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

Short Communication

Comparison of methods for estimates of molecular genetic diversity in genus *Croton*: influence of coefficients, clustering strategies and data projection

M.M. Scaldaferri[1,2], J.S. Freitas[1], J.G.P. Vieira[1], Z.S. Gonçalves[1], A.M. Souza[1] and C.B.M. Cerqueira-Silva[1,2]

[1]Laboratório de Genética Molecular Aplicada,
Departamento de Ciências Exatas e Naturais
Universidade Estadual do Sudoeste da Bahia, Itapetinga, BA, Brasil
[2]Laboratório de Biologia Geral e Botânica,
Departamento de Ciências Exatas e Naturais
Universidade Estadual do Sudoeste da Bahia, Itapetinga, BA, Brasil

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

Genet. Mol. Res. 13 (3): 5566-5573 (2014)
Received July 1, 2013
Accepted November 12, 2013
Published July 25, 2014
DOI http://dx.doi.org/10.4238/2014.July.25.11

ABSTRACT. We investigated 10 similarity (and disimilarity) coefficients in a set of 40 wild genotypes of *Croton linearifolius* subjected to analyses using hierarchical grouping methods, grouping methods by optimization and data projection in two-dimensional space. Genotypes were characterized by analyzing DNA polymorphism with the use of 15 ISSR and 12 RAPD markers. The distance measurements were compared by the Spearman correlation test, projection in two-dimensional space and grouping efficiency evaluation. The Spearman correlation coefficients between the 10 coefficients evaluated were

significant ($P < 0.001$) and indicated significant changes in genotype ranking due to type of coefficient used ($0.76 \leq r_s \leq 1$). Wide variation was also observed in the efficiency of clustering methods, where the unweighted pair group method with arithmetic mean was the most suitable ($0.3 \leq D \leq 1.5$; $0.41 \leq r_c \leq 0.77$; $5.99 \leq S \leq 12.61$). Projection efficiencies in two-dimensional space showed high-stress values ($65 < S < 89\%$). Similar to the results observed for hierarchical clustering methods and for projection in two-dimensional space, the formation of groups with grouping methods by optimization showed variations when using different coefficients. We believe that the results confirm the influence of coefficients in studies of genetic diversity, showing the need to use criteria and standards for selecting appropriate methods for genetic studies of the genus *Croton*.

Key words: Genetic divergence; Multivariate statistics; Similarity coefficients; Molecular markers

INTRODUCTION

The Caatinga biome is distributed across the States of Bahia, Sergipe, Alagoas, Pernambuco, Paraíba, Rio Grande do Norte, Ceará, Piauí, and the north of Minas Gerais in Brazil (Almeida-Cortez et al., 2007). Although there is no authoritative list with the number of Caatinga species, studies indicate that the flora of this biome consists of approximately 600 recorded species, including about 200 endemic species (Tabarelli and Gascon, 2005).

Among the wild species of the Caatinga are those that belong to the genus *Croton*. Species of this genus are known to have chemical and/or pharmacological properties giving them medicinal potential (Souza et al., 2006; Palmeira Júnior et al., 2006). Among the species of *Croton*, *Croton linearifolius* is commonly used as a natural insecticide in Bahia's semiarid region (Cunha e Silva et al., 2010).

However, despite its potential use as a natural insecticide, genetic studies of *C. linearifolius* are limited to testing methods for DNA extraction (Scaldaferri et al., 2013) and preliminary studies of genetic diversity (Cerqueira-Silva CBM, personal communication). The growth and improvement of molecular techniques have been remarkable, especially for genetic studies aimed at characterizing the diversity of plants. In this context, it is possible to highlight the use of co-dominant and dominant markers.

Numerous statistical methodologies for data analysis for determining genetic diversity, as well as for graphic presentation of these estimates, are also available. Among the statistical methods used for diversity studies are i) similarity (and dissimilarity) coefficients, ii) hierarchical grouping methods and grouping methods by optimization, and iii) data projection in two-dimensional space (Duarte et al., 1999; Mohammadi and Prasanna, 2003). In view of the large number of statistical methods available in the literature and the fact that existing comparative studies reveal that the choice of statistical methods can affect the results (Meyer et al., 2004; Gonçalves et al., 2008; Cerqueira-Silva et al., 2009; Sesli and Yegenoglu, 2010; Alves et al., 2012), it is important to establish criteria for choosing statistical methods that are appropriate for the reality of each study to be conducted.

Considering that the coefficients and cluster methods used in the analysis of genetic data may influence the results obtained and that there are no studies related to the efficiency of these methods for species of the genus *Croton*, we evaluated the influence of 10 coefficients and 8 clustering methods, as well as the efficiency of projection in two-dimensional space in the characterization of the diversity of 40 wild genotypes of *C. linearifolius*, on the basis of ISSR and RAPD markers.

MATERIAL AND METHODS

Obtaining genotypes and DNA extraction

Young leaf tissue was collected from 40 wild genotypes of *C. linearifolius* (popularly known in Brazil as 'velame pimenta') in the National Forest Contendas Sincorá, city of Contendas do Sincorá, Bahia, Brazil, and stored in an ultra-cold freezer at the Laboratory of Applied Molecular Genetics of Universidade Estadual do Sudoeste da Bahia - UESB, Itapetinga, BA. DNA extraction was standardized by using 0.2 g leaf tissue in each extraction.

Genomic DNA used in this study was extracted according to the CTAB method, using routines and modifications described by Scaldaferri et al. (2013), and deposited in the genomic DNA bank of Laboratory of Applied Molecular Genetics, UESB.

Obtaining molecular data

Genotyping was performed by analyzing DNA polymorphism with the use of 15 ISSR primers (DiCA 3G, DiCA 3RG, DiGA 3C, DiGA 3RC, DiGA 3T, TriCAC 3RC, TriCAC 5CY, TriCAG 3RC, TriTGT 3YC, TriAAR 3RC, TriAAG 3RC, TriACG 3RC, TriAGA 3RC, TriCGC 3RC, and TriGAC 3RC) and 12 RAPD primers (OPD-01, OPD-03, OPD-05, OPD-06, OPD-08, OPD-10, OPD11, OPD13, OPD15, OPD-16, OPD-18, and OPD-20). These primers were preliminarily selected among the 23 ISSR primers and 40 RAPD primers that comprise the Operon© OPD series, since they result in the best standards of amplification (data not shown). Genotyping was always performed by two researchers to enhance the reliability of the data.

Amplification reactions with ISSR and RAPD primers were performed according to dos Santos et al. (2011) and Williams et al. (1990), respectively, using an MJ 96 thermocycler (Biocycler). Amplification products were then mixed with an EZ Vision buffer (according to manufacturer specifications) and separated by electrophoresis on 2% (w/v) agarose gel submerged in Tris-borate and EDTA buffer (1X TBE), at a constant voltage of 120 V for approximately 2 h. Finally, the ethidium bromide-stained gels were exposed to ultraviolet light and visualized in a Kodak photodocumentation system. The gels were then evaluated by two researchers to construct a binary data matrix (0 for absence and 1 for the presence of bands).

Statistical analysis

For genetic estimates, with binary data, coefficients were obtained for Sokal distance (Sd), as well as simple matching (SM), Rogers and Tanimoto (RT), Sokal and Sneath (SS), Russell and Rao (RR), Jaccard (J), Sorensen-Dice (SD), Ochiai (O), Baroni, Urbani and Buser (BUB) and Index II (I-II) (Table 1). Aiming to verify the existence of difference in genotype

ranking, we determined the Spearman (r_s) correlation for the 10 coefficients used in this study.

Different hierarchical clustering methods were used: closest neighbor; farthest neighbor, Ward method, weighted pair-group method using arithmetic averages (WPGMA), average linkage in groups, unweighted pair-group method using arithmetic averages (UPGMA), Gower method (WPGMC), and centroid method (UPGMC). Distortion values (D), cophenetic correlation coefficient (r_c) and stress (S) were used as parameters to evaluate the efficiency of grouping methods and data projection in two-dimensional space. Stress level (%) observed from the grouping and projections were classified according to Kruskal (1964) as follows: unsatisfactory ($S \geq 40\%$), regular ($20\% \leq S < 40\%$), good ($10\% \leq S < 20\%$), excellent ($5\% \leq S < 10\%$), and perfect ($0\% \leq S < 5\%$). Grouping methods by optimization (Tocher method and Tocher modified method), contained in the Genes program (Cruz, 2006), was also performed.

Estimates of similarity, projection in two-dimensional space, estimates of D, r_c, S, and grouping methods by optimization were carried out using the Genes program (Cruz, 2006). In turn, the BioEstat 5.0 program was used for correlation analyses (Ayres et al., 2005).

Table 1. Similarity (and dissimilarity) coefficients used among 40 wild genotypes of *Croton linearifolius* genotyped with ISSR and RAPD markers.

Coefficient	Expression*	Interval
Sokal distance (Sd)	[(b+c)/(a+b+c+d)]1/2	0-1
Simple matching (SM)	(a+d)/(a+b+c+d)	0-1
Rogers and Tanimoto (RT)	(a+d)/[a+2(b+c)+d]	0-1
Sokal and Sneath (SS)	2(a+d)/[2(a+d)+b+c]	0-1
Russell and Rao (RR)	a/(a+b+c+d)	0-1
Jaccard (J)	a/(a+b+c)	0-1
Sorensen-Dice (SD)	2a/(2a+b+c)	0-1
Ochiai (O)	a/[(a+b)(a+c)]1/2	0-1
Baroni, Urbani and Buser (BUB)	[a+(ad)1/2]/[a+b+c(ad)1/2]	0-1
Index II (I-II)	0.5[a/(a+b)+a/(a+c)]	0-1

*a = 1-1; b = 1-0; c = 0-1; d = 0-0.

RESULTS AND DISCUSSION

The Spearman correlation for the 10 coefficients evaluated were high and significant ($r_s \geq 0.95$; $P < 0.001$) for most of the correlations evaluated, and examples of values observed between Sd and SM and RT and SS are given in Table 2. However, significant changes in genotype ranking were expected from the correlation values observed between RR and Sd, SM, RT, and SS, among others ($0.76 \leq r_s \leq 0.89$, $P < 0.001$). Variations in correlation values were also observed in studies comparing similarity coefficients in maize (*Zea mays* L.) (Meyer et al., 2004) and passion fruit (*Passiflora edulis* Sims) (Cerqueira-Silva et al., 2009), namely 0.74 to 1 and 0.31 to 1, respectively.

In relation to distances estimated on the basis of the 10 coefficients examined, the projection efficiencies in two-dimensional space displayed high-stress values ($65 < S < 89\%$) (Table 3). These stress values, according to the classification of Kruskal (1964), are considered to be inadequate for projection in two-dimensional space, based on the matrix of binary data from dominant markers in *C. linearifolius*. However, considering the values of cophenetic correlation for *C. linearifolius*, the best results were observed in the Russell and Rao (r_c =

0.57) and Ochiai (r_c = 0.55) coefficients. Contrasting results in relation to the projection of the genetic distances in two-dimensional are available in the literature, such as the high-stress levels of these projections in studies with passion fruit (54 < S < 75), based on RAPD markers (Cerqueira-Silva et al., 2009), and lower values in studies with common bean (*Phaseolus vulgaris* L.) (11 < S < 57), based on RAPD markers (Duarte et al., 1999), and with mango (*Mangifera indica* L.) (17 < S < 23), based on fruit physicochemical descriptors (Alves et al., 2012).

Table 2. Spearman correlation coefficients* between 10 coefficients related to DNA amplifications with 15 ISSR and 12 RAPD markers, detected in 40 wild genotypes of *Croton linearifolius*.

Coefficient	Sd	SM	RT	SS	RR	J	SD	O	BUD	I-II
Sokal distance (Sd)	1.00									
Simple matching (SM)	1.00	1.00								
Rogers and Tanimoto (RT)	1.00	1.00	1.00							
Sokal and Sneath (SS)	1.00	1.00	0.99	1.00						
Russell and Rao (RR)	0.77	0.76	0.77	0.76	1.00					
Jaccard (J)	0.89	0.89	0.89	0.89	0.97	1.00				
Sorensen-Dice (SD)	0.88	0.89	0.88	0.89	0.97	1.00	1.00			
Ochiai (O)	0.88	0.88	0.88	0.88	0.96	1.00	1.00	1.00		
Baroni, Urbani e Buser (BUB)	0.98	0.98	0.98	0.98	0.86	0.96	0.96	0.96	1.00	
Index II (I-II)	0.88	0.88	0.88	0.88	0.96	0.99	0.99	1.00	0.95	1.00

*All correlation coefficients in the table are significant (P < 0.001).

Table 3. Efficacy of the projection of similarity (and dissimilarity) coefficients in two-dimensional space, in wild genotypes of *Croton linearifolius*, based on distortion percentage (D), correlation between the original and the projected (r_c) distance (D) and stress (S) values.

Coefficients	*Croton linearifolius*		
	r_c	D	S
Sokal distance (Sd)	0.22	78.93	81.24
Simple matching (SM)	0.28	67.32	70.37
Rogers and Tanimoto (RT)	0.24	74.73	77.18
Sokal and Sneath (SS)	0.30	60.86	65.03
Russell and Rao (RR)	0.57	79.15	81.20
Jaccard (J)	0.37	72.47	75.11
Sorensen-Dice (SD)	0.42	65.24	68.64
Ochiai (O)	0.55	64.68	67.53
Baroni, Urbani e Buser (BUB)	0.32	67.10	70.28
Index II (I-II)	0.53	64.64	67.69

The different combinations between 10 coefficients and the 8 hierarchical clustering methods generated distinct results concerning the efficacy of the grouping data in presenting the original distance data (-32370 ≤ D ≤ 55.1; 0.10 ≤ r_c ≤ 0.77; 5.9 ≤ S ≤ 1703) (Table 4). In summary, the methods WPGMA (0.41 < r_c < 0.75; -26.19 < D < 2.52; 6.3 < S < 18.95) and UPGMA (0.41 < r_c < 0.77; 0.3 < D < 1.5; 5.99 < S < 12.61) were the ones that showed the best results for *C. linearifolius*. In contrast, the Ward method showed the most unsatisfactory results (0.27 < r_c < 0.62; -20379 < D < 32370; 1333 < S < 1703). Similar results were found by Gonçalves et al. (2008) in studies with tomato accessions and by Sesli and Yegenoglu (2010) in studies with wild olives. In general, our results support the findings reported by Mohammadi and Prasanna (2003) in that the UPGMA is among agglomerative hierarchical methods most commonly used.

Table 4. Efficacy of five grouping methods (closest neighbor - CN; farthest neighbor - FN; Ward method - W; weighted pair-group method using arithmetic averages - WPGMA; average linkage in groups - ALG; unweighted pair-group method using arithmetic averages - UPGMA; Gower method - WPGMC; and centroid method - UPGMC) from different similarity (and dissimilarity) coefficients in wild genotypes of *Croton linearifolius*, based on criteria of distortion percentage (D), cophenetic correlation (r_c) and stress percentage values (S).

Coefficient*	CN			FN			W			WPGMA			ALG			UPGMA			WPGMC			UPGMC		
	r_c	D	S	r_c	D	S	r_c	D	S	r_c	D	S	r_c	D	S	r_c	D	S	r_c	D	S	r_c	D	S
Sd	0.24	38.0	23.9	0.25	-80.5	38.7	0.30	-31829	1688	0.41	-3.0	11.6	0.32	6.4	12.1	0.41	1.3	11.41	0.37	55.1	35.24	0.10	18.86	16.90
SM	0.25	32.4	19.9	0.26	-57.4	28.8	0.31	-31822	1688	0.42	-2.5	9.4	0.33	5.0	9.8	0.42	0.8	9.19	0.37	54.7	34.14	0.11	15.33	13.82
RT	0.24	41.2	26.2	0.25	-91.7	43.3	0.29	-30800	1659	0.41	-3.1	12.9	0.32	7.2	13.4	0.41	1.5	12.61	0.37	54.8	35.50	0.10	20.80	18.61
SS	0.26	22.7	13.5	0.28	-32.8	17.4	0.33	-32370	1703	0.42	-2.6	6.3	0.35	3.2	6.4	0.44	0.3	5.99	0.34	55.3	33.78	0.12	10.09	9.17
RR	0.71	49.0	30.9	0.67	-58.8	30.8	0.62	-20379	1333	0.75	-26.1	18.9	0.74	26.7	19.0	0.77	1.5	12.52	0.74	42.5	27.22	0.69	48.31	30.60
J	0.54	45.2	28.5	0.35	-78.2	38.6	0.28	-30531	1652	0.64	-17.9	16.1	0.42	13.7	15.8	0.66	1.4	12.12	0.64	47.9	30.32	0.50	43.72	27.78
SD	0.56	34.3	20.7	0.36	-46.8	24.7	0.27	-31681	1684	0.64	-11.8	10.9	0.43	9.3	11.0	0.66	0.7	8.51	0.64	50.9	31.00	0.51	32.90	20.14
O	0.61	32.4	19.5	0.24	-50.7	26.8	0.32	-30828	1660	0.67	-11.6	10.4	0.52	11.2	10.9	0.70	0.6	7.99	0.66	51.4	31.26	0.57	30.94	18.87
BUD	0.35	32.5	19.8	0.25	-56.1	28.3	0.29	-32357	1703	0.50	-7.9	10.2	0.33	5.6	9.9	0.53	0.7	8.68	0.46	53.7	33.22	0.23	27.98	17.80
I-II	0.65	30.4	18.3	0.24	-46.7	25.2	0.33	-30595	1654	0.70	-10.8	9.8	0.55	11.7	10.8	0.72	0.5	7.62	0.70	51.6	31.30	0.62	28.95	17.61

*The coefficients presented follow the order and nomenclature presented in the previous tables.

Similar to the results observed for hierarchical clustering methods and for projection in two-dimensional space, the formation of groups with grouping methods by optimization (Tocher method and Tocher modified method) showed variations when using different coefficients (data not shown). Variations in the results from clustering by optimization methods were also observed by Alves et al. (2012) and Duarte et al. (1999), demonstrating the importance of the choice of coefficients to be used in estimating similarity (or dissimilarity), since this choice influences the results of grouping, regardless of the method chosen for training groups. It should be noted that the combined use of hierarchical clustering methods and optimization is commonly observed in the literature, for example, in the articles of Duarte et al. (1999) with common bean, Bertan et al. (1999) with wheat genotypes, and Alves et al. (2012) with progeny of mango.

To check the consistency of these results with other species of the genus *Croton*, the evaluations performed with *C. linearifolius* were repeated in wild genotypes of *C. heliotropifolius* (belonging to the genomic DNA bank of the Laboratory of Applied Molecular Genetics, UESB), and we observed the same pattern of results (data not shown).

Considering the results obtained in this study, as well as discussions available in literature about the choice of coefficients and clustering methods for genetic studies of plants, we believe that the present results confirm the influence of coefficients on genetic diversity, which can lead to difficulties in comparisons between different research results. Therefore, it is necessary to use criteria and standards for selecting appropriate methods for genetic studies of the genus *Croton*, where it is possible to clearly identify inadequate methodological strategies for evaluating genetic data for this genus, at least on the basis of the molecular data considered in this study.

ACKNOWLEDGMENTS

Research supported by UESB, CNPq and Programa de Pós-Graduação em Ciências Ambientais of UESB (Campus Itapetinga, BA). We thank the researchers Elisa S.L. Santos (UESB), Fernanda A. Gaiotto (UESC) and Onildo N. Jesus (EMBRAPA) for their contribution during the writing and review of this article. This study is part of the Master's dissertation of M.M. Scaldaferri.

REFERENCES

Almeida-Cortez JS, Cortez PHM, Franco JMV and Uzunian A (2007). Caatinga - Coleção Biomas do Brasil. Harbra, São Paulo.

Alves EO, Cerqueira-Silva CB, Souza AM, Santos CA, et al. (2012). Comparison of efficiency of distance measurement methodologies in mango (*Mangifera indica*) progenies based on physicochemical descriptors. *Genet. Mol. Res.* 11: 591-596.

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.

Bertan I, Carvalho FIF, Oliveira AC, Vieira EA, et al. (1999). Comparison of clustering methods representing morphological distances between wheat genotypes. *Rev. Bras. Agroc.* 12: 279-286.

Cerqueira-Silva CBM, 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.

Cruz CD (2006). Programa Genes: Aplicativo Computacional em Genética e Estatística. Editora UFV, Viçosa.

Cunha e Silva SL, Carvalho MG, Gualberto SA, Carneiro-Torres DS, et al. (2010). Bioatividade do extrato etanólico do caule de *Croton linearifolius* Mull. Arg. (Euphorbiaceae) sobre *Cochliomyia macellaria* (Diptera: Calliphoridae). *Acta Vet. Bras.* 4: 252-258.

dos Santos LF, de Oliveira EJ, dos Santos SA, de Carvalho FM, et al. (2011). ISSR markers as a tool for the assessment of genetic diversity in *Passiflora*. *Biochem. Genet.* 49: 540-554.

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.

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.

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

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.

Palmeira Júnior SF, Alves VL, Moura FS, Vieira LFA, et al. (2006). Constituintes químicos das folhas e caule de *Croton sellowii* (Euphorbiaceae). *Rev. Bras. Farmacogn.* 16: 397-402.

Scaldaferri MM, Freitas JS, Santos ESL, Vieira JGP, et al. (2013). Comparison of protocols for genomic DNA extraction from 'velame pimenta' (*Croton linearifolius*), a species native to the Caatinga, Brazil. *Afr. J. Biotechnol.* 12: 4761-4766.

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.

Souza MAA, Souza SR, Veiga Júnior VF, Cortez JKPC, et al. (2006). Composição química do óleo fixo de *Croton cajucara* e determinação das suas propriedades fungicidas. *Rev. Bras. Farmacogn.* 16 (Suppl): 599-610.

Tabarelli M and Gascon C (2005). Lessons from fragmentation research: Improving management and policy guidelines for biodiversity conservation. *Conserv. Biol.* 19: 734-739.

Williams JG, Kubelik AR, Livak KJ, Rafalski JA, et al. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18: 6531-6535.