horticultura brasileira

Research

GOMES, GP; BABA, VY; SANTOS, OP; SUDRÉ, CP; BENTO, CS; RODRIGUES, R; GONÇALVES, LSA. 2019. Combinations of distance measures and clustering algorithms in pepper germplasm characterization. *Horticultura Brasileira* 37: 172-179. DOI - http://dx.doi.org/10.1590/S0102-053620190207

Combinations of distance measures and clustering algorithms in pepper germplasm characterization

Gisely Paula Gomes [1]; Viviane Yumi Baba [1]; Odair P dos Santos [1]; Cláudia P Sudré [2]; Cintia dos S Bento [3]; Rosana Rodrigues [2]; Leandro SA Gonçalves [1]

[1]Universidade Estadual de Londrina (UEL), Londrina-PR, Brazil; gipgomes@yahoo.com.br; vybaba15@gmail.com; odairjap@gmail.com; leandrosag@uel.br (correspondence author); [2]Universidade Estadual do Norte Fluminense (UENF), Campos dos Goytacazes-RJ, Brazil; claudia.pombo@yahoo.com.br; rosana@uenf.br; [3]Universidade Federal do Espirito Santo (UFES), Alegre-ES, Brazil; cdossantosbento@yahoo.com.br

ABSTRACT

Characterization and evaluation of genotypes conserved in the germplasm banks have become of great importance due to gradual loss of genetic variability and search for more adapted and productive genotypes. This can be obtained through several ways, generating quantitative and qualitative data. Joint analysis of those variables may be considered a strategy for an accurate germplasm characterization. In this study we aimed to evaluate different clustering techniques for characterization and evaluation of *Capsicum* spp. accessions using combinations of specific measures for quantitative and qualitative variables. A collection of 56 *Capsicum* spp. accessions was characterized based on 25 morphoagronomic descriptors. Six quantitative distances were used [A1) average of the range-standardized absolute difference (Gower), A2) Pearson correlation, A3) Kulczynski, A4) Canberra, A5) Bray-Curtis, and A6) Morisita] combined with distance for qualitative data [Simple Coincidence (B1)]. Clustering analyses were performed using agglomerative hierarchical methods (Ward, the nearest neighbor, the farthest neighbor, UPGMA and WPGMA). All combined distances were highly correlated. UPGMA clustering was the most efficient through cophenetic correlation and 2-norm analyses, showing a concordance between the two methods. Six clusters were considered an ideal number by UPGMA clustering, in which Gower distance showed a better adjustment for clustering. Most combined distances using UPGMA clustering allowed the separation of the accessions in relation to species, using both quantitative and qualitative data, which could be an alternative for simultaneous joint analysis, aiming to compare different clusters.

Keywords: *Capsicum* spp., multivariate analysis, clustering methods, genetic diversity, qualitative and quantitative descriptors.

RESUMO

Combinações de medidas de distância e algoritmos de agrupamento na caracterização de germoplasma de pimenta

Com o aumento da perda da variabilidade genética e a procura por genótipos mais adaptados e produtivos, a caracterização e a avaliação dos genótipos conservados em um banco de germoplasma são de elevada importância. Essas podem ser obtidas de várias formas, gerando dados quantitativos e qualitativos. A análise conjunta dessas variáveis pode ser considerada uma estratégia para a avaliação precisa do germoplasma. O presente trabalho teve como objetivo avaliar diferentes técnicas de agrupamento para caracterização e avaliação de acessos de *Capsicum* spp. utilizando combinações de medidas específicas para as variáveis quantitativas e qualitativas. Foram caracterizados 56 acessos de *Capsicum* spp. com base em 25 descritores morfoagronômicos. As distâncias analisadas foram seis quantitativas [A1) média das diferenças absolutas dos rank-padronizados (Gower), A2) correlação de Pearson, A3) Kulczynski, A4) Canberra, A5) Bray-Curtis, e A6) Morisita] combinadas com a distância para dados qualitativos [Coincidência Simples (B1)]. Os agrupamentos foram realizados pelos métodos hierárquicos aglomerativos (Ward, Vizinho Mais Próximo, Vizinho Mais Distante, UPGMA e WPGMA). Todas as distâncias combinadas foram altamente correlacionadas. O agrupamento UPGMA obteve maior eficiência pelas análises de correlação cofenética e 2-norm, indicando uma concordância entre os dois métodos. Seis grupos foram considerados como número ideal pelo agrupamento UPGMA, no qual a distância de Gower apresentou um melhor ajuste para formação dos grupos. A maioria das distâncias combinadas utilizando o agrupamento UPGMA permitiu a separação dos acessos em relação às espécies, utilizando simultaneamente dados quantitativos e qualitativos podendo ser uma alternativa para análise simultânea de dados conjuntos, visando uma comparação entre diferentes agrupamentos.

Palavras-chave: *Capsicum* spp., análise multivariada, métodos de agrupamento, divergência genética, descritores qualitativos e quantitativos.

Received on July 12, 2018; accepted on March 4, 2019

Capsicum is a highly diversified genus, in which sweet and chili peppers are inserted, being widely cultivated both in tropical and subtropical regions. This genus is a vegetable of great economic importance, mainly due to versatility in cuisine, industry, pharmacy and ornamental use. Besides being segmented and diverse, *Capsicum* genus has a great variety of products and by-products, uses and forms of consumption (Sudré *et al*., 2010; Cardoso *et al*., 2018). According to FAOSTAT (2016), the production of fresh and dehydrated sweet and

chili peppers was estimated in over 38 million tons, in a total cultivated area of 3.7 million ha. So far, 38 species of *Capsicum* were described (USA-ARS, 2015), in which only five are cultivated for commercial purposes: *C. annuum*, *C. frutescens*, *C. chinense*, *C. pubescens* and *C. baccatum*.

With increasing extinction risks and loss of genetic variability, centers for plant genetic resource conservation (CPGRC) have been established worldwide. These CPGRC can be conserved as seed and pollen collections, in the field and *in vitro*, constituting what is called germplasm bank (Engels & Visser, 2003). CPGRV conserved in germplasm banks include newly breeding and obsolete cultivars, local varieties, breeding lines obtained as intermediate products and genetic stocks, such as gene, chromosomal, and genomic mutants and wild relative *(*Ríos, 2015).

Many useful traits such as nutritional quality, resistance and/or tolerance to biotic and abiotic stresses are found among the accessions conserved in the germplasm bank. However, characterization and evaluation of these accessions are essential aiming to make them useful, in order to contribute to agricultural productivity (Dulloo *et al*., 2013). Characterization and evaluation of germplasm can be obtained through agronomic, morphological, cytological, biochemical and molecular information, in which numeric and categorical measurements are frequently involved and, in many cases, types of different variables combinations (Gonçalves *et al*., 2008; Sudré *et al*., 2010).

Different studies of *Capsicum* spp. characterization were carried out (Signorini *et al*., 2013; Araújo *et al*., 2018; Cardoso *et al*., 2018; Moreira *et al*., 2018). Nevertheless, the generation of a large number of data from different categories may be a factor which makes it difficult to analyze and interpret the results, resulting frequently in an incomplete distinction of the accessions (Oliveira *et al*., 2016). Thus, a joint analysis of variables may provide a more complete indicator of the variability in germplasm banks. Few studies have used this strategy mainly due to the lack of knowledge of which statistics techniques allow this approach, in addition to the tendency of researchers to give more importance to those variables which are directly related to traits to be improved in a breeding program (Gonçalves *et al*., 2008; Moura *et al*., 2010).

Gower (1971) proposed a joint similarity measure of variables, being widely adopted in several studies on characterization and evaluation of germplasm of different species (Gonçalves *et al*., 2008; Moura *et al*., 2010; Brandão *et al*., 2013; Kyriakopoulou *et al*., 2014; Abid *et al*., 2015; Oliveira *et al*., 2016). Another way to study the variables together is to combine specific measures for quantitative and qualitative variables using a pre-determined weight. Sarkar *et al*. (2015) proposed a mix of six measures of combined distance, considering three for quantitative data (a1: average of the range-standardized absolute difference, a2: Pearson correlation and a3: scaling based on standard score) and two for qualitative data (b1: standardized simple coincidence and b2: distance based on the average absolute difference). The authors verified that combined distance a1b2 using k-means clustering method was the one which presented better allocation of the evaluated rice accessions.

Clustering methods which are usually used for RGV are the agglomerative hierarchical clustering UPGMA and Ward, and non-hierarchical analysis of k-means (Mohammadi & Prasanna, 2003; Crossa & Franco, 2004). Agglomerative hierarchical clusterings consist of considering that each individual is considered an individual cluster. At each step of the algorithm, the individuals are clustered, forming new clusterings until the moment when all the considered individuals will be in a single group. K-means method partitions n individuals into k groups in which each individual belongs to the group closest to the average (Mingoti, 2005).

One of the main advantages of k-means method in relation to the hierarchical methods is the possibility of a pattern changes clustering with algorithm evolution. However, the disadvantage is that the number of clusterings has to be chosen *a priori*, which may infer in misinterpretations about data structure if the number of clusters is not optimal.

In agglomerative hierarchical clusterings the definition of the best method is often performed by the co-phenotype correlation coefficient (CCC) based on Pearson's correlation. However, CCC may not always be a reliable measure of distortions generated by algorithms (Mérigot *et al*., 2010; Carteron *et al*., 2012). Thus, Mérigot *et al*. (2010) proposed a methodology based on a *norm* matrix between dissimilarity matrices (D) and clustering (U). One *norm* allows to define one distance between D and U which verifies general properties of non-negativity, symmetry, and certainty.

This study aims to evaluate different clustering techniques for characterizing and evaluating *Capsicum* spp. accessions using combinations of specific measures for quantitative and qualitative variables. The joint analysis of these variables can be considered one strategy for an accurate evaluation and knowledge of variability of species in germplasm banks.

MATERIAL AND METHODS

We evaluated 56 *Capsicum* spp. accessions of the Germplasm Bank of Universidade Estadual do Norte Fluminense Darcy Ribeiro (UENF), belonging to 17*C. annuum* species (G1 - G17), 15 to *C. baccatum* (G18 - G32), 18 to *C. chinense* (G33 – G50) and six to *C. frutescens* (G51 – G56). The experimental arrangement was in randomized blocks, with three replicates and ten plants per plot.

The accessions were characterized and evaluated based on morphological and agronomic descriptors proposed by Bioversity International (htpp://www.bioversityinternational.org) for *Capsicum* spp. For morphoagronomic characterization, the experiment was carried out in the municipality of Campos dos Goytacazes, Rio de Janeiro (21º45'S, 41º18'W).

We evaluated 25 descriptors, being 14 morphological (qualitative variables) and 11 agronomic descriptors (quantitative variables). The morphological descriptors were: *i)* stem color (1= green, 2= green with purple stripes, and 3= purple), *ii)* anther color (1= yellow, 2= pale blue, 3= blue, and 4= purple), *iii)* corolla color (1= white, 2= purple, 3= white with yellow-green spots, 4= white-green, 5= yellow with purple base, 6= purple with yellow base) *iv)* number of flowers per axil (1= one, 2= two, and 3= three), *v)* flower position (1= pendant, 2= intermediate, and 3= erect), *vi)* plant growth habit (1= intermediate and 2= erect), *vii)* fruit color at intermediate stage (1= yellow, 2= green, 3= orange, 4= purple, and 5= other), *viii)* fruit color at mature stage (1= white, 2= pale orange-yellow, 3= orange-yellow, 4= pale orange, 5= orange, 6= light red, 7= red, 8= dark red, and 9= purple), *ix)* fruit shape (1= elongated, 2= round, 3= triangular, 4= campanulated, 5= blocky, 6= pitanga, 7= oval, and 8= scotch bonner), *x)* fruit surface (1= smooth, 2= semi-wrinkled, and 3= wrinkled), *xi)* number of locules per fruit (1= two, 2= three, and 3= four), *xii)* cotyledoneous leaf color (1= green, 2= purple and 3= variegated), *xiii)* calyx annular constriction (1= present and 2= absent), and *xiv)* neck at base of fruit (1= present and 2= absent). The agronomic descriptors were: *i)* fruit length (cm), *ii)* fruit width (cm), *iii)* number of seeds per fruit, *iv)* plant height (cm), *v)* plant canopy width (cm), *vi)* 1000-seed weight (g), *vii)* days to flowering, *viii)* days to fruiting, *ix)* number of fruits per plant, *x)* fruit weight per plant, and *xi)* average weight per fruit.

For combined analysis of distances (quantitative and qualitative), six distance measures for quantitative data were considered, such as:

i) Distance based on the average of the range-standardized absolute difference (Gower):

$$A_1 = \frac{1}{p}\sum_{k=1}^{p}\frac{|X_{ik} - x_{jk}|}{r_k}$$

where x_{ik} and x_{jk} are i_{th} and j_{th} accessions of k_{th} quantitative variables; r_k ranking of k_{th} variables; and p is the total number of quantitative variables (Gower, 1971).

ii) Distance based on Pearson correlation:

$$A_2 = (1 - r_{ij}^2)$$

where r_{ij} is the correlation product (similarity) between i_{th} and j_{th} accessions, so dissimilarity = 1-similarity.

iii) Kulczynski distance:

$$A_3 = 1 - 0.5\left(\frac{\sum \min(x_{ij}, x_{ik})}{\sum X_{ij}} + \frac{\sum \min(x_{ij}, x_{ik})}{\sum X_{ik}}\right)$$

where x_{ij} and x_{ik} are i_{th} and j_{th} accessions;

iv) Canberra distance:

$$A_4 = \frac{1}{p}\sum_{k=1}^{p}\frac{|x_{ik} - x_{jk}|}{x_{ik} + x_{jk}}$$

where x_{ik} and x_{jk} are i_{th} and j_{th} accessions of k_{th} quantitative variables; and p is the total number of quantitative variables.

v) Bray-Curtis distance:

$$A_5 = \frac{\sum_{k=1}^{p}|x_{ik} - x_{jk}|}{\sum_{k=1}^{p}(x_{ik} + y_{ik})}$$

where x_{ik} and x_{jk} are i_{th} and j_{th} accessions of k_{th} quantitative variables; and p is the total number of quantitative variables.

vi) Morisita distance:

$$A_6 = 1 - \frac{2\sum_{k=1}^{p} x_{ik}x_{jk}}{(\lambda_i + \lambda_j)\sum_{k=1}^{p} x_{ik}\sum_{k=1}^{p} x_{jk}}$$

where x_{ik} and x_{jk} are i_{th} and j_{th} accessions of k_{th} quantitative variables; p is the total number of quantitative variables, e .

$$\lambda_i = \frac{\sum_{k=1}^{p} x_{ik}(x_{ik}-1)}{\sum_{k=1}^{p} x_{ik}\sum_{k=1}^{p}(x_{ik-1})} \text{ e}$$

$$\lambda_j = \frac{\sum_{k=1}^{p} x_{jk}(x_{jk}-1)}{\sum_{k=1}^{p} x_{jk}\sum_{k=1}^{p}(x_{jk-1})}$$

For qualitative data, the distance based on simple coincidence was used:

$$B_1 = \frac{1}{m}\sum_{k=1}^{m} d_k$$

Where $dk = 0$ if $yik = yjk$, else $dk = 1$ (Gower, 1971).

The amplitude of the six matrix elements of quantitative distance (A_1-A_6) and qualitative distance (B1) is between 0 and 1. Thus, combination of several distance matrices was calculated with the sum of the distance corresponding to qualitative and quantitative data, such as:

$$A_1B_1 = ((a_{1ij}) + (b_{1ij}))$$

$$A_2B_1 = ((a_{2ij}) + (b_{1ij}))$$

$$A_3B_1 = ((a_{3ij}) + (b_{1ij}))$$

$$A_4B_1 = ((a_{4ij}) + (b_{1ij}))$$

$$A_5B_1 = ((a_{5ij}) + (b_{1ij}))$$

$$A_6B_1 = ((a_{6ij}) + (b_{1ij}))$$

Where (a_{1ij}), (a_{2ij}), (a_{3ij}), (a_{4ij}), (a_{5ij}), (a_{6ij}) and (b_{1ij}) represent the ij_{th} matrix elements A_1, A_2, A_3, A_4, A_5, A_6 and B_1, respectively. These combined matrices were correlated using Mantel test (1000 permutations).

Capsicum spp. accessions were clustered using different agglomerative hierarchical clusterings (Ward, the nearest neighbor method, the farthest neighbor method, UPGMA and WPGMA). Afterwards, we used cophenetic correlation coefficient (based on Pearson correlation) between combined distance matrices with grouping matrix and 2-norm analysis (Mérigot *et al*., 2010).

The optimal number of clusters (k) was determined by Frey's analyses (Frey & Van Groenewoud, 1972), pseudo-t^2 (Duda & Hart, 1973), dunn (Dunn, 1974), mcclain (McClain *et al*., 1975), cindex (Hubert & Levin, 1976), cc (Sarle, 1983), and silhouette (Rousseeuw, 1987). All these analyses were performed in R (R Core Team, 2018) using cluster, clue, and Nbclust packages.

RESULTS AND DISCUSSION

In the correlation of combined distance matrices, we noticed a high association, considering that all of them were significant at 1% probability using Mantel test (Table 1). The highest values of correlation (0.98) were observed between A1B1 x A4B1, A2B1 x A6B1, and A3B1 x A5B1, whereas the lowest value observed (0.77) was between A1B1 x A2B1.

The high correlation between combined distances is due to different factors, like similarity between some distances from quantitative data, such as, Canberra, Bray-Curtis and Gower.

The difference between Bray-Curtis and Canberra is the sum of distances *ij*, considering that in Bray-Curtis the sum is inside the fraction, whereas in Canberra, it is out of the fraction. In relation to Gower distance, the difference is the denominator, being determined by the amplitude of the accessions studied in a certain variable *k*, whereas for Bray-Curtis and Canberra this denominator is the sum of *i* and *j* for variable *k*. In relation to Pearson, Morisita and Kulczynski distances, a greater dissimilarity between them and in comparison to Canberra, Bray-Curtis and Gower is observed. Only Pearson combined distance/Simple Coincidence (A2B1) obtained correlation inferior to 0.9, when associated with the other combined distances (A1B1 x A2B1 = 0.77; A2B1 x A3B1 = 0.88; A2B1 x A4B1 = 0.84 and A2B1 x A5B1 = 0.85) (Table 1).

For most studies of plant germplasm characterization, using joint analyses of quantitative and qualitative data, Gower distance (A1B1) is widely used (Gonçalves *et al*., 2008; Adewale *et al*., 2012; Sartie *et al*., 2012; Silva *et al*., 2015). However, other combinations can be used aiming to define more reliably dissimilarity/similarity among accessions.

Evaluating cophenetic correlation coefficient (CCC) between agglomerative hierarchical clustering and combined distance matrices, UPGMA clustering obtained the highest values, ranging from 0.77 (A6B1) to 0.84 (A4B1) (Table 2). The lowest values were verified for Ward clustering which ranged from 0.60 (A2B1) to 0.76 (A1B1). According to Sokal & Rohlf (1962), values 0.9≥CCC show a very good adjustment, 0.8≤CCC<0.9 good adjustment, 0.7≤CCC<0.8 a bad adjustment and <0.7 very bad adjustment. Using this classification in the obtained results, we observed that the majority of the values obtained by UPGMA method showed a good adjustment between clustering and dissimilarity matrices.

Mérigot *et al*. (2010) have raised three criticisms about the reliability of the information obtained from CCC analysis: *i)* is a measure of intensity of monotonic relationship between dissimilarity (D) and clustering matrices (U); *ii)* is sensitive to extreme values; and *iii)* the CCC close to 1 shows a perfect correspondence between D and U, whereas the correspondence between the two matrices may indeed be weak. However, when 2-norm analysis was performed, lower values of UPGMA clustering and higher values for Ward clustering were observed, showing an agreement between CCC and 2-norm methods (Table 2).

Carteron *et al*. (2012), studying the comparison of 15 distance measures and seven agglomerative hierarchical clusterings, observed that 2-norm analysis and CCC were not in accordance with efficiency of the clustering algorithm, considering that CCC did not provide any clear indication of the efficiency of clustering algorithms. Using 2-norm analysis, UPGMA was the most efficient algorithm, whereas Ward was the least efficient (Table 2).

Despite the high difference of values observed in the 2-norm analysis between UPGMA and Ward clusterings, just little distortion between UPGMA and Ward clusterings in the Gower distance (A1B1) was observed (Figure 1). In UPGMA clustering, based on the seven criteria

Table 1. Correlation between joint distance matrices (quantitative: A1: Gower distance, A2: Pearson, A3: Kulczynski, A4: Canberra, A5: Bray-Curtiz, A6: Morisita) and qualitative (B1: Simple coincidence). Campos dos Goytacazes, UENF, 2015.

Matrices	A2B1	A3B1	A4B1	A5B1	A6B1
A1B1	0.77**	0.94**	0.98**	0.90**	0.85**
A2B1		0.88**	0.84**	0.85**	0.98**
A3B1			0.96**	0.98**	0.94**
A4B1				0.92**	0.90**
A5B1					0.92**

**,*Significant at 1 and 5% probability, respectively, by Mantel test based on 1000 simulations.

Table 2. Cophenetic correlation coefficient (CCC) and 2-norm analysis between hierarchical clustering matrices (Ward, Nearest neighbor (VMP), Farthest neighbor (VMD), UPGMA and WPGMA) and joint distance (quantitative: A1: Gower distance, A2: Pearson, A3: Kulczynski, A4: Canberra, A5: Bray-Curtis, A6: Morisita, and qualitative B1: Simple coincidence. Campos dos Goytacazes, UENF, 2015.

Genetic distance	Hierarchical clustering				
	Ward	VMP	VMD	UPGMA	WPGMA
			CCC		
A1B1	0.76	0.79	0.83	0.84	0.81
A2B1	0.60	0.68	0.73	0.78	0.77
A3B1	0.68	0.77	0.78	0.81	0.79
A4B1	0.75	0.79	0.77	0.83	0.79
A5B1	0.65	0.76	0.75	0.82	0.77
A6B1	0.64	0.72	0.76	0.81	0.76
			2-norm		
A1B1	134.51	9.09	9.50	1.69	2.40
A2B1	135.25	10.91	13.41	3.11	3.50
A3B1	115.30	8.80	9.23	1.83	2.04
A4B1	123.91	8.65	9.03	1.71	2.35
A5B1	119.10	9.89	11.52	2.14	2.18
A6B1	120.32	9.47	11.46	2.17	2.37

(frey, pseudot2, dunn, mcclain, cindex, cc and silhouette), six was the optimal number of clusters observed, being groups I and II formed by *C. chinense* accessions, group III was formed by *C. baccatum* accessions, group IV was formed by *C. frutescens* accessions and groups V and VI were formed by *C. annuum* accessions. In relation to Ward clustering analysis, separation of species into groups was also verified, except for group I which was formed by four *C. annuum* accessions and six *C. frutescens* accessions.

Comparing different combinations of distances, using UPGMA clustering, we observed that Gower distance showed better adjustment for group formation, that means, separation of species when compared with the others (Figures 2 and 3). We also observed better adjustment of groups formed by Gower distance when groups were validated by program fpc (*Flexible Procedures for Clustering* – R program).

Genus *Capsicum* species are distributed in three distinct gene complexes based on crossability. Annuum complex consists of *C. annuum, C. chinense* and *C. frutescens*. These species are integrated by morphological characteristics, derived from wild relatives of different species; they are potentially easily crossed (Onus & Pickersgill, 2004) and they have the capacity to produce interspecific hybrids (Hill *et al*., 2013); Baccatum complex consists of *C. baccatum var. baccatum, C. baccatum var. pratermissum* and *C. baccatum var. pendulum*; and pubescens complex consists of some wild species and only one cultivated species, *C. pubescens*. Thus, most of combined distances using UPGMA method allowed the separation of *Capsicum*

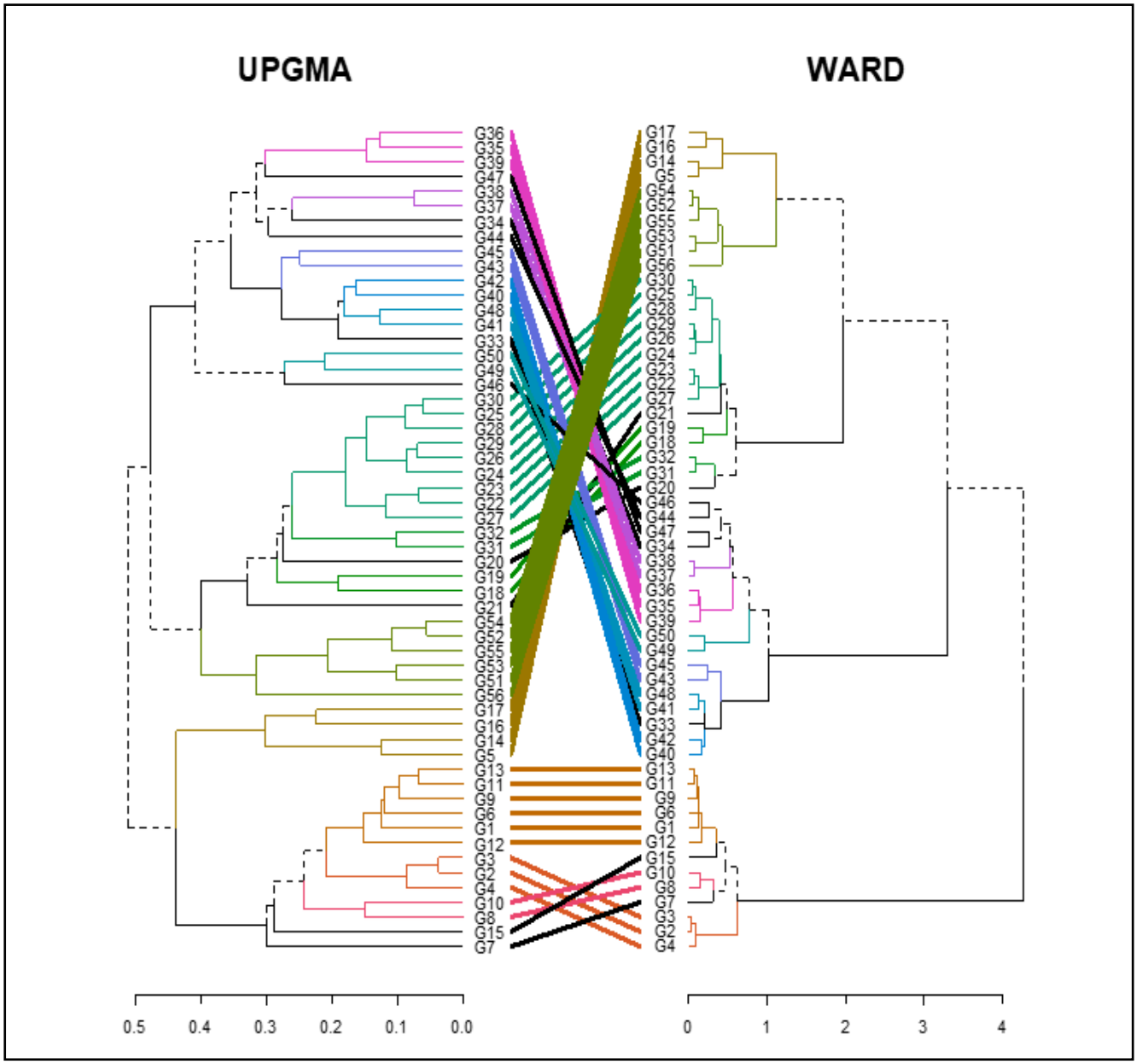

Figure 1. Genetic dissimilarity dendrogram among 56 Capsicum spp. accessions obtained by UPGMA and Ward Clustering Methods based on Gower Dissimilarity Matrix. Campos dos Goytacazes, UENF, 2015.

species, with greater efficiency in maximizing the dissimilarities between annuum and baccatum complexes. However, we did not observe any correct separation of complexes of the genus when joint analyses of quantitative and qualitative variables of the descriptors proposed by Bioversity International for *Capsicum* spp. was used. Data obtained in this study show a viable alternative in determining genetic variability and divergence among the evaluated accessions with the generation of more accurate and more complete information.

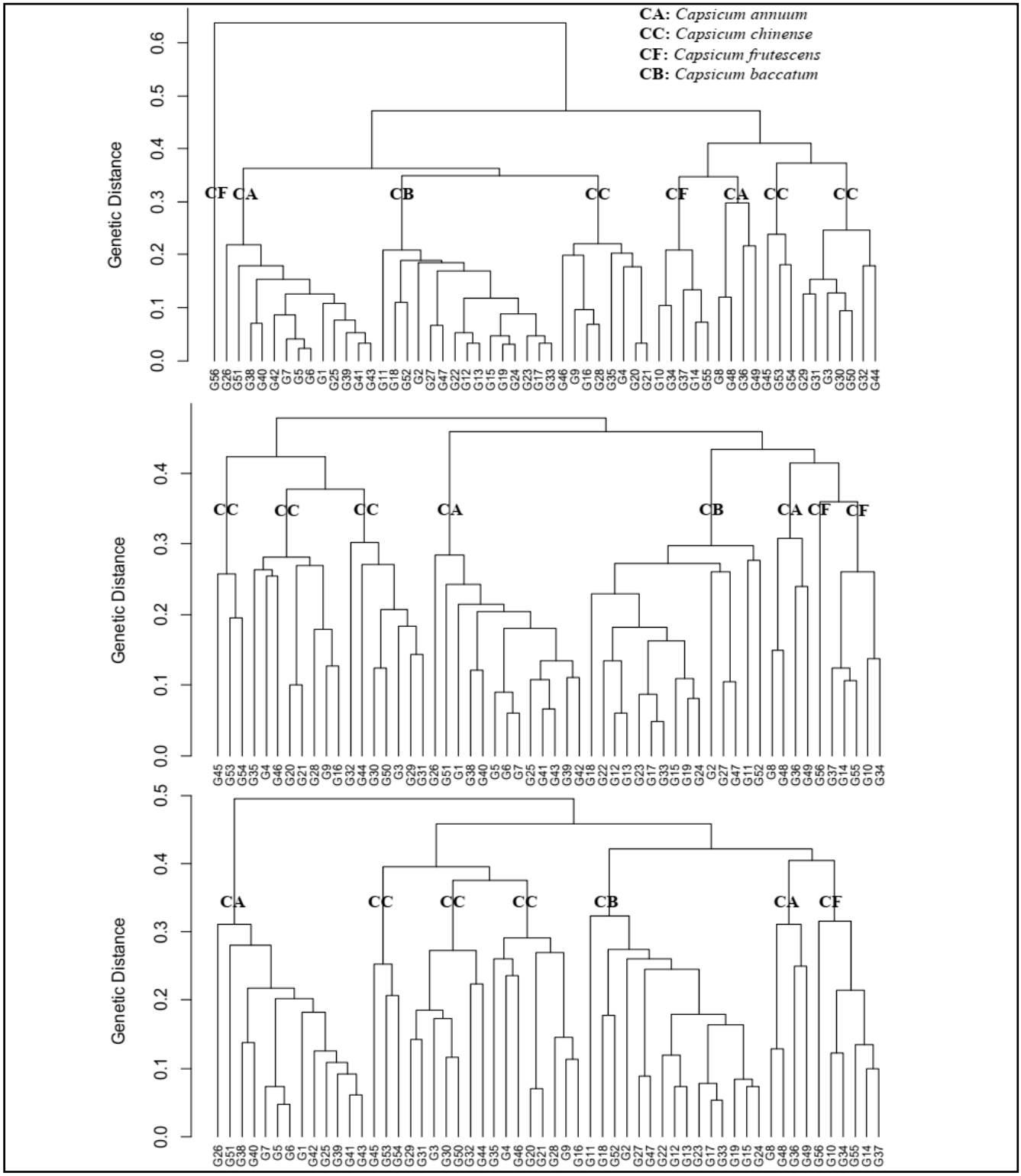

Figure 2. Genetic dissimilarity dendrogram among 56 *Capsicum* spp. accessions obtained by UPGMA clustering based on the dissimilarity matrices of joint distances (quantitative: A2: Pearson, A3: Kulczynski and A4: Canberra, and qualitative B1: Simple coincidence). Campos dos Goytacazes, UENF, 2015.

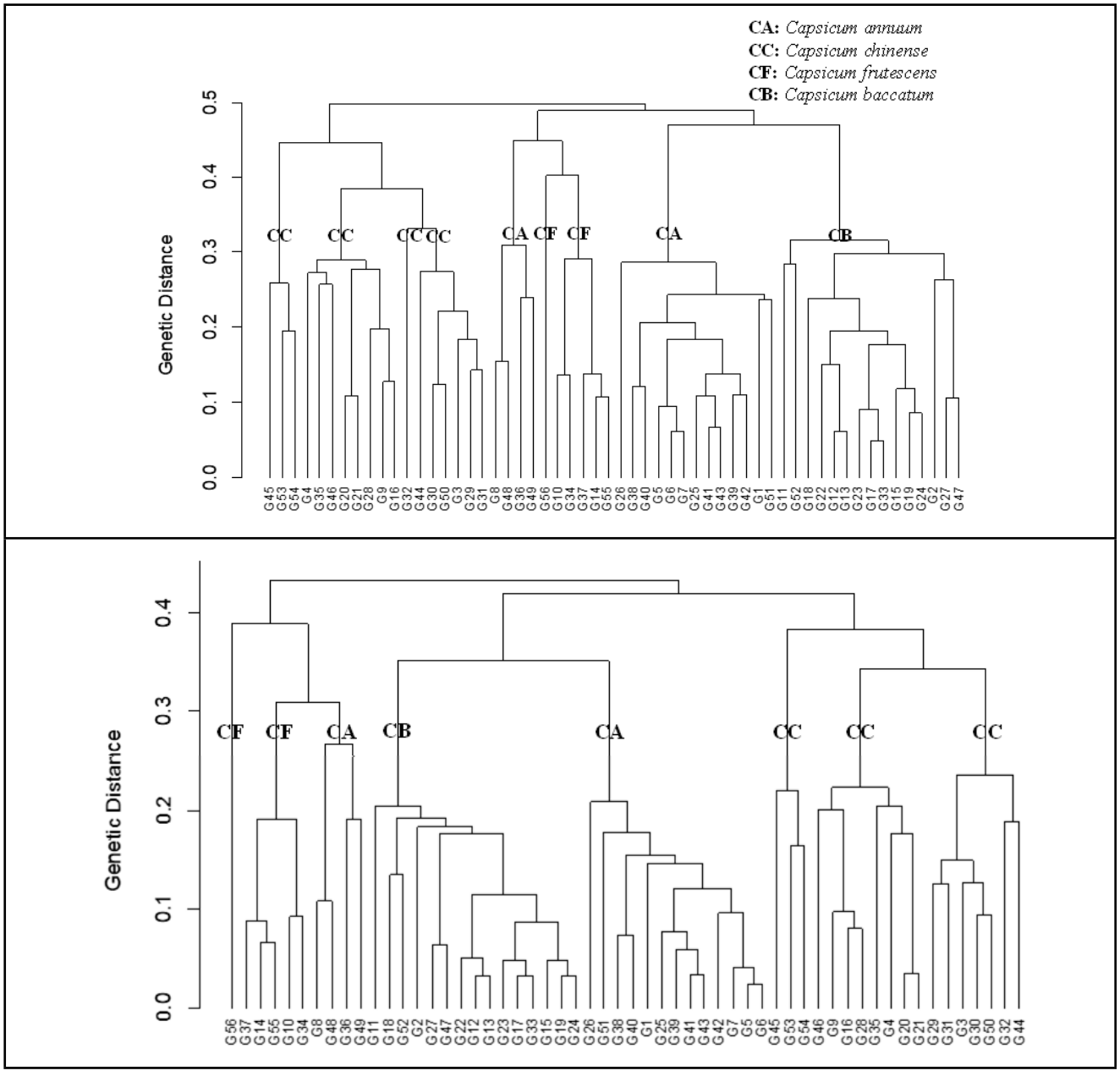

Figure 3. Genetic dissimilarity dendrogram among 56 accessions of *Capsicum* spp. obtained by UPGMA clustering based on the dissimilarity matrices of joint distances (quantitative: A5: Bray-Curtis and A6: Morisita, and qualitative B1: Simple coincidence). Campos dos Goytacazes, UENF, 2015.

REFERENCES

ABID, G; MINGEOT, D; UDUPA, SM; MUHOVSKI, Y; WATILLON, B; SASSI, K; M'HAMDI, M; SOUISSI, F; MANNAI, K; BARHOUMI, F; JEBARA, M. 2015. Genetic relationship and diversity analysis of faba bean (*Vicia Faba* L. var. *Minor*) genetic resources using morphological and microsatellite molecular markers. *Plant Molecular Biology Reporter* 33: 1755-1767.

ADEWALE, BD; DUMET, DJ; VROH-BI, I; KEHINDE, OB; OJO, DK; ADEGBIT, AE; FRANCO, J. 2012. Morphological diversity analysis of African yam bean (*Sphenostylis stenocarpa* Hochst. Ex A. Rich.) Harms and prospects for utilization in germplasm conservation and breeding. *Genetics Resource and Crop Evolution* 59: 927-936.

ARAÚJO, CMM; SILVA FILHO, DF; TICONA-BENAVENTE, CA; BATISTA, MRA. 2018. Morphoagronomic characteristics display high genetic diversity in Murupi chili pepper landraces. *Horticultura Brasileira* 36: 083-087.

BRANDÃO, LP; SOUZA, CPF; PEREIRA, VM; SILVA, SO; SANTOS-SEREJO, JA; LEDO, CAS; AMORIM, EP. 2013. Descriptor selection for banana accessions based on univariate and multivariate analysis. *Genetics and Molecular Research* 12: 1603-1620.

CARDOSO, R; RUAS, CF; GIACOMIN, RM; RUAS, PM; RUAS, EA; BARBIERI, RL; RODRIGUES, R; GONÇALVES, LSA. 2018. Genetic variability in Brazilian *Capsicum baccatum* germplasm collection assessed by morphological fruit traits and AFLP markers. *PLoS ONE* 13: e0196468.

CARTERON, A; JEANMOUGIN, M; LEPRIEUR, F; SPATHARIS, S. 2012. Assessing the efficiency of clustering algorithms and goodness-of-fit measures using phytoplankton field data. *Ecological Informatics* 9: 64-68.

CROSSA, J; FRANCO, J. 2004. Statistical methods for classifying genotypes. *Euphytica* 137: 19-37.

DUDA, RO; HART, PE. 1973. *Pattern*

classification and scene analysis. New York: John Wiley & Sons. 482p.

DULLOO, ME; THORMANN, I; FIORINO, E; FELICE, S DE; RAO, VR; SNOOK, L. 2013. Trends in research using plant genetic resources from germplasm collections: from 1996 to 2006. *Crop Science* 53: 1217-1227.

DUNN, J. 1974. Well separated cluster and optimal fuzzy partitions. *Journal Cybern* 4: 95-104.

ENGELS, JMM; VISSER, L. 2003. *A guide to effective management of germplasm collections*. IPGRI Handbooks for Genebanks, IPGRI, Rome, No. 6, 172p.

FAO - Food and Agriculture Organization of the United Nations. Statistics: FAOSTAT Domains/Production/Crops. Available http://www.fao.org/faostat/en/#data/QC. Acessed on Demberce 14, 2018.

FREY, T; VAN GROENEWOUD, H. 1972. A cluster analysis of the D-squared matrix of white spruce stands in Saskatchewan based on the maximum-minimum principle. *Journal of Ecology* 1: 873-886.

GONÇALVES, LSA; RODRIGUES, R; AMARAL JÚNIOR, AT; KARASAWA, M; SUDRÉ, CP. 2008. Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. *Genetics and Molecular Research* 7: 1289-1297.

GOWER, JC. 1971. A general coefficient of similarity and some of its properties. *Biometrics* 27: 857-874.

HILL, TA; ASHRAFI, H; WO, RCS; YAO, J; STOFFEL, K; TRUCO, JM; KOZIK, A; MICHELMORE, RW; DEYNZE, AV. 2013. Characterization of *Capsicum annuum* genetic diversity and population structure based on parallel polymorphism discovery with a 30k unigene pepper gene chip. *PLoS ONE*8: 1-16.

HUBERT, LJ; LEVIN, JR. 1976. A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin* 83: 1072-1080.

KYRIAKOPOULOU, O; ARENS, P; PELGROM, KTB; KARAPANOS, I; BEBELI, P; PASSAM, HC. 2014. Genetic and morphological diversity of okra (*Abelmoschus esculentus* [L.] Moench.) genotypes and their possible relationships, with particular reference to Greek landraces. *Scientia Horticulturae* 171: 58-70.

MCCLAIN, JO; RAO, VR. 1975. CLUSTISZ: a program to test for the quality of clustering of a set of objects. *Journal of Marketing Research* 12: 456-460.

MÉRIGOT, B; DURBEC, JB; GAERTNER, JC. 2010. On goodness-of-fit measure for dendrogram-based analyses. *Ecology* 91: 1850-1859.

MINGOTI, SA. 2005. Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada. Belo Horizonte: UFMG. 297p.

MOHAMMADI, SA; PRASANNA, BM. 2003. Analysis of genetic diversity in crop plants – salient statistical tools and considerations. *Crop Science* 43: 1235-1248.

MOREIRA, AFP; RUAS, PM; RUAS, CF; BABA, VY; GIORDANI, W; ARRUDA, IM; RODRIGUES, R; GONÇALVES, LSA. 2018. Genetic diversity, population structure and genetic parameters of fruit traits in *Capsicum chinense*. *Scientia Horticulturae* 236: 1-9.

MOURA, MCCL; GONÇALVES, LSA; SUDRÉ, CP; RODRIGUES, R; AMARAL JÚNIOR, AT; PEREIRA, TNS. 2010. Algoritmo de Gower na estimativa da divergência genética em germoplasma de pimenta. *Horticultura Brasileira* 28: 155-161.

OLIVEIRA, RL; GONÇALVES, LSA; RODRIGUES, R; BABA, VY; SUDRÉ, CP; SANTOS, MH; ARANHA, FM. 2016. Genetic divergence among pumpkin landraces. *Semina Ciências Agrárias* 37: 547-556.

ONUS, AN; PICKERSGILL, B. 2004. Unilateral incompatibility in *Capsicum* (Solanaceae): ocurrence and taxonomic distribution. *Annals of Botany* 94: 289-295.

R Core Team. 2018. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

RÍOS, RO. 2015. *Plant breeding in the omics era*. Springer. 249p.

ROUSSEEUW, P. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20: 53-65.

SARKAR, RK; MEHER, PK; WAHI, SD; MOHAPATRA, T; RAO, AR. 2015. An approach to the development of a core set of germplasm using a mixture of qualitative and quantitative data. *Plant Genetic Resources: Characterization and Utilization* 13: 96-103.

SARLE, WS. 1983. *SAS Technical Report*. Cubic clustering criterion. Cary, NC: SAS Institute Inc. 56p.

SARTIE, A; ASIEDU, R; FRANCO, J. 2012. Genetic and phenotypic diversity in a germplasm working collection of cultivated tropical yams (*Dioscorea* spp.). *Genetic Resources and Crop Evolution* 59: 1753-1765.

SIGNORINI, T; RENESTO, E; MACHADO, MFPS; BESPALHOK, DN; MONTEIRO, ER. 2013. Diversidade genética de espécies de *Capsicum* com base em dados de isozimas. *Horticultura Brasileira* 31: 534-539.

SILVA, CQ; JASMIM, JM; SANTOS, JO; BENTO, CS; SUDRÉ, CP; RODRIGUES, R. 2015. Phenotyping and selecting parents for ornamental purposes in pepper accessions. *Horticultura Brasileira* 33: 66-73.

SOKAL, RR; ROHLF, FJ. 1962. The comparison of dendrograms by objective methods. *Taxon* 9: 33-40.

SUDRÉ, CP; GONÇALVES, LSA; RODRIGUES, R; AMARAL JÚNIOR, AT; RIVA-SOUZA, EM; BENTO, CS. 2010. Genetic variability in domesticated *Capsicum* spp. as assessed by morphological and agronomic data in mixed statistical analysis. *Genetics and Molecular Research* 9: 283-294.

USDA, ARS, National Genetic Resources Program. Germplasm Resources Information Network (GRIN). National Germplasm Resources Laboratory, Beltsville, Maryland. http://www.ars-grin.gov/cgi-bin/npgs/html/exsplist.pl. Accessed on August 20, 2017.