FARIA PN; CECON PR; SILVA AR; FINGER FL; SILVA FF; CRUZ CD; SÁVIO FL. 2012. Métodos de agrupamento em estudo de divergência genética de pimentas. *Horticultura Brasileira* 30: 428-432.

Métodos de agrupamento em estudo de divergência genética de pimentas

Priscila N Faria[1]; Paulo R Cecon[2]; Anderson R da Silva[2]; Fernando L Finger[2]; Fabyano F e Silva[2]; Cosme D Cruz[2]; Filipe L Sávio[1]

[1]USP-ESALQ, 13418-900 Piracicaba-SP; [2]UFV, 36570-000 Viçosa-MG; anderson.rodrigo@ufv.br

RESUMO

O objetivo deste trabalho foi comparar três métodos para determinação do número de grupos em estudos com aplicação de métodos hierárquicos de agrupamentos, baseando-se em dados obtidos a partir da caracterização de acessos de *Capsicum*, de modo a identificar aquele com maior poder de discriminação. Os métodos de Mojena, de Tocher e o método RMSSTD foram aplicados com a finalidade de determinar o número ideal de grupos formados na fase final do procedimento de agrupamento com o método UPGMA. Foram analisados 49 acessos da espécie *Capsicum chinense* do Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa, em relação a dez características morfológicas com o intuito de identificar e agrupar os acessos mais similares, tornando possível a seleção de genótipos superiores, ou seja, com as características comerciais de interesse. Os resultados mostraram que o método RMSSTD permitiu concluir sobre a existência de sete grupos, evidenciando um maior poder de discriminação para este método, em relação ao método de otimização de Tocher e ao método de Mojena, que formaram respectivamente, quatro e três grupos.

Palavras-chave: *Capsicum chinense,* método de Mojena, método de Tocher, RMSSTD.

ABSTRACT

Clustering methods in a study of genetic diversity of peppers

The objective of this study was to compare three methods to determine number of groups in studies with hierarquical cluster analysis, based at data from characterization of *Capsicum* accessions, in order to identify those with the greatest power of discrimination. Mojena method, Tocher method and RMSSTD method were applied to determine the optimal number of groups formed in final stage of the UPGMA method. Forty nine *Capsicum chinense* accessions from the Vegetable Germplasm Bank of the Universidade Federal de Viçosa were analyzed, in relation to ten morphological characters for identifying the most similar accessions group, making possible the selection of superior genotypes, i.e., of commercial interest. The RMSSTD method allowed to conclude on the existence of seven groups, with a greater power of discrimination for this method, compared to the Tocher optimization method and the Mojena method, which formed, respectively, four and three groups.

Keywords: *Capsicum chinense*, Mojena method, Tocher method, RMSSTD.

(Recebido para publicação em 20 de abril de 2011; aceito em 25 de junho de 2012)
(*Received on April 20, 2011; accepted on June 25, 2012*)

A diversidade genética pode ser avaliada de forma simultânea em relação às várias características, e para isso recomenda-se a utilização de medidas de dissimilaridade (Cruz & Carneiro, 2003). Uma forma prática e eficiente de se obter essas medidas é por meio da análise de agrupamentos (ou análise de *cluster*), a qual tem por finalidade reunir os indivíduos em grupos, de forma que exista a máxima homogeneidade dentro do grupo e a máxima heterogeneidade entre os grupos (Johnson & Wichern, 1992; Cruz & Regazzi, 2001). Dos métodos de agrupamento, os mais utilizados são os de otimização e os hierárquicos (Cruz *et al*., 2011).

Uma dificuldade na etapa final dos estudos que se utilizam de algoritmos hierárquicos de agrupamento é a falta de critérios objetivos para identificar o número ideal de grupos formados, uma vez que na prática este número é dado simplesmente por uma inspeção gráfica visual ou estabelecido em pontos de alta mudança de nível dos dendrogramas (Milligan, 1981). Milligan & Cooper (1985) apresentaram um estudo comparativo de 30 critérios de corte para determinação do número de agrupamentos, e utilizando dados artificiais com número conhecido de agrupamentos, mostraram que critérios diferentes podem conduzir a resultados muito diferentes.

Um dos métodos considerados como "objetivos", dentre os poucos existentes, é o de Mojena (1977). Este método recorre a um critério estatístico e propõe um procedimento de cálculo baseado no tamanho relativo dos níveis de fusões ou distâncias no dendrograma. Dessa forma, o método não necessita do conhecimento prévio da conformação dos grupos, ao contrário do método subjetivo, com base na inspeção visual das ramificações do dendrograma.

Vários trabalhos têm utilizado a análise de agrupamentos no estudo da diversidade genética e relatado a subjetividade no estabelecimento do ponto de corte no dendrograma. Dentre eles, Sudré *et al*. (2005) analisaram a divergência genética entre acessos de pimenta e pimentão utilizando técnicas multivariadas, obtendo a separação dos acessos em sete grupos distintos, utilizando pontos de mudança brusca de nível como referência para estabelecer o corte. Karasawa *et al*. (2005) relataram que a utilização do procedimento subjetivo, baseado no exame visual do dendrograma, pode gerar alguma dificuldade na tomada de decisão quanto ao número de grupos gerados, uma vez que qualquer inferência rígida sobre este número pode não ser produtiva.

Na tentativa de solucionar esse problema, Faria (2009) propôs que o Método da Máxima Curvatura Modificado (Lessman & Atkins, 1963) fosse utilizado juntamente com o índice RMSSTD (*Root Mean Square Standard Deviation*). A metodologia proposta é fundamentada na obtenção do ponto de máxima curvatura para a trajetória

do índice RMSSTD dada em função do número de grupos, onde o referido ponto é que determina o número ideal de grupos.

O objetivo deste trabalho foi comparar três métodos aplicados com a finalidade de determinar o número de grupos de agrupamentos hierarquicos, quais sejam: Método de Mojena, Método de otimização de Tocher e o método que utiliza o índice RMSSTD, a fim de se detectar aquele com maior poder em discriminar grupos homogêneos em relação a caracteres do fruto de pimentas *Capsicum chinense*.

MATERIAL E MÉTODOS

O Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa (BGH/UFV) possui cerca de 100 acessos de *C. chinense* com grande potencial para contribuir em programas de melhoramento, uma vez que grandes quantidades de informações a respeito de características agronômicas, químicas, morfológicas e moleculares já foram coletadas. Neste trabalho foram utilizados resultados de caracterização feita para 49 desses acessos por Faria (2009).

O experimento foi conduzido em área experimental da UFV no ano de 2009 sob delineamento em blocos casualizados com três repetições, utilizando espaçamento de 1x1 m, sendo cada linha constituída por três plantas de cada acesso. Os tratos culturais seguiram as recomendações feitas por Filgueira (2000). Foram utilizados dados de caracteres quantitativos de fruto, os quais foram colhidos no estádio maduro e imediatamente caracterizados de acordo com os descritores de *Capsicum* (IPGRI, 1995). Dez características foram avaliadas: comprimento do fruto (mm); maior diâmetro do fruto (mm); espessura do pericarpo (mm); porcentagem de matéria seca (g/100g); massa seca do fruto (g); massa fresca do fruto (g); capsaicina total (mg/gMS); teor de sólidos solúveis (ºBrix); Vitamina C (mg/100 g de fruto fresco); cor extraível (unidades ASTA de cor). A identificação dos referidos acessos pode ser verificada em Faria (2009).

Realizou-se análise de variância univariada para se verificar a existência de variabilidade genética significativa entre os 49 acessos, em relação às características avaliadas com base na média das parcelas.

Foram calculadas as médias de cada acesso para cada variável analisada. A matriz de correlações residuais entre as variáveis foi estimada por ocasião da realização de análise de variância univariada. Foi aplicado o teste de esfericidade de Bartlet (Rencher, 2002) para testar a hipótese de que esta é uma matriz diagonal, isto é, se as dez variáveis apresentadas são independentes. Tendo encontrado resultado não significativo ($p>0,05$) pela aproximação de qui-quadrado com 45 graus de liberdade, a distância generalizada de Mahalanobis coincidiu com a distância euclidiana como medida de dissimilaridade. Pode-se concluir então, que a utilização de quaisquer das distâncias supracitadas na análise de agrupamento realizada neste trabalho não altera significativamente os resultados obtidos. Desta forma, optou-se pelo uso da distância Euclidiana. A diversidade genética entre os acessos foi avaliada utilizando-se o método da ligação média entre grupos ou UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*).

Três métodos foram utilizados para determinação do número ótimo de grupos no dendrograma, a saber: o Método de Mojena, o Método de Tocher e o método que utiliza o índice RMSSTD.

O método de Mojena (1977) é um procedimento baseado no tamanho relativo dos níveis de fusões (distâncias) no dendrograma. Este método consiste em selecionar o número de grupos no estágio *j* que, primeiramente, satisfizer à seguinte inequação: $\alpha_j > \theta_k$, em que α_j é o valor de distâncias dos níveis de fusão correspondentes ao estágio *j* (*j* = 1, 2, ..., n) e θ_k é o valor referencial de corte, expresso por: $\theta_k = \overline{\alpha} + k\hat{\sigma}_\alpha$, em que α e $\hat{\sigma}_\alpha$ são, respectivamente, as estimativas não viesadas da média e do desvio padrão dos valores de α; k é uma constante. Adotou-se k= 1,25 como regra de parada na definição do número de grupos, como sugerem Milligan & Cooper (1985).

O método de otimização de Tocher é um método de agrupamento que se baseia na formação de grupos cujas distâncias dentro dos grupos sejam menores que as distâncias entre grupos. Ao final do processo obtém-se o número de grupos e os acessos contidos em cada grupo. A aplicação deste método foi feita conforme sugerem Cruz & Carneiro (2003).

A determinação do número de grupos pelo índice RMSSTD foi realizada por meio de análise gráfica (determinação geométrica) dos valores da raiz quadrada do desvio padrão médio (valores de RMSSTD). De acordo com Sharma (1996) o valor do índice deve ser tão baixo quanto possível para um agrupamento, isto é, quanto menor o RMSSTD, mais homogêneo ou compacto é o agrupamento formado a um determinado passo.

Juntamente com o índice RMSSTD, utilizou-se o Método da Máxima Curvatura Modificado (Lessman & Atkins, 1963), que se baseia na modificação do método de Smith (1938). Relacionaram-se então os valores de RMSSTD em função do número de grupos. Assim, o ponto que estabelece declínio acentuado do RMSSTD indica no eixo das abscissas o número ótimo de grupos. A metodologia pode ser expressa pela seguinte equação,

$$RMSSTD = \frac{a}{X^{b}}$$

em que X representa o número de *clusters* (grupos), a e b são constantes estimadas. A partir desta função de curvatura, determinou-se o valor da abscissa onde ocorre o ponto de máxima curvatura ou ponto crítico da função de curvatura, conforme apresentado por Meier & Lessman (1971), por meio da seguinte equação,

$$X_{MC} = \left[\frac{a^2 b^2 (2b+1)}{(b+2)}\right]^{\frac{1}{2b+2}}$$

em que X_{MC} é o número de grupos que representa a máxima curvatura em RMSSTD; a e b são constantes estimadas.

As análises foram realizadas utilizando-se o módulo 'Diversidade Genética' do programa GENES (Cruz, 2006),

para obtenção do número de grupos pelo Método de Mojena e pelo Método de Tocher; e SAS® (SAS, 2003), para o método de agrupamento UPGMA e para empregar a metodologia proposta por Faria (2009), que utiliza o método da máxima curvatura modificado e o índice RMSSTD.

RESULTADOS E DISCUSSÃO

Pela análise de variância ($p<0,01$), foi possível detectar diferenças significativas entre os acessos para todas as características avaliadas, indicando a presença de variabilidade genética. Os coeficientes de variação CV (%) do experimento variaram de 5,34% (capsaicina total) a 36,82% (cor extraível).

De acordo com a Tabela 1 é possível verificar que os acessos 5 e 28 foram os mais similares geneticamente, possuindo a menor distância (9,47); entre os acessos 1 e 27 houve a maior magnitude

Tabela 1. Distâncias dos níveis de fusão entre acessos de *C. chinense* obtidas pelo método UPGMA para estimação do valor referencial de corte (θ_k) no dendrograma por meio do Método de Mojena (1977), utilizando k= 1,25 (distance of fusion levels between *C. chinense* accessions by UPGMA method to estimate the cut reference value (θ_k) in dendrogram using the Mojena method (1977), with k= 1.25). Viçosa, UFV, 2009.

Estágio	Acesso x	Acesso y	Distância (α_j)	Distância (%)	θ_k
1	5	28	9,47	2,04	
2	2	12	11,48	2,47	12,25
3	9	10	11,90	2,56	12,57
4	1	48	12,56	2,71	13,02
5	23	34	12,73	2,74	13,26
6	17	49	14,90	3,21	14,39*
7	3	46	16,40	3,53	15,62*
8	11	15	16,71	3,60	16,43*
9	41	47	16,87	3,63	16,98
10	33	42	18,64	4,02	17,85*
11	21	23	18,91	4,08	18,53*
12	30	36	19,32	4,16	19,11*
13	4	43	19,57	4,22	19,59*
14	19	26	21,11	4,55	20,27*
15	8	9	21,37	4,61	20,84*
16	29	37	21,74	4,69	21,36*
17	1	4	22,96	4,95	21,98*
18	3	11	23,89	5,15	22,63*
19	21	31	25,73	5,55	23,46*
20	7	20	26,82	5,78	24,32*
21	5	32	27,04	5,83	25,06*
22	6	24	27,47	5,92	25,73*
23	19	39	28,30	6,10	26,42*
24	38	41	29,23	6,30	27,12*
25	16	33	29,61	6,38	27,76*
26	1	35	31,93	6,88	28,60*
27	18	38	33,15	7,15	29,47*
28	6	8	33,32	7,18	30,24*
29	17	30	33,62	7,25	30,95*
30	7	16	34,89	7,52	31,70*
31	5	19	37,60	8,11	32,65*
32	3	18	40,02	8,63	33,75*
33	6	21	40,12	8,65	34,72*
34	6	7	42,35	9,13	35,81*
35	1	17	45,03	9,71	37,04*
36	5	25	46,63	10,05	38,30*
37	2	40	48,18	10,39	39,56*
38	1	3	50,73	10,94	40,93*
39	14	29	52,06	11,23	42,28*
40	6	22	57,14	12,32	43,96*
41	2	5	66,01	14,23	46,35*
42	1	6	70,06	15,11	48,85*
43	2	44	76,05	16,40	51,66*
44	14	45	80,03	17,26	54,53*
45	13	14	94,34	20,35	58,51*
46	1	2	108,79	23,46	63,50*
47	1	13	210,88	45,48	81,51*
48	1	17	463,59	100,00	135,27*

* $\alpha_j > \theta_k$

Tabela 2. Agrupamento de 49 acessos de *C. chinense* do Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa pelo método de Tocher (grouping of 49 *C. chinense* accessions from Vegetable Germplasm Bank of the Universidade Federal de Viçosa by Tocher method). Viçosa, UFV, 2009.

Grupo	Acessos
I	3, 43, 47, 28, 5, 15, 48, 38, 11, 32, 23, 4, 7, 21, 13, 6, 1, 16, 26, 9, 10, 14, 17, 2, 12, 46, 37, 30, 19, 41, 22, 42, 34, 8, 31, 44, 39, 24, 36, 29, 49, 33, 25, 40, 20, 18
II	35
III	27
IV	45

Tabela 3. Agrupamento de 49 acessos de *C. chinense* do Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa pelo método RMSSTD juntamente com o da máxima curvatura modificado (grouping of 49 *C. chinense* accessions from Vegetable Germplasm Bank of the Universidade Federal de Viçosa by RMSSTD method with modified maximum curvature method). Viçosa, UFV, 2009.

Grupo	Acessos
I	1, 48, 4, 43, 35, 17, 49, 30, 36, 21, 23, 34, 31, 6, 24, 8, 9, 10, 7, 20, 16, 33, 42, 22
II	3, 46, 11, 15, 18, 38, 41, 47, 44, 5, 28, 32, 19, 26, 39, 25, 40
III	2, 12, 14
IV	45
V	13
VI	29, 37
VII	27

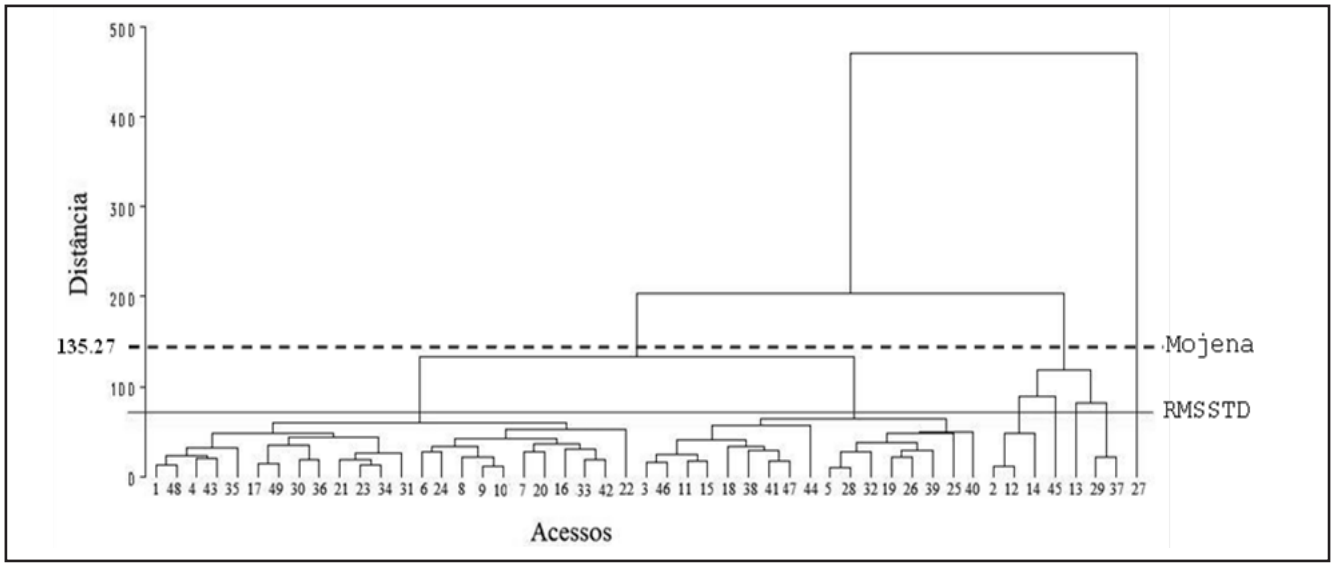

Figura 1. Dendrograma obtido pelo método UPGMA, a partir das medidas de dissimilaridade entre 49 acessos de *C. chinense*, baseado na distância Euclidiana (dendrogram obtained by UPGMA method from the dissimilarity between 49 *C. chinense* accessions, based on Euclidean distance). Viçosa, UFV, 2009.

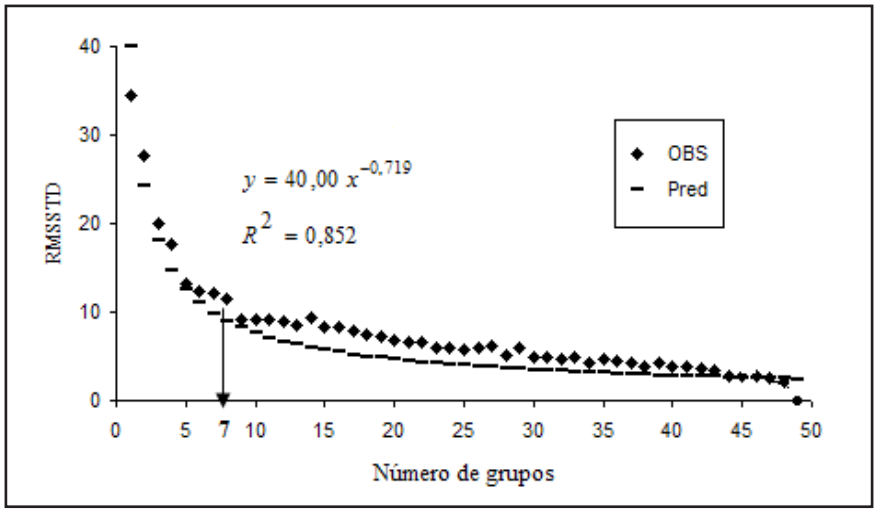

Figura 2. Regressão das estimativas de RMSSTD em função do número de grupos de acessos de pimenta e determinação do ponto de máxima curvatura geométrica (regression of the estimates of RMSSTD depending on the number of accessions groups and determination of the point of maximum geometric curvature). Viçosa, UFV, 2009.

(463,59), sendo portanto os acessos mais dissimilares. Observa-se ainda que é viável um corte no dendrograma na altura de $\theta = 135{,}27$ (critério geral: $\theta_k = \bar{\alpha} + 1{,}25\hat{\sigma}_\alpha$), indicando que o número ideal de grupos deve ser igual a três. Sendo assim, o primeiro grupo foi composto somente pelo acesso 27, o segundo grupo pelos acessos 2, 12, 14, 45, 13, 29, 37 e o terceiro grupo contendo o restante dos acessos (Figura 1).

Pelo método de Tocher, foi detectada a formação de quatro grupos distintos, sendo os grupos II, III e IV formados por apenas um acesso cada (Tabela 2). O grupo I foi composto por 46 acessos, aproximadamente 94%. Verifica-se, portanto, a dificuldade em analisar a divergência entre os acessos, visto que a maioria deles encontra-se em apenas um grupo. Em estudo sobre dissimilaridade genética entre 17 acessos de pimentas *Capsicum* spp., Neitzke *et al.* (2010) encontraram resultados semelhantes, quando verificaram a formação de quatro grupos distintos pelo método de Tocher para dados quantitativos. Os mesmos autores relatam ainda que o primeiro grupo também reuniu a grande maioria dos acessos (13 acessos, cerca de 76%), e que, quando o método foi aplicado a dados qualitativos, foi maior o número de grupos formados (oito grupos distintos).

A informação do índice RMSSTD permitiu a identificação do número ótimo de grupos por meio da máxima curvatura (Figura 2). De posse desta informação, os 49 acessos foram agrupados, e a utilização do índice RMSSTD, juntamente com o método da máxima curvatura modificado possibilitou a formação de sete grupos (Tabela 3). Posteriormente foi possível identificar um ponto de corte no dendrograma, obtido pela utilização do método UPGMA com base na distância Euclidiana (Figura 2). Observou-se que o maior grupo formado foi composto por 24 acessos e os grupos IV, V e VII formados apenas por um acesso cada: o acesso 45 pertencente ao grupo IV, o acesso 13 pertencente ao grupo V e o acesso 27 pertencente ao grupo VII (Tabela 3). O acesso 27 mostrou-se bastante divergente dos demais, visto que formou um grupo exclusivo e permaneceu isolado dos demais acessos no dendrograma. Sua divergência genética pode ser explorada em programas de melhoramento com base em descritores quantitativos para

obtenção de novas cultivares.

Utilizando os métodos de agrupamento de Tocher e UPGMA em estudo sobre diversidade genética entre 23 acessos de espécies cultivada de pimentas *Capsicum* spp., Monteiro *et al*. (2010) relataram que a espécie *C. chinense* foi a mais divergente em todos os agrupamentos, com as maiores distâncias intragrupo. Este fato está de acordo com os resultados obtidos no presente trabalho, evidenciando o elevado número de acessos num mesmo grupo. Nesse caso, tendo a espécie *C. chinense* apresentado grande divergência genética, é esperado que haja a formação de mais grupos e até mesmo com alguns dos grupos contendo muitos acessos, uma vez que apenas poucos destes acessos apresentam grande dissimilaridade. Assim, um método mais sensível para determinação do número de grupos consegue captar mais eficientemente a existência de acessos discrepantes.

A aplicação da metodologia do RMSSTD aos dados de pimenta permitiu concluir sobre a existência de sete grupos. Este número de grupos foi superior ao obtido pelo método de Tocher (quatro grupos) e ao método de Mojena (três grupos), sendo, portanto de maior poder de discriminação, possibilitando a identificação de mais grupos contendo acessos similares.

REFERÊNCIAS

CRUZ CD. 2006. *Programa Genes*: Biometria. Viçosa: Editora UFV. 382p.

CRUZ CD; CARNEIRO PCS. 2003. *Modelos biométricos aplicados ao melhoramento genético*. Volume 2. Viçosa: UFV. 585p.

CRUZ CD; FERREIRA FM; PESSONI LA. *Biometria aplicada ao estudo da diversidade genética*. Visconde do Rio Branco: Suprema, 2011. 620p.

CRUZ CD; REGAZZI AJ. 2001. *Modelos biométricos aplicados ao melhoramento genético.* 2 ed. Viçosa: UFV. 390p.

FARIA PN. 2009. *Avaliação de métodos para determinação do número ótimo de clusters em estudo de divergência genética entre acessos de pimenta*. Viçosa: UFV. 67p. (Dissertação mestrado).

FILGUEIRA FAR. 2000. *Novo Manual de Olericultura*. Viçosa: UFV. 402p.

IPGRI AVRDC; CATIE. 1995. *Descriptors for Capsicum* (*Capsicum* spp.). International Plant Genetic Resources Institute, Rome, Italy; the Asian Vegetable Research and Development Center, Taipei, Taiwan, and the Centro Agronómico Tropical de Investigación y Enseñanza, Turrialba, Costa Rica. 110p.

JOHNSON RA; WICHERN DW. 1992. *Applied Multivariate Statistical Analysis.* New Jersey: Englewood Cliffs. 642p.

KARASAWA M; RODRIGUES R; SUDRÉ CP; SILVA MP; RIVA EM; AMARAL JÚNIOR AT. 2005. Aplicação de métodos de agrupamento na quantificação da divergência genética entre acessos de tomateiro. *Horticultura Brasileira* 23: 1000-1005.

LESSMAN KJ; ATKINS RE. 1963. Optimum plot size and relative efficiency of lattice designs for grain sorghum yield tests. *Crop Science* 3: 477-481.

MEIER VD; LESSMAN RJ. 1971. Estimation of optimum field plot shape and size for testing yield in *Crambe abyssinica* Hochst. *Crop Science* 11: 648-645.

MILLIGAN GW. 1981. A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika* 46: 187-199.

MILLIGAN GW; COOPER MC. 1985. An examination of procedures for determining the number of clusters in a data set. *Psychometrika* 50: 159-179.

MOJENA R. 1977. Hierarchical grouping methods and stopping rules: an evaluation. *The Computer Journal* 20: 359-363.

MONTEIRO ER; BASTOS EM; LOPES ACA; GOMES RLF; NUNES JAR. 2010. Diversidade genética entre acessos de espécies cultivadas de pimentas. *Ciência Rural* 40: 288-283.

NEITZKE RS; BARBIERI RL; RODRIGUES WF; CORRÊA IV; CARVALHO FIF. 2010. Dissimilaridade genética entre acessos de pimenta com potencial ornamental. *Horticultura Brasileira* 28: 47-53.

RENCHER AC. 2002. *Methods of multivariate analysis*. New York: John Wiley. 708p.

SUDRÉ CP; RODRIGUES R; RIVA EM; KARASAWA M; AMARAL JÚNIOR AT. 2005. Divergência genética entre acessos de pimenta e pimentão utilizando técnicas multivariadas. *Horticultura Brasileira* 23: 22-27.

SAS INSTITUTE. 2003. *SAS System:* SAS/STAT version 9.1 (software). Cary.

SHARMA S. 1996. *Applied multivariate techniquies.* NewYork: John Wiley & Sons. 493p.

SMITH HF. 1938. An empirical law describing heterogeneity in the yields of agricultural crops. *Journal of Agricultural Science* 28: 1-23.