GMR Genetics and Molecular Research
Online Journal
ISSN:1676-5680

Short Communication

Comparison of efficiency of distance measurement methodologies in mango (*Mangifera indica*) progenies based on physicochemical descriptors

E.O.S. Alves[1], C.B.M. Cerqueira-Silva[2], A.M. Souza[2], C.A.F. Santos[3], F.P. Lima Neto[3] and R.X. Corrêa[1]

[1]Departamento de Ciências Biológicas, Universidade Estadual de Santa Cruz, Ilhéus, BA, Brasil
[2]Departamento de Estudos Básicos e Instrumentais,
Universidade Estadual do Sudoeste da Bahia, Itapetinga, BA, Brasil
[3]Empresa Brasileira de Pesquisa Agropecuária, EMBRAPA Semi-Árido, Petrolina, PE, Brasil

Corresponding author: C.B.M. Cerqueira-Silva
E-mail: csilva@uesb.edu.br

Genet. Mol. Res. 11 (1): 591-596 (2012)
Received July 2, 2011
Accepted October 26, 2012
Published March 14, 2012
DOI http://dx.doi.org/10.4238/2012.March.14.2

ABSTRACT. We investigated seven distance measures in a set of observations of physicochemical variables of mango (*Mangifera indica*) submitted to multivariate analyses (distance, projection and grouping). To estimate the distance measurements, five mango progeny (total of 25 genotypes) were analyzed, using six fruit physicochemical descriptors (fruit weight, equatorial diameter, longitudinal diameter, total soluble solids in °Brix, total titratable acidity, and pH). The distance measurements were compared by the Spearman correlation test, projection in two-dimensional space and grouping efficiency. The Spearman correlation coefficients between the seven distance measurements were, except for the Mahalanobis' generalized distance ($0.41 \leq rs \leq 0.63$), high and significant ($rs \geq 0.91$; $P < 0.001$). Regardless of the origin of the distance matrix, the unweighted pair group method with arithmetic mean grouping method proved to be the most adequate. The various distance measurements and grouping methods gave different values for distortion ($-116.5 \leq D \leq 74.5$), cophenetic correlation ($0.26 \leq rc \leq 0.76$) and stress ($-1.9 \leq S \leq 58.9$). Choice of distance measurement and analysis methods influence the characterization of genetic variability, and this should be taken into account for studies of mango.

Key words: Grouping analysis; Multivariate statistics;
Genetic divergence; *Mangifera indica* L.

INTRODUCTION

Genetic diversity is determined using different statistical methods. Among the methods used for diversity studies are cluster analysis, calculation of measurements and data projection in two-dimensional space (Dias, 1998; Duarte et al., 1999; Mohammadi and Prasanna, 2003).

Although there are numerous statistical methods available to analyze and show data regarding genetic diversity, care should be taken in selecting these methods. Studies involving different species have shown that the choice of different statistical methods can affect the results (Gower and Legendre, 1986; Meyer et al., 2004; Cerqueira-Silva et al., 2009; Sesli and Yegenoglu, 2010). In this regard, studies have been conducted to identify the similarity coefficient (or dissimilarity) most suitable for studies of genetic diversity in species such as corn, passion fruit, olive, and wheat (Meyer et al., 2004; Balestre et al., 2008; Cerqueira-Silva et al., 2009; Sesli and Yegenoglu, 2010). In some cases, it is possible to indicate the most appropriate methods for studies of genetic diversity, whereas other studies cannot discriminate them.

In studies of the genetic diversity of mango (*Mangifera indica* L.), different measures and ratios have been applied based on different genetic markers and agronomic traits (Kumar et al., 2001; Ravishankar et al., 2004; Duval et al., 2005; Pradeepkumar et al., 2006; Díaz-Matallana et al., 2009). However, there are no studies that demonstrate which statistical method is most appropriate to assess the genetic diversity of this species.

The use of different approaches regarding method and statistical analysis in the characterization of diversity may influence the results obtained. Additionaly, there are no studies related to the efficiency of these methods for mango crops. We evaluated the influence of seven distance measurements and five clustering methods and the efficiency projection in two-dimensional space in the characterization of the diversity of five mango progeny, based on data taken from six fruit physicochemical descriptors.

MATERIAL AND METHODS

The mango fruit genotypes (genus) used in this study belong to the Collection of Mango Germplasm from Empresa Brasileira de Pesquisa Agropecuária - Semi-Árido (EMBRAPA Semi-Árido). These genotypes are part of the genetic improvement program of Embrapa Semi-Árido. To estimate the distance measurements, a total of five mango progeny (total of 25 genotypes) were analyzed during the months of August and September of 2009, using six fruit physicochemical descriptors (fruit weight, equatorial diameter, longitudinal diameter, total soluble solids in °Brix, total titratable acidity, and pH), with the use of a digital scale (precision of 0.01 g) for weight.

To estimate the distances, based on results derived from physicochemical descriptors, the following strategies for the calculation of measurements were employed: Coler-Rodgers distance (C-DR); Euclidean distance (ED); average Euclidean distance (AED); Gower distance (GD); Mahalanobis' generalized distance (MGD); weighted distance by squared residuals (WDSR); Euclidean distance squared (EDS).

Alterations in genotype ranking, obtained by the distance measurements, were analyzed by Spearman's correlation coefficient (rs). The efficacy of the measurements and coefficients, with regard to the different grouping methods (nearest neighbor, farthest neighbor, Ward, Gower, and unweighted pair group method with arithmetic mean - UPGMA), was estimated based on original and simplified dissimilarity matrices; the last resulting from the use of one grouping method with the following parameters: distortion (D) values, cophenetic correlation coefficient (rc) and stress (S).

Alterations in the efficiency of data projection in two-dimensional space, due to the choice of different measures and coefficients, were also assessed. For this, the D values, rc and S were calculated based on the original and graphic distances (adjusted for two-dimensional space). The grouping method proposed by Tocher, contained in the Genes program (Cruz, 2006), was also considered in this study. The Tocher method used the criterion of maintaining the average distance intragroup always less than any distance between groups (Rao, 1952).

The similarity and genetic distance analyses, as well as estimates of D, rc, S, and projection efficiency in two-dimensional space, were carried out using the Genes program (Cruz, 2006). Additionally, the BioEstat 5.0 program was chosen for Spearman correlation analyses (Ayres et al., 2005).

RESULTS AND DISCUSSION

The Spearman correlation coefficients between the seven distance measurements were, except for MGD ($0.41 \leq rs \leq 0.63$), elevated and significant ($rs \geq 0.91$; $P < 0.001$), indicating that the calculated distances were highly correlated and showed few changes in genotype ranking (Table 1). The $rs = 1$ ($P < 0.001$) observed between ED, AED, and EDS stands out, allowing us to infer that these three measurements show the same distance ranking between the genotypes and have differences with regard to the ranking obtained using GD and MGD. A low correlation value between distance measures was also found by Benin et al. (2003) when assessing oat genotypes, Cerqueira-Silva et al. (2009) when assessing *Passiflora* genotypes, and Sesli and Yegenoglu (2010) when assessing olive genotypes, all using agronomic descriptors or molecular markers. This low correlation may be associated, for example, with the fact that the Euclidean distances assess phenotype variation based on the average of the characteristics, while the Mahalanobis distance assesses genetic variation, based on repetitions and their deviations (Cruz, 1990; Benin et al., 2003).

The different combinations between the distance measurements and the seven grouping methods produced different results concerning the effectiveness of the grouping matrix in presenting the original distance matrix ($-116.5 \leq D \leq 74.5$; $0.26 \leq rc \leq 0.76$; $-1.9 \leq S \leq 58.9$) (Table 2). UPGMA was the most efficient among the grouping methods assessed, showing all distance measurements, distortion and stress values as being closer to zero ($3 \leq D \leq 10.8$; 17.2

≤ S ≤ 32.2) and the highest cophenetic correlation values (0.66 ≤ rc ≤ 0.76). Thus, the Ward method showed, based on all the distance measurements, the lowest efficiency in the genotype grouping (-2311.0 ≤ D ≤ -1477.0; 0.57 ≤ rc ≤ 0.70; 395.7 ≤ S ≤ 312.7). These results indicate that UPGMA has the highest/best efficiency as a grouping method, in regard to the assessment of mango quantitative variables, as well as the inefficiency of the Ward grouping method for this purpose. Similar results were found by Cerqueira-Silva et al. (2009) and Gonçalves et al. (2008) in assessing the genetic distance of passion fruit and tomato genotypes, respectively, with agronomic descriptors. Also, Cerqueira-Silva et al. (2009) and Sesli and Yegenoglu (2010) obtained similar results in studies with DNA amplification using RAPD.

Table 1. Spearman correlation coefficients between seven distance measurements of six fruit physicochemical descriptor variables measured in five populations of mango genotypes (each population consists of five genotypes).

Distance measurements	C-DR	ED	AED	GD	MGD	WDSR	EDS
Coler-Rodgers distance (C-DR)							
Euclidean distance (ED)	0.99	-					
Average Euclidean distance (AED)	0.99	1	-				
Gower distance (GD)	0.96	0.94	0.94	-			
Mahalanobis generalized distance (MGD)	0.52	0.59	0.59	0.41	-		
Weighted distance by squared residuals (WDSR)	0.95	0.95	0.95	0.91	0.63	-	
Euclidean distance squared (EDS)	0.99	1	1	0.95	0.56	0.96	-

Table 2. Efficacy of five grouping methods [nearest neighbor (NN); farthest neighbor (FN); Ward (W); Gower (WPGMC), and unweighted pair group method with arithmetic mean (UPGMA)] from different similarity (and dissimilarity) measures and coefficients, based on criteria of distortion (D) percentage, cophenetic correlation coefficient (rc) and stress (S) percentage values.

		C-DR	ED	AED	GD	MGD	WDSR	EDS
NN	D	73.3	43.3	43.3	40	70.6	74.5	73.4
	S	56.9	31.2	31.2	30.1	52.4	58.7	56.9
	rc	0.26	0.33	0.33	0.39	0.75	0.53	0.29
FN	D	66.9	-30.8	-30.8	-44	-71.5	-63	-63
	S	47.4	24.2	24.2	29.6	49.2	46.8	46.4
	rc	0.66	0.65	0.65	0.68	0.75	0.76	0.66
W	D	-72.3	-116.5	-116.5	-118.6	-47.7	-47.9	-72.1
	S	51.4	58.9	58.9	58.1	44.6	43.6	51.2
	rc	0.65	0.64	0.64	0.68	0.75	0.76	0.66
WPGMC	D	33.8	27.3	27.3	40.7	18.7	21.3	32.1
	S	34.7	39	39	29	33.2	32.9	34.2
	rc	0.65	0.58	0.58	0.64	0.75	0.76	0.65
UPGMC	D	73.3	27.3	27.3	37.4	53.5	74.5	73.4
	S	56.9	39	39	30.1	42.3	58.7	56.9
	rc	0.26	0.58	0.58	0.36	0.74	0.53	0.29
WPGMA	D	32.3	-4.4	-4.4	0.55	-15.8	-11.4	-4.2
	S	-1.9	17.8	17.8	18.9	35.6	34.2	32.5
	rc	0.66	0.65	0.65	0.69	0.75	0.76	0.66
UPGMA	D	10	3	3	3.5	10.8	10.3	9.9
	S	31.6	17.2	17.2	18.6	32.8	32.2	31.5
	rc	0.66	0.65	0.65	0.7	0.75	0.76	0.66

For other abbreviations, see Table 1.

With regard to the projection efficiency in two-dimensional space, the distance measurements showed wide variation, with stress values oscillating between 17.2 and 25% (Table 3). The highest stress values were observed based on ED and AED (25) distances, while the lowest stress values were obtained based on GD (17.2) and MGD (19.2). Because stress values were obtained, it is possible to classify the distance measures as “adequate”, according to Kruskal’s classification (1964), for the representation of distance in the assessment of mango by quantitative variables. These results differ from those presented by Cerqueira-Silva et al. (2009) for evaluations of *Passiflora*, where the use of projection in two-dimensional space was inadequate according to Kruskal’s classification (1964).

Table 3. Efficacy of the projection of similarity measurements in two-dimensional space, in mango genotypes, based on distortion (D) percentage, correlation coefficient (rc) between the original and the projected distance and stress (S) values.

Distance measurements	D	rc	S
Coler-Rodgers distance (C-DR)	5.8	0.88	20.5
Euclidean distance (ED)	16.8	0.82	25
Average Euclidean distance (AED)	16.8	0.82	25
Gower distance (GD)	10.6	0.88	17.2
Mahalanobis generalized distance (MGD)	15.4	0.99	19.2
Weighted distance by squared residuals (WDSR)	14.7	0.95	22
Euclidean distance squared (EDS)	4.8	0.87	21.1

Similar to the results observed for the two-dimensional representation (forming dendrograms and projections of the data), the formation of groups by Tocher’s modified method showed variations when using different distance measures (data not shown). Most changes were observed in the array of genotypes observed for the MGD compared with other measures.

The results confirm the hypothesis that the choice of statistical methods for analysis and presentation of diversity data in mango, from physical-chemical descriptors, influence the results. These findings are consistent with studies available for other plant species (Benin et al., 2003; Gonçalves et al., 2008; Cerqueira-Silva et al., 2009; Sesli and Yegenoglu, 2010).

We understand that although the results do not permit a final resolution on the most appropriate combinations of methods to be used in diversity studies of mango, it is possible to compare methods and at least exclude some methods. As discussed by Corrêa et al. (1999), the researcher should be aware that the efficiency of methods may be influenced by the level of heterozygosity associated with the population studied, since the number of heterozygous loci is usually different between natural and improved populations, as well as autogamous and allogamous species.

ACKNOWLEDGMENTS

Research supported by EMBRAPA, CNPq and Programa de Pós-Graduação em Genética e Biologia Molecular of UESC. Elisa Susilene Lisboa dos Santos contributed to this study during the writing of this article. This study is part of the Master’s dissertation of E.O.S. Alves.

REFERENCES

Ayres M, Ayres Junior M, Ayres DL and Santos AS (2005). Programa BioEstat 5.0. Aplicações Estatísticas nas Áreas das Ciências Biológicas e Biomédicas. Sociedade Civil Mamirauá, Belém.

Balestre M, Von Pinho RG, Souza JC and Lima JL (2008). Comparison of maize similarity and dissimilarity genetic coefficients based on microsatellite markers. *Genet. Mol. Res.* 7: 695-705.

Benin G, Carvalho FIF, Oliveira AC, Marchioro VS, et al. (2003). Comparisons among dissimilarity measures and multivariate statistics as criterions for directing hibridizations in oat. *Cienc. Rural* 33: 657-662.

Cerqueira-Silva CB, Cardoso-Silva CB, Conceicao LD, Nonato JV, et al. (2009). Comparison of coefficients and distance measurements in passion fruit plants based on molecular markers and physicochemical descriptors. *Genet. Mol. Res.* 8: 870-879.

Corrêa RX, Abdelnoor RV, Faleiro FG, Cruz CD, et al. (1999). Genetic distances in soybean based on RAPD markers. *Bragantia* 58: 15-22.

Cruz CD (1990). Aplicação de Algumas Técnicas Multivariadas no Melhoramento de Plantas. Doctoral thesis, Departamento de Genética e Melhoramento, Escola Superior de Agronomia Luiz de Queiroz, Universidade de São Paulo, Piracicaba.

Cruz CD (2006). Programa Genes: Análise Multivariada e Simulação. Editora da Universidade Federal de Viçosa, Viçosa.

Dias LAS (1998). Análises Multidimensionais. In: Eletroforese de Isoenzimas e Proteínas Afins: Fundamentos e Aplicações em Plantas e Microrganismos (Alfenas AC, ed.). Editora da Universidade Federal de Viçosa, Viçosa, 405-476.

Díaz-Matallana M, Schuler-García I, Ruiz-Garcia M and Jaramillo EH (2009). Analysis of diversity among six populations of Colombian mango (*Mangifera indica* L. cvar. Hilacha) using RAPDs markers. *Electron. J. Biotechnol.* 12: 1-8.

Duarte JM, Santos JB and Melo LC (1999). Comparison of similarity coefficients based on RAPD markers in the common bean. *Genet. Mol. Biol.* 22: 427-432.

Duval MF, Bunel J, Sitbon C and Risterucci M (2005). Development of microsatellite markers for mango (*Mangifera indica* L.). *Mol. Ecol. Notes* 5: 824-826.

Gonçalves LS, Rodrigues R, Amaral AT Jr, Karasawa M, et al. (2008). Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. *Genet. Mol. Res.* 7: 1289-1297.

Gower JC and Legendre P (1986). Metric and Euclidean properties of dissimilarity coefficients. *J. Classif.* 3: 5-48.

Kruskal J (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika* 29: 1-27.

Kumar NVH, Narayanaswamy P, Prasad DT, Mukunda GK, et al. (2001). Estimation of genetic diversity of commercial mango (*Mangifera indica* L.) cultivars using RAPD markers. *J. Hortic. Sci. Biotechnol.* 76: 529-533.

Meyer AS, Garcia AAF, Souza AP, Souza CL Jr, et al. (2004). Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L). *Genet. Mol. Biol.* 27: 83-91.

Mohammadi SA and Prasanna BM (2003). Analysis of genetic diversity in crop plants - salient statistical tools and considerations. *Crop Sci.* 43: 1235-1248.

Pradeepkumar T, Philip J and Johnkutty I (2006). Variability in physico-chemical characteristics of mango genotypes in northern Kerala. *J. Trop. Agric.* 44: 57-60.

Rao RC (1952). Advanced Statistical Methods in Biometric Research. John Wiley, New York.

Ravishankar KV, Chandrashekara P, Sreedhara SA, Dinesh MR, et al. (2004). Diverse genetic bases of Indian polyembryonic and monoembryonic mango (*Mangifera indica* L) cultivars. *Curr. Sci.* 87: 870-871.

Sesli M and Yegenoglu ED (2010). Comparison of similarity coefficients used for cluster analysis based on RAPD markers in wild olives. *Genet. Mol. Res.* 9: 2248-2253.