se apenas um subconjunto do total de dados são utilizados. As demais variáveis sendo redundantes ou contribuindo pouco para a variabilidade, podem ser descartadas. O objetivo desse trabalho foi comparar dois métodos de descarte de variáveis: o método de Singh e o método dos componentes principais, e avaliar o efeito do descarte de variáveis sobre o padrão inicial de agrupamento. Nas análises foram utilizados dados referentes a seis características de frutos de pimentas. Outras características, como variabilidade e colineariedade previamente conhecidas, foram adicionadas. Foi considerado mais eficiente o método que indicou como menos importante para divergência a(s) variável(is), que ao serem descartadas, não alteraram o padrão inicial de agrupamento. O método de Singh não detectou diferenças quando foram usados dados padronizados e o padrão inicial de agrupamento foi o mesmo obtido pelo método dos componentes principais. Quando a medida de distância foi obtida a partir de dados não padronizados, o método de Singh apontou as variáveis que menos contribuíram para a diversidade entre acessos, como sendo: espessura do pericarpo (0.018%), sólidos solúveis totais (0.1668%) e menor diâmetro do fruto (2.99%). Por outro lado, o método dos componentes principais apontou como passíveis de descarte as variáveis comprimento do fruto, sólidos solúveis totais e número de sementes por fruto. Quando a variável espessura do pericarpo foi descartada houve alterações no padrão original de agrupamento, o mesmo não ocorreu quando o descarte feito incluiu comprimento do fruto. Os dados mostraram não concordância entre os dois métodos comparados, uma vez que esta variável foi apontada pelo método de Singh como a de maior importância para a divergência. Este método não foi eficiente em detectar multicolineariedade entre as variáveis utilizadas. Conclui-se que o método dos componentes principais foi mais eficiente que o método de Singh em apontar as variáveis a serem descartadas, sem alterações no agrupamento inicial. Recomenda-se que, em futuros trabalhos, a divergência genética seja calculada com base nos escores de componentes principais, para evitar o descarte de variáveis importantes na determinação a diversidade. Entretanto, se forem utilizadas variáveis independentes, o método Singh, baseado em cálculo de dist6ancia Euclideana pode ser utilizado.

REFERENCES

Arunachalam, V. 1981. Genetic distance in plant breeding. Indian Journal Genetics & Plant Breeding. 41:226-236.

Beale, E.M.L., Kendall, M.G. and Mann, D.W. 1967.

The discarding of the variables in multivariate analysis. Biometrika. 54:357-365.

Cruz, C.D. 1997. Programa GENES, aplicativo computacional em genética e estatística. UFV, Viçosa.

Garcia, S.L.R. 1998. Importância de características de crescimento, de qualidade da madeira e da polpa na diversidade genética de clones de eucalipto. M. S. Thesis. Universidade Federal de Viçosa, Viçosa.

Jolliffe, I.T. 1972. Discarding variables in a principal component analysis; I Artificial data. Applied Statistics. 22:160-173.

Jolliffe, I.T. 1973. Discarding variables in a principal component analysis; II Real data. Applied Statistics. 22:21-31.

Rao, A.V. 1952. Advanced statistical methods in biometrics research. John Wiley & Sons, New York.

Rêgo, E.R. 2001. Diversidade, herança e Capacidade Combinatória em Pimenta (*Capsicum baccatum*). D.S. Thesis. Universidade Federal de Viçosa, Viçosa.

Singh, D. 1981. The relative importance of characters affecting genetic divergence. Indian Journal Genetics & Plant Breeding. 41:237-45.

Received: July 05, 2001;

Accepted: May 22, 2002.