academicJournals

Vol. 10(11), pp. 1331-1337, 12 March, 2015
DOI: 10.5897/AJAR2014.9477
Article Number: 614937251470
ISSN 1991-637X

http://www.academicjournals.org/AJAR

African Journal of Agricultural Research

Full Length Research Paper

Comparison of clustering methods for study of genetic dissimilarity in soybean genotypes

P. E. Teodoro[1]*, J. P. G. Rigon[2], F. E. Torres[1], L. P. Ribeiro[1], C. C. G. Corrêa[1], F. A. Silva[1], A. Zanuncio[1], D. P. Capristo[1], M. S. Simões[1], M. C. Souza[1] and E. C. Souza[1]

[1]Departament of Fitotecnia, Universidade Estadual de Mato Grosso do Sul, Aquidauana-MS, 79-200, Brazil.
[2]Universidade Estadual Paulista "Júlio de Mesquita Filho", Botucatu-SP, 18610-307, Brazil.

Received 28 December, 2014; Accepted 11 March, 2015

This study aimed to compare clustering methods based on the dissimilarity measures and get information on genetic diversity in twelve soybean genotypes. The experiment was conducted at the State University of Mato Grosso do Sul - Unit Aquidauna. The genotypes were grown in a randomized complete block design with four replications. The following quantitative traits were measured: Plant height, first pod insertion height, number of branches and pods, mass of hundred grains and grain yield. It were evaluated the clustering methods Ward, complete linkage, median, mean linkage within group and mean linkage between groups using as dissimilarity measures the mean standardized Euclidean distance (D) and the Mahalanobis's generalized distance (D^2). The diagnosis of multicollinearity revealed adequacy of data to the proposed study. Clustering based on standardized mean Euclidean distance is distinct from those formed based on the Mahalanobis's generalized distance, being this measure most recommended to quantify the genetic diversity in soybean genotypes based on morphological traits because it presents higher values cophenetic correlation coefficient (CCC) for all clustering methods. The mean linkage between groups's hierarchical method formed concordant groups for D and D^2, being recommended for these dissimilarity measures. According to both methods, the cross between the genotypes CD238 with SYN3358 and CD238 with Potência can generate hybrids with high heterotic effect due to different numbers of loci in which the dominance effects are evident.

Key words: Dissimilarity measures, genetic divergence, *Glycine max*, quantitative descriptors.

INTRODUCTION

Among the major oilseeds grown in the world, the soybean crop (*Glycine max* (L.) Merrill) excels with production of 253 million tons of grain (harvest 2012), with Brazil accounting for 25% of production total, characterizing it as the second largest crop producer (Fao, 2013). The average increase of 36 kg ha^{-1} $year^{-1}$ in the yield, between 1976/77 up to 2012/13 (Conab, 2013), was provided mainly by genetic improvement that has obtained highly productive genotypes and adapted diverse edaphoclimatic conditions of the country.

*Corresponding author. E-mail: eduteodoro@hotmail.com

The establishment of homogeneous classes with heterogeneity among groups is a starting point in plant breeding programs. Cultivars in more distant groups provides the indicative of being dissimilar, may be considered more suitable to artificial crosses (Abreu et al., 1999), and once these materials are already on the market, these dissimilarities can be used in hybridization and/or selection schemes, aiming to incorporate the favorable traits and generate superior soybean genotypes (Cruz et al., 2004). For these measures, the clustering analysis stands out by having the purpose to gather, by some classification criterion, the parents in groups so that there is homogeneity within the group and heterogeneity between groups, being adequate to identify divergent genotypes and with most probability to succeed in the crosses (Bussab et al., 1990; Barroso and Artes, 2003).

The standardized mean Euclidean and Mahalanobis's generalized distances between pairs of genotypes are widely used as a measure of dissimilarity. The first, by not requiring repetitions, is widely used in germplasm collections, where the large number of genotypes compromises their utilization of experimental design. On the other hand the Mahalanobis distance offers the advantage of taking into account the correlations between the traits analyzed through the residual variances and covariance matrices, however, requires experiments with repetitions (Cruz et al., 2004).

The dissimilarity measure and the method used should ensure security in selection of parents for breeding. In cases of concordance of clustering, the choice of method should fall those simple to implement and easy to interpret. However, Mingoti (2005) reports that if there is disagreement between the methods, the choice of parents becomes dependent on the method used, with the need to choose the most efficient method.

Currently, research with genotypes of crops such as beans (Cargnelutti Filho et al., 2008), wheat (Bertan et al., 2006) and tomato (Karasawa et al., 2005) has been conducted aiming at establishing the most efficient dissimilarity measure and, consequently, the clustering method. However, there is heterogeneity between the results and lack of information for the soybean crop.

Given the above, the aim with this study was to compare clustering methods based on dissimilarity measures (standardized mean Euclidean and Mahalanobis's generalized), and get information about genetic divergence in twelve soybean cultivars based on six quantitative descriptors.

MATERIALS AND METHODS

The experiment was installed in the experimental area of the State University of Mato Grosso do Sul, Unit of Aquidauana (UEMS/UUA), in the city of Aquidauana-MS, whose coordinates comprise 20°27'S and 55°40'W, with an average elevation of 170 m. The soil was classified as Ultisol sandy loam texture, with the following chemical features at layer 0 - 0.20 m: pH (H_2O) = 6.2; Al exchangeable ($cmol_c\ dm^{-3}$) = 0.0; Ca+Mg ($cmol_c\ dm^{-3}$) = 4.31; P ($mg\ dm^{-3}$) = 41.3; K ($cmol_c\ dm^{-3}$) = 0.2; Organic matter ($g\ dm^{-3}$) = 19.7; V (%) = 45.0; m (%) = 0.0; Sum of bases ($cmol_c\ dm^{-3}$) = 2.3; cation exchange capacity (or CEC) ($cmol_c\ dm^{-3}$) = 5.1. The climate of the region, according to the classification described by Köppen-Geiger, is Aw (Savanna Tropical) with a cumulative rainfall during the experiment of 450 mm. The minimum and maximum temperatures over experiment were 19 and 33°C, respectively.

Twelve commercial soybean genotypes were grown (P98Y70, CD238, CD241, BRS255, VMAX, NK7059, Magna, BRS245, Potência, SY3358, MS7909 and SY5909) in a randomized blocks with four replications. *Brachiaria decumbens* was grown as predecessor plant and desiccated with the active ingredient glyphosate, at dose of 1 kg ha^{-1}. Soybean seeds were inoculated with *Bradyrhizobium japonicum* at dose of 240 g 100 kg^{-1} of seeds. The sowing was done manually on December 08th, 2011, under spacing of 0.45 m, with 16 seeds per meter. The base fertilization corresponded to 400 kg ha^{-1} in formulation 4-20-20.

At 80 days after emergence (DAE) were measured plant height (PH) and first pod insertion height (PIH). At complete maturity (115 DAE) the number of branches (NB) and pods per plan (NPP) were assessed. All variables above mentioned were analyzed in ten plants per plot, while in two central lines the mass of hundred grains (MHG) and grain yield (GY) were evaluated, with humidity corrected to 13%.

For evaluating the variability among genotypes based on these quantitative traits was used analysis of variance, with the F-test at 1% probability. Subsequently, was determined the coefficient matrix of Pearson correlation among traits (phenotypic matrix) and realized the diagnosis of multicollinearity, as recommended by Montgomery and Peck (1982).

From these traits were determined matrices standardized mean Euclidean distance (D) and Mahalanobis's distance (D^2) among genotypes. These distance matrix, in relative scale, were used as dissimilarity measure for clustering analysis in cultivars by Ward's, hierarchical method of the single linkage (nearest neighbor), complete linkage (farthest neighbor), median (WPGMA), mean linkage within group and mean linkage between group (UPGMA) (Cruz and Carneiro, 2003).

The two matrices (D and D^2) were compared using the Pearson's correlation coefficient. To validate the clustering, it was verified the ability of the dendrogram to reproduce the dissimilarity matrices (D and D^2), and the cophenetic correlation coefficient (CCC) was calculated, on which values close to unity indicate the better representation (Barroso and Artes, 2003; Cruz and Carneiro, 2003). The concordance among hierarchical methods and measures of dissimilarity was verified by CCC (Bussab et al., 1990; Mingoti, 2005). Data were analyzed using the statistical software Genes (Cruz, 2013).

RESULTS AND DISCUSSION

The F-test revealed the existence of genetic variability among the cultivars evaluated, which indicates that the population under study is promising for breeding (Table 1). Furthermore, all the traits presented heritability above 80%, value considered high by Cruz et al. (2004) which denotes low environmental influence on these traits. Similar results were also observed in other works with soybean genotypes (Sihag et al., 2004; Chettri et al., 2005; Malik et al., 2007; Rigon et al., 2012).

The diagnosis of multicollinearity, in Pearson's linear correlation coefficient among the traits revealed that the condition number (CN) was 42, featuring weak

Table 1. Values of average square and coefficient of variation of the variables plant height (PH) pod insertion height (PIH), number of branches (NB), number of pods per plant (NPP), mass of hundred grains (MHG) and grain yield (GY), evaluated in twelve soybean genotypes.

Sources of variation	PH	PIH	NB	NPP	MGH	GY
Blocks	4.37^{ns}	2.98^{ns}	8.41^{ns}	12.13^{ns}	23.44^{ns}	56.05^{ns}
Genotypes	44.09**	75.01**	91.14**	101.20**	134.05**	201.07**
Mean	0.80^{m}	0.12^{m}	15.13	84.07	22.32 g	3230.12 kg ha^{-1}
Heritability (%)	81.32	80.04	83.06	91.14	90.87	86.17
Coefficient of variation (%)	3.75	6.99	4.77	8.99	6.01	9.97

^{ns} and **: not significant and significant, respectively, by F-test at 1% probability.

Table 2. Mahalanobis's generalized distance (D^2 – upper diagonal) and standardized mean Euclidean distance (D – lower diagonal) among twelve soybean genotypes based in six quantitative descriptors.

	P98Y70	CD238	CD241	BRS255	VMAX	NK7059	Magna	BRS245	Potência	SY3358	MS7909	SY5909
P98Y70		47.21	44.71	67.99	60.91	54.25	209.94	314.42	352.27	343.81	175.76	242.50
CD238	1.77		18.91	57.58	17.85	58.52	223.15	340.74	499.72	485.85	190.17	263.79
CD241	1.38	1.25		25.46	20.41	50.60	167.63	277.62	332.02	320.64	129.45	199.51
BRS255	1.32	1.39	0.64		82.91	27.00	85.32	166.38	212.15	205.13	67.34	110.00
VMAX	1.58	0.89	0.49	0.92		112.12	298.29	437.92	401.11	484.60	242.97	339.88
NK7059	0.97	1.06	1.18	1.12	1.20		92.76	167.43	215.94	213.51	96.04	123.46
Magna	1.41	1.72	1.10	0.82	1.43	1.37		14.23	33.84	32.46	10.12	2.75
BRS245	1.60	1.94	1.50	1.17	1.78	1.59	0.41		8.04	8.69	36.77	8.41
Potência	1.76	2.42	1.89	1.56	2.21	2.01	0.87	0.64		0.84	52.83	20.34
SY3358	1.84	2.36	1.90	1.57	2.19	2.03	0.87	0.58	0.21		47.54	19.27
MS7909	1.65	2.07	1.23	1.03	1.61	1.80	0.52	0.70	0.84	0.85		1.29
SN5909	1.55	2.06	1.30	1.04	1.68	1.69	0.38	0.52	0.66	0.70	0.28	

collinearity (CN <100), according Cruz et al. (2004), showing adequacy of the data the proposed study. In the presence of multicollinearity, the use of all traits in cluster analysis is not an appropriate procedure, because the multicolinear traits are implicitly weighted with the highest weight (Cruz and Carneiro, 2003).

Dissimilarity measures estimated from the standardized mean Euclidean distance (lower diagonal) and Mahalanobis's generalized distance (upper diagonal) are presented in Table 2. Similar groups were formed between SY3358 and Potência (D=0.21 and D^2=0.84), as well between SY5909 and MS7909 (D=0.28 and D^2=1.29). Such pairs, for having the same similarity standards, are not recommended for use in breeding programs by hybridization, avoiding restriction in the genetic variability, in order to derail the gains to be obtained by selection. More distances were observed between CD238 and Potência (D=2.42 and D^2=499.72) and between CD238 and SYN3358 (D=2.36 and D^2=485.85). This high divergence, in principle, allows recommend the crossing among these pairs in order to maximize the heterosis in progenies and increase the possibility of segregants in advanced generations (Cruz et al., 2004).

The matrix D and D^2 showed significant (P <0.05) and medium magnitude (r = 0.79) linear correlation. Correlations in higher magnitude were observed by Cargnelutti Filho et al. (2008) (r=0.92) and lower by Benin et al. (2003) (r=0.529). Cruz et al. (2004) highlight that D should be used in experiments that do not include repetition, because it is difficult to quantify the environmental act influences on genetic constitutions. These techniques are recommended for the evaluation of genotypes in germplasm collections, where the large number of genotypes compromises the experimental design utilization. However, the D^2 can be estimated only when the experimental design includes repetition, allowing the environmental effects quantification on genetic constitutions.

It is important mention that the quantification of genetic similarity of soybean genotypes based on molecular information is more accurate compared to morphological traits, because there is no influence of the environment (Cruz et al., 2004). However, the acquisition of molecular information demands a high financial cost for breeding programs and is not always possible in these institutions. Thus, it is necessary to seek alternatives to classify the most divergent genotypes, as the dissimilarity based on

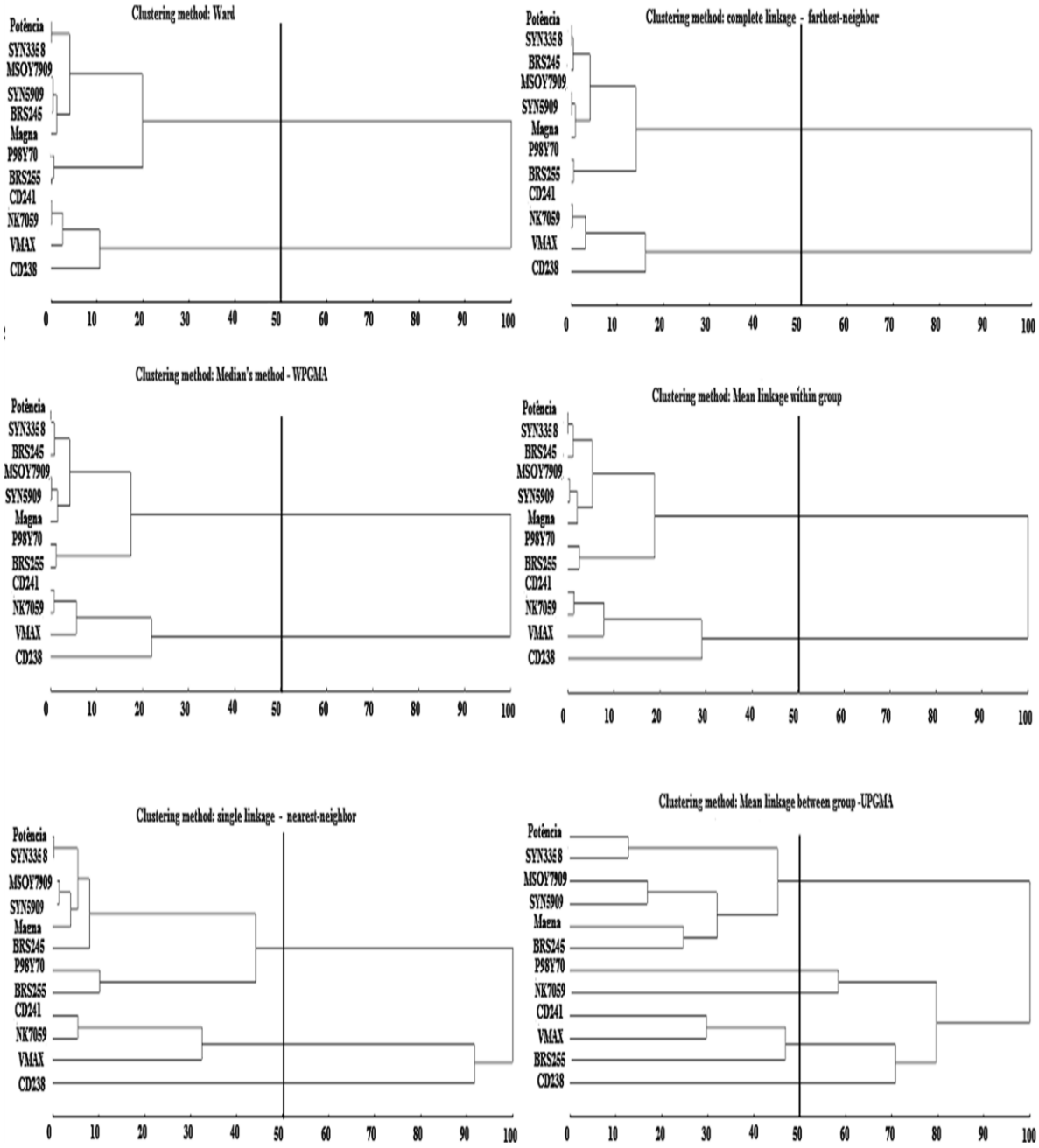

Figure 1. Dendrograms of hierarchical clustering method from the Mahalanobis's generalized distance (D^2) for twelve soybean cultivars based on six quantitative traits.

morphological traits.

In a dendrogram, large level change indicates heterogeneous cultivars union (Bonett et al., 2006). Thus, using 50% similarity as criterion for defining the groups (Cruz and Carneiro, 2003; Cruz et al., 2004) based on Mahalanobis's generalized distance (D^2), there was the formation of two identical groups by methods: hierarchical Ward's method, complete linkage, median and mean linkage within group (Figure 1), indicating a good correlation among them. These results are similar to those obtained by Cargnelutti Filho et al. (2008), Cargnelutti Filho et al. (2011) and Araújo et al. (2014), who verified a good correlation among the clustering methods evaluated, based on Mahalanobis's generalized distance (D^2).

Group 1 was formed by cultivars P98Y70, BRS255, Magna, BRS245, Potência, SYN3358, MS7909 and SYN5909, and the group 2 was formed by others

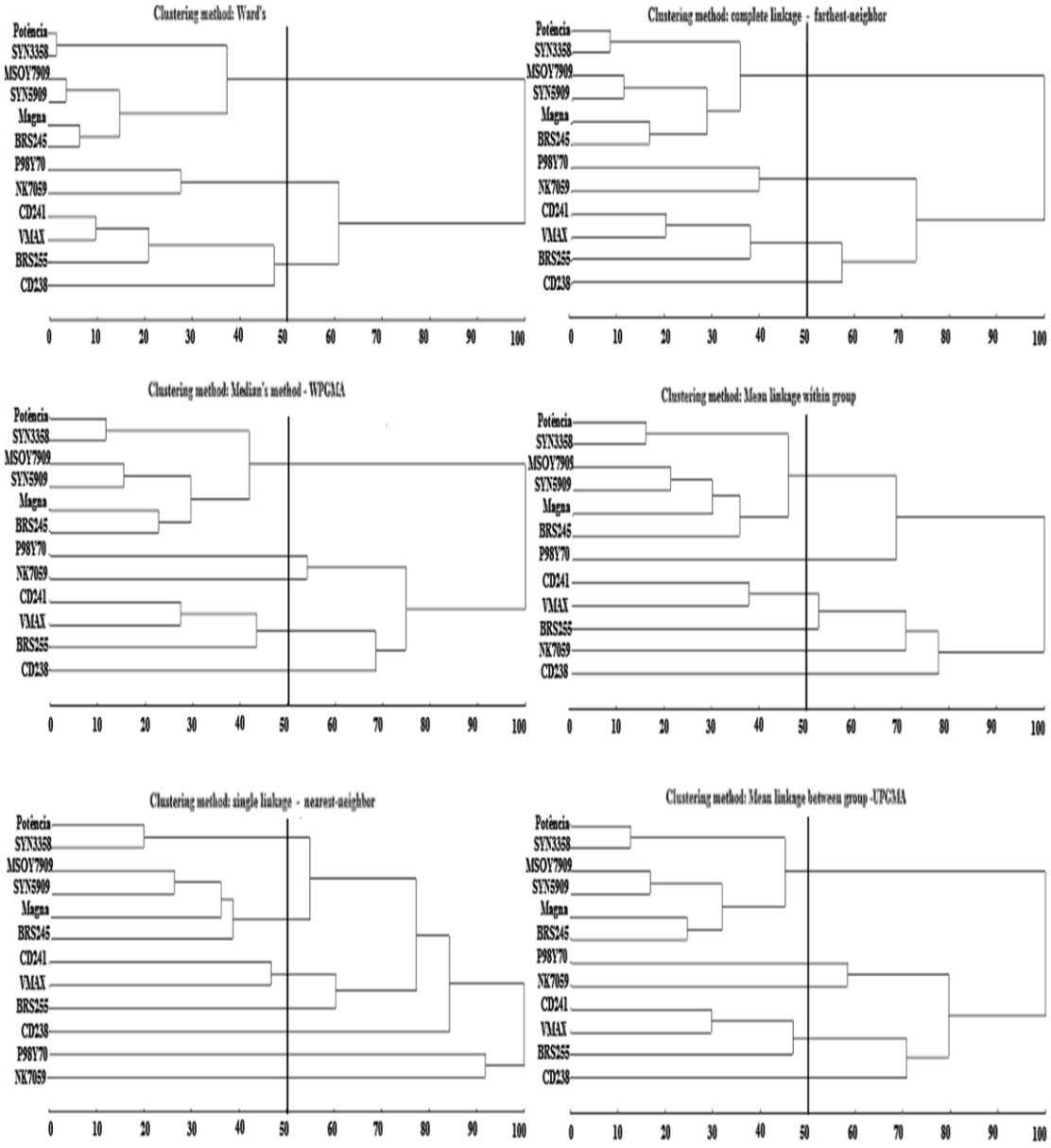

Figure 2. Dendrograms of hierarchical clustering method from standardized mean Euclidean distance (D) for twelve soybean cultivars based on six quantitative traits.

cultivars (CD241, CD238, VMAX and NK7059). There was, respectively, the formation of three and five groups using the hierarchical single linkage and mean linkage between groups, and in both cases the cultivar CD238 was part of an isolated group, so it can be used for obtaining of segregating populations with higher variability. The improvement with inbred lines extracted from improved cultivars is more favorable because such materials already possess high portion of favorable alleles and, concurrently, being evaluated in multiple environments (Amorim and Souza, 2005).

In contrast, the groups formed based on mean Euclidean distance (D) at 50% similarity were highly discordant among hierarchical methods (Figure 2). There was formation of seven, three, four, five and six groups of cultivars by the hierarchical methods single linkage, Ward, complete linkage, median, mean linkage within group and mean linkage between groups, respectively. These results are explained by the low magnitude ($0.6748 < r < 0.7183$) of the cophenetic correlation coefficient (CCC) (Table 3).

The CCC among the matrix of Mahalanobis's generalized distance (D^2) and C were significant with high magnitude, ranging from 0.7113 (single linkage) and 0.8384 (mean linkage between group). These results reinforce the hypothesis of greater consistency of groups

Table 3. Cophenetic correlation coefficient among matrix of standardized mean Euclidean (D - upper diagonal), Mahalanobis's generalized distance (D^2 - lower diagonal) and hierarchical clustering methods.

	D^2	SL	WA	CL	ME	MLWG	MLBG
D^2	0.6748**	0.7162**	0.7170**	0.7180**	0.7181**	0.7183**	0.6748**
SL	0.7113**		0.8897**	0.6755**	0.6800**	0.6164**	0.6784**
WA	0.8215**	0.9155**		0.9892**	0.9858**	0.8432**	0.9800**
CL	0.8332**	0.9177**	0.9982**		0.9963**	0.8726	0.9938**
ME	0.8373**	0.9293**	0.9981**	0.9995**		0.8894**	0.9987**
MLWG	0.7797**	0.9388**	0.9961**	0.9984**	0.9996**		0.8973**
MLBG	0.8384**	0.9392**	0.9970**	0.9982**	0.9996**	0.9998**	

** Significant at 1% probability by t-test with 64 degrees of freedom; SL, single linkage; WA, Ward's method; CL, complete linkage; ME, median; MLWG, mean linkage within group; MLBG, mean linkage between group.

formed based on Mahalanobis's generalized distance (D^2) regarding mean Euclidean distance (D).

Correlation coefficient among the clustering methods, obtained based on matrix D, ranged from 0.6164 (mean linkage within group and single linkage) and 0.9987 (mean linkage between group and median). While based on the matrix Mahalanobis's generalized distance (D^2), ranged from 0.9155 (single linkage and Ward) to 0.9998 (mean linkage between group and mean linkage within group). Mean linkage between group's method showed greater CCC in both dissimilarity measures.

In the conception of Sokal and Rohlf (1962), CCC values higher than 0.8 indicates good adjustment between the distance original matrix and the graphical distances derived. This proves the greater reliability of the clustering of mean linkage between groups, both for mean Euclidean distance (D) as well as for Mahalanobis's generalized distance (D^2), similar to results obtained by Cargnelutti Filho et al. (2008), Cargnelutti Filho et al. (2011) and Araújo et al. (2014). The best fit of the data when using this method can be explained by the fact that this method is based on arithmetic means of dissimilarity measures (Rocha et al., 2010).

From the plant breeder's perspective, the data processing by clustering methods and based on various dissimilarity measures and the consideration of particularities each one is suitable for better decision about crossings (Cargnelutti Filho et al., 2008). The clustering based on Mahalanobis's generalized distance (D^2) was consistent and all methods can be used. However, additional comparisons are needed before these considerations be generalized and also comparison with other methods based on further dissimilarity measures.

Conclusion

Clustering based on standardized mean Euclidean distance is distinct from those formed based on the Mahalanobis's generalized distance, being this measure most recommended to quantify the genetic diversity in soybean genotypes based on morphological traits because it presents higher values cophenetic correlation coefficient (CCC) for all clustering methods. The mean linkage between groups's hierarchical method formed concordant groups for D and D^2, being recommended for these dissimilarity measures. According to both methods, the cross between the genotypes CD238 with SYN3358 and CD238 with Potência can to generate hybrids with high heterotic effect due to different numbers of loci in which the dominance effects are evident.

Conflict of Interest

The author(s) have not declared any conflict of interest.

REFERENCES

Abreu AFB, Ramalho MAP, Ferreira DF (1999). Selection potential for seed yield from intra and inter-racial populations in common bean. Euphytica 108(1):121-127.

Amorim EP, Souza JC (2005). Híbridos de milho inter e intrapopulacionais obtidos a partir de populações S0 de híbridos simples comerciais. Bragantia, 64:561-567. http://dx.doi.org/10.1590/S0006-87052005000400005

Araújo LF, Almeida WS, Bertini CHCM, Vidal Neto FC, Bleicher E (2014). The use of different clustering methods in the evaluation of genetic diversity in upland cotton. Rev. Cien. Agron. 45:312-318. http://dx.doi.org/10.1590/S1806-66902014000200012

Barroso LP, Artes R (2003). Análise multivariada. UFLA, Lavras P. 151.

Benin G, Carvalho FIF, Oliveira AC, Marchioro VS, Lorencetti C, Kurek AJ, Silva JAG, Cruz PJ, Hartwig I, Schmidt DAM (2003). Comparações entre medidas de dissimilaridade e estatísticas multivariadas como critérios no direcionamento de hibridações em aveia. Ciência Rural 33:657-662. http://dx.doi.org/10.1590/S0103-84782003000400011

Bertan I, Carvalho FIF, Oliveira AC, Vieira EA, Hartwig I, Silva JAG, Shimidt DAM, Valério IP, Busato CC, Ribeiro G (2006). Comparação de métodos de agrupamento na representação da distância morfológica entre genótipos de trigo. Rev. Bras. Agrociência. 12: 279-286.

Bonett LP, Gonçalves-Vidigal MC, Schuelter AR, Vidigal Filho PS, Gonela A, Lacanallo GF (2006). Divergência genética em germoplasma de feijoeiro comum coletado no estado do Paraná,

Brasil. Semina, 27:547-560. http://dx.doi.org/10.5433/1679-0359.2006v27n4p547.
Bussab WO, Miazaki ES, Andrade DF (1990). Introdução à análise de agrupamentos. ABE, São Paulo, P.105.
Cargnelutti Filho A, Ribeiro ND, Burin C (2011). Consistência do padrão de agrupamento de cultivares de feijão conforme medidas de dissimilaridade e métodos de agrupamento. Pesq. Agropec. Bras. 45: 236-243. http://dx.doi.org/10.1590/S0100-204X2010000300002.
Cargnelutti Filho A, Ribeiro ND, Reis RCP, Souza JR, Jost E (2008). Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. Ciência Rural, 38:2138-2145.http://dx.doi.org/10.1590/S0103-84782008000800008
Chettri M, Mondal S, Nath R (2005). Studies on genetic variability in Soybean (*Glycine max* (L.) Merril) in the mid hills of Darjeeling District. J. Interac. 9:175-178.
CONAB – Companhia Nacional de Abastecimento (2013). Acompanhamento de safra brasileira: grãos, nono levantamento. Available at: http://www.conab.gov.br. Accessed 06 Nov, 2014.
Cruz CD (2013). GENES - a software package for analysis in experimental statistics and quantitative genetics. Acta Sci. Agron. 35: 271-276.http://dx.doi.org/10.4025/actasciagron.v35i3.21251
Cruz CD, Carneiro PCS (2003). Modelos biométricos aplicados ao melhoramento genético. UFV, Viçosa, P. 579.
Cruz CD, Regazzi AJ, Carneiro PCS (2004). Modelos biométricos aplicados ao melhoramento genético. UFV, Viçosa,P. 480.
FAO–Food and Agriculture Organization of the United Nations (2013). Production- Crops. Available at: http://www.fao.org. Accessed 26 Oct, 2014.
Karasawa M, Rodrigues R, Sudré CP, Silva MP, Riva EM, Amaral Jr AT (2005). Aplicação de métodos de agrupamento na quantificação da divergência genética entre acessos de tomateiro. Hortic. Bras. 23:1000-1005. http://dx.doi.org/10.1590/S0102-05362005000400028.
Malik MFA, Ashraf M, Qureshi AS, Ghafoor A (2007). Assessment of genetic variability, correlation and path analyses for yield and its components in soybean. Pakist. J. Bot. 39:405-413.
Mingoti SA (2005). Análise de dados através de métodos de estatística multivariada. UFMG, Belo Horizonte, P. 297.
Montgomery DC, Peck EA (1982). Introduction to linear regression analysis. John Wiley & Sons, New York,P. 504.
Rigon JPG, Capuani S, Brito Neto JF, Rosa GM, Wastowski AD, Rigon CAG (2012). Dissimilaridade genética e análise de trilha de cultivares de soja avaliada por meio de descritores quantitativos. Rev. Ceres 59:233-240. http://dx.doi.org/10.1590/S0034-737X2012000200012.
Rocha MC, Gonçalves LSA, Rodrigues R, Silva PRA, Carmo MGF, Abboud ACS (2010). Uso do algoritmo de Gower na determinação da divergência genética entre acessos de tomateiro do grupo cereja. Acta Sci. Agron. 32:423-431. http://dx.doi.org/10.4025/actasciagron.v32i3.4888
Sihag R, Hooda JS, Vashishtha RD, Malik RS (2004). Genetic divergence in soybean [*Glycine max* (L.) Merrill]. Annals Biol. 20:17-21.
Sokal RR, Rohlf FJ (1962). The comparison of dendrograms by objective methods. Taxon, 11:33-40. http://dx.doi.org/10.2307/1217208