

Seleção de descritores morfológicos para a análise discriminante de *Capsicum* spp.

Tânia Nunes Rodrigues – UNIVERSIDADE FEDERAL DO AMAZONAS
Fabio Medeiros Ferreira – UNIVERSIDADE FEDERAL DO AMAZONAS
E-mail para contato: tanianunes.masterpiece@mail.com

Agência Financiadora: FAPEAM

Eixo Temático: 2.1.5 Ciências Agrárias

Categoria: Comunicação oral

1. Resumo

O objetivo do presente trabalho foi avaliar a eficiência da seleção de variáveis pelos procedimentos *forward*, *backward* e *stepwise* aplicados a descritores morfológicos de pimentas *Capsicum* spp., a fim de avaliar a diversidade genética e a classificação dos acessos em espécies e morfotipos. Foram analisados dados de 58 acessos de pimentas e pimentões do gênero *Capsicum* spp., da coleção de trabalho da Faculdade de Ciências Agrárias da Universidade Federal do Amazonas, previamente agrupados em morfotipos e espécies, com cerca de cinquenta descritores morfológicos, estabelecido pelo International Plant Genetic Resources Institute (IPGRI) para o gênero *Capsicum* spp. Para a seleção dos descritores morfológicos, além do critério de Akaike (AIC), foram utilizadas variantes e combinações de níveis de significância (α) de 5, 10, 15, 20 e 25%. Para a avaliar a diversidade genética e checar a classificação dos acessos foi empregada a análise discriminante não paramétrica dos k-vizinhos mais próximos ($k = 1$), usando como medida de dissimilaridade o complemento aritmético do coeficiente de coincidência simples. A partir de três subconjuntos de dados, pôde-se verificar que a seleção de variáveis ou uso daquelas com maior poder discriminante, no geral, proporcionam melhores classificações dos acessos em suas respectivas populações (pré-estabelecidas). Verificou-se uma diminuição significativa do número de variáveis utilizadas para as análises discriminantes, o método estatístico selecionou 14 de um total de cinquenta variáveis, o que equivale a uma redução expressiva de 72% do

número total de variáveis avaliadas, sendo que oito das 14 variáveis selecionadas coincidiram com aquelas indicadas pelo IPGRI como os descritores com maior poder discriminatório para o gênero *Capsicum spp.*, sendo estas: cor do caule, altura da planta, número de folhas por axila, cor do fruto maduro, forma do fruto, comprimento do fruto largura do fruto, apêndice na ponta do fruto, pungência do fruto. Na seleção de variáveis, os descritores cor do fruto e forma do fruto foram escolhidos para os três conjuntos de dados avaliados neste estudo, sendo as que mais se destacaram, portanto, tais descritores morfológicos são de suma importância na discriminação de acessos de pimenta.

Palavras-chave: k-vizinhos mais próximos, diversidade genética, pimentas.

2. Introdução

As pimentas do gênero *Capsicum spp.* têm um enorme potencial de melhoramento no sentido nutricional, pois dispõem de altos teores de vitaminas A e C (PESSOA *et al.*, 2015). Melhoristas têm tido a preocupação de selecionar cultivares que tenham maior produtividade e um fruto de melhor qualidade, em conformidade com a utilidade destinada para indústria ou para consumo *in natura* (RÊGO *et al.*, 2011a). Portanto, o melhorista durante o estágio de formação das populações base deve saber agrupar bem os acessos – ou genótipos – similares e discriminar bem aqueles que são diferentes, além de determinar a superioridade de desempenho e de comportamento de cada genótipo. (CRUZ, 2001).

Nos estudos genéticos de alocação e discriminação de genótipos em suas respectivas populações, o primeiro quesito a considerar é definir ou examinar se as populações avaliadas são diferentes o suficiente para que a investigação tenha significado. O segundo aspecto importante é saber se as variáveis mensuradas contém informações úteis ou se apenas algumas delas são suficientes para discriminar as populações (KHATTREE & NAIK, 2000). Existe muitos descritores botânicos, morfológicos e agrônômicos para o gênero *Capsicum spp.*

A eliminação daqueles considerados redundantes ou com menor capacidade discriminatória pode tornar-se uma estratégia mais adequada por permitir que a caracterização de maior número de acessos seja feita de modo mais rápido e sem comprometer a precisão da classificação dos genótipos e avaliação da diversidade. Sudré *et al.* (2006, p.729) avaliaram a divergência genética entre

acessos de sete espécies do gênero *Capsicum* spp., com base em 13 variáveis multicategóricas mais discriminantes, definidas pelo *International Plant Genetic Resources Institute* (IPGRI 1995). Segundo os autores as funções discriminantes de Anderson identificaram 100% das espécies e variedades botânicas estudadas. Em Costa *et al.* (2015) foram caracterizados 40 acessos de *Capsicum* spp. da coleção de germoplasma da Universidade Federal do Amazonas, com 56 descritores morfológicos, as três espécies que foram identificadas são: *C. chinense* com 35 acessos *C. baccatum* com um e *C. frutescens* com quatro acessos, também observou-se variabilidade fenotípica principalmente nas características e frutos que mostraram diferenças quanto ao tamanho, formato, coloração e pungência.

Métodos de seleção ou descarte de variáveis para o estudo da diversidade genética foram estabelecidos por meio da determinação da importância relativa de cada uma delas pelo critério proposto por Singh (1981, p.186), baseado na distância de Mahalanobis e o critério baseado na técnica de componentes principais Jolliffe (1972, p. 160; 1973, p.21), amplamente utilizados em dados fenotípicos contínuos – como componentes de produção e variáveis morfoagronômicas.

Outros procedimentos como a seleção *forward*, a eliminação *backward*, ou a combinação de ambos – seleção *stepwise* – (McKay & Campbell, 1982) também têm sido empregados na avaliação de marcadores moleculares mais discriminantes e informativos para estudos genéticos (Zhang *et al.*, 2005; Miano *et al.* 2008). Estes são processos de seleção de variáveis semelhantes aqueles usados em regressão múltipla, cuja critério de decisão baseia-se no teste F da análise multivariada de variância (MANOVA, do inglês *multivariate analysis of variance*). Considerando que dados de marcadores moleculares também são categóricos, sugere-se o uso da seleção automática *forward*, *backward* e *stepwise* também para descritores morfológicos qualitativos. Portanto, o objetivo do presente trabalho foi avaliar a eficiência dos procedimentos de seleção de variáveis *forward*, *backward* e *stepwise* para descritores morfológicos de pimentas do gênero *Capsicum* spp., na investigação da diversidade genética e classificação de acessos em espécies e morfotipos.

3. Material e Métodos

Foram analisados dados de 58 acessos de pimentas e pimentões do gênero *Capsicum* spp., da coleção de trabalho da Faculdade de Ciências Agrárias da Universidade Federal do Amazonas, previamente agrupados em morfotipos –

populações *a priori* – pimenta-de-cheiro, olho de peixe, murupi, curiabá, dedo-de-moça e pimenta de passarinho, cumari-do-Pará, malagueta, pimenta-de-mesa, pimentão, pimenta doce ou; agrupados em espécies – populações *a priori* – *C. chinense*, *C. frutescens*, *C. annuum* var. *glabriusculum* e *C. annuum*. Mais de cinquenta descritores morfológicos, estabelecido pelo International Plant Genetic Resources Institute (IPGRI) para o gênero *Capsicum* spp. (IPGRI 1995), com algumas adaptações propostas por Carvalho *et al.* (2003, p 203), foram considerados como variáveis qualitativas categóricas (multicategóricas ou binárias).

As variáveis contempladas foram referentes à plântula (fase de muda), planta, folha, flor, fruto e semente (IPGRI, 1995). Os conjuntos de dados avaliados neste trabalho foram coletados mediante dois experimentos desenvolvidos em épocas diferentes. Para as análises foram utilizados os seguintes subconjuntos de dados: 33 acessos e 50 variáveis, destas, 16 mais discriminantes conforme IPGRI, com duas espécies; ii) 38 acessos e 50 variáveis, destas 17 variáveis mais discriminantes, com seis morfotipos e; iii) com 56 acessos e 50 variáveis, destas 16 variáveis mais discriminantes, com duas espécies. Em cada subgrupo, as variáveis monomórficas foram retiradas da análise. Os subconjuntos i) e iii) foram representados por acessos dos dois conjuntos de dados unificados.

Para os três subconjuntos de dados supracitados, foi realizado o descarte de variáveis por meio de critério estatístico, a fim de comparar os resultados de classificação de genótipos e análise de diversidade quando usada todas as variáveis e quando usada aquelas mais discriminantes definidas pelo IPGRI. Para tal, foi utilizado para importância relativa dos descritores morfológicos o processo de seleção automática de variáveis, *stepwise*, *forward* e *backward* tendo como critérios para decisão de descarte ou inclusão, o índice de Akaike (AIC) e o teste F com variações e combinações dos níveis de significância iguais a 0,05, 0,10, 0,15, 0,20 e 0,25. As variáveis selecionadas foram aquelas mais coincidentes quando empregados os diferentes procedimentos e critérios adotados.

Após seleção das variáveis pelo critério estatístico, foram estimadas a dissimilaridade entre acessos, dando origem a uma matriz de dissimilaridade para cada subconjunto de dados, cujos valores foram obtidos pelo complemento aritmético da coincidência simples, definido a seguir:

$$d_{ii'} = \frac{D}{C+D}$$

em que, $d_{ii'}$ corresponde a dissimilaridade entre os acessos i e i' , para $i = i'$ ou todo $i \neq i'$; C corresponde a concordância entre os valores de categorias dos v descritores morfológicos estudados; D : discordância entre os valores de categorias dos v descritores morfológicos estudados.

Cada acesso possuía um único padrão fenotípico – obtido pelo valor da moda em relação a cinco plantas – para cada descritor morfológico considerado. Os valores dos coeficientes de dissimilaridade utilizados variaram de 0 (zero, para distância genética mínima) a 1 (um, para distância genética máxima).

Com base nos dados foram realizadas as classificações dos acessos em suas respectivas populações, por meio da análise discriminante dos k -vizinhos mais próximos, considerando $k = 1$. O vizinho mais próximo (k) foi aquele acesso que apresentou a menor dissimilaridade genética em relação ao acesso que se deseja classificar em uma das populações. Deste modo, após a estimação de todas as distâncias genéticas possíveis de um acesso i com os demais, à exceção dele com ele mesmo ($i = i'$), o vizinho mais próximo de cada acesso, serviu como regra decisória para (re) classificar qualquer acesso. Para verificar a eficiência da seleção de variáveis, estimou-se as porcentagens de classificação correta (P_c 's), como o complemento aritmético das taxas de erro aparente (TEA's), calculadas por:

$$P_c = 1 - \sum_{l=1}^L \pi_l \frac{m_l}{n_l} = 1 - \text{TEA},$$

Em que: π_l é o número de genótipos classificados erroneamente na l -ésima população, uma vez que estavam alocados previamente em outra; n corresponde ao número total de acessos do respectivo conjunto de dados; m_l corresponde ao número de classificações erradas de acessos pertencentes a população L . As análises foram executadas pelo software GENES, versão 1990.2018.18 (CRUZ, 2016), software ACTION STAT, versão 3.5.152.34 (*Equipe Estatcamp*, 2014).

4. Resultados e Discussão

O gênero *Capsicum spp.* se caracteriza por uma grande diversidade genética, o que é de suma importância para projetos de melhoramento genético, essa divergência é exposta em trabalhos, como o trabalho realizado por Souza *et al.*

(2015, p.667) em que foi investigado parâmetros referente ao crescimento de plantas e produção de frutos, os resultados encontrados pelos pesquisadores melhoram a utilidade de um importante recurso genético e atestam uma base genética complexa para distinções com base em descritores morfológicos relacionados ao fruto entre populações naturais.

Na avaliação do subconjunto formado por 33 acessos alocados em dois morfotipos: murupi e pimenta de cheiro, ao utilizar-se o total de 50 variáveis obteve-se porcentagem de classificação correta (P_c) dos acessos em suas respectivas populações igual a 42,43%, pelo método dos k-vizinhos mais próximos, indicando, no geral, má classificação global. Entretanto, o uso dos 16 descritores morfológicos com maior poder discriminatório, de acordo com o IPGRI, obteve-se uma significativa elevação de P_c para 85% (Tabela 1), reforçando a hipótese de que variáveis redundantes ou pouco discriminatórias podem atrapalhar a classificação e discriminação de acessos.

Contudo, pelo procedimento estatístico empregado na seleção de variáveis, em que foram selecionados sete descritores morfológicos, a quantidade de classificações corretas diminuiu drasticamente ($P_c = 30,31\%$), em que todas elas são variáveis de maior poder discriminatório, conforme a definição biológica dada pelo IPGRI, sendo elas: comprimento do fruto, cor da corola, cor do fruto maduro, forma do fruto, largura do fruto, pescoço na base do fruto e pungescencia do fruto (Tabela 1).

Para o subconjunto com 38 acessos alocados em seis morfotipos (curiabá, dedo-de-moça, murupi, olho de peixe, pimenta de cheiro e pimenta passarinho), todos pertencentes a espécie *Capsicum chinense*, as variáveis selecionadas foram: altura da planta, dia do florescimento, cor da corola, cor do fruto maduro, forma do fruto e peso do fruto, todas dentre aquelas de maior poder discriminatório. Neste conjunto de dados o valor de $P_c = 57,90\%$ também foi alcançado quando foram usadas somente as variáveis mais discriminantes considerando o critério biológico (IPGRI, 1995), superando os valores de P_c quando se fez uso de todas as variáveis ou promoveu-se seleção estatística (Tabela 1). A escolha das variáveis mais coincidentes dentro dos cenários testados nos procedimentos de seleção pode ter deixado para trás outras variáveis importantes na discriminação e classificação dos acessos em suas respectivas populações.

Tabela 1. Descritores morfológicos selecionados pelo critério estatístico a partir dos procedimentos *forward*, *backward* e *stepwise* para avaliação dos agrupamentos e classificação de acessos em suas respectivas populações e; porcentagens de classificação correta (Pc's) considerando todos os descritores de cada subconjunto de acessos[@], ou apenas aqueles com maior poder discriminatório conforme definido pelo IPGRI[&], ou aqueles selecionadas pelo critério estatístico[#].

Conjuntos de acessos		Variáveis selecionadas#	Taxa de acerto	
N° acessos	N° de populações	Nomes das variáveis #	N° de variáveis	%
33	Dois morfotipos	comprimento do fruto*, cor da corola*, pescoço na base do fruto*, pungescencia do fruto*,	50 @	43,33
		cor do fruto maduro*, forma do fruto*, largura do fruto*.	7#	30,31
		altura da planta*, dia do florescimento*,	51 @	44,74
38	Seis morfotipos	cor da corola*, cor do fruto maduro*, forma do fruto* e peso do fruto*	17*	57,90
		cor do caule*, pubescência do caule, densidade da folha,	50 @	82,15
56	Duas espécies	altura da planta*, cor da folha cotiledonar, pubescência da folha,	16*	92,86
		cor do filamento, número de folhas por axila*, cor do fruto maduro*,	14#	96,57
		forma do fruto*, comprimento do fruto*, largura do fruto*,		
		apêndice na ponta do fruto, pungência do fruto*.		

*Os descritores mais discriminantes, definidos pelo IPGRI (1995).

O subconjunto com 56 acessos, separados em duas espécies, *C. annum* com cinco acessos e *C. chinense* para os demais, foi aquele que apresentou as melhores taxas de classificações corretas, cujos valores se elevaram de 82,15% para 96,57%, com a diminuição do número de variáveis (Tabela 1). Pinto *et al.* (2010) realizaram um trabalho tendo como objetivo a seleção e determinação da importância de descritores morfológicos vegetativos para a caracterização de acessos de guariroba, em que foram descartados 55% dos descritores sem ocasionar perda de informação, por haver correlação significativa entre os descritores selecionados.

Das 50 variáveis utilizadas neste último conjunto de dados, 14 foram selecionadas, dentre as quais, oito foram consideradas de maior poder

discriminatório biológico (Tabela 1). Portanto, para este conjunto de dados, outras seis variáveis não definidas pelo IPGRI como mais discriminantes foram importantes na classificação dos acessos nas duas espécies *C. annum* e *C. chinense*. As variáveis selecionadas pelo método estatístico que coincidiram nos três subconjuntos de dados foram a cor do fruto maduro e a forma do fruto, estas variáveis fazem parte do grupo de variáveis com maior poder discriminatório de acordo com o IPGRI considerando o critério biológico. Mais simulações com diferentes números de variáveis, acessos, populações e grau de associação entre elas devem ser realizadas, a fim de se investigar a existência de padrões quanto ao uso ou descarte de variáveis com maior capacidade discriminatória em estudos de diversidade genética.

5. Conclusão

No geral, a seleção de variáveis com os procedimentos *forward*, *backward* e *stepwise* apresentaram porcentagens de classificação correta dos acessos em suas respectivas populações mais baixas ou próximas aquelas porcentagens de acertos quando fez-se uso de variáveis com maior poder discriminatório. Contudo, houve uma tendência de que a seleção de variáveis ou uso daquelas com maior poder discriminatório permitisse agrupar e classificar melhor os acessos em suas populações.

Na seleção de variáveis, cor do fruto e forma do fruto foram escolhidas para os três conjuntos de dados avaliados neste estudo, destacando, assim, a importância delas na discriminação de acessos de pimenta.

6. Referências

CARVALHO SIC; BIANCHETTI LB; BUSTAMANTE PG; SILVA DB. **Catálogo de germoplasma de pimentas e pimentões (*Capsicum* spp.) da Embrapa Hortaliças**. Brasília: Embrapa Hortaliças, (Embrapa Hortaliças, Documentos 49). 2003.

COSTA, L. V. et al. Caracterização de acessos de pimentas do Amazonas. **Horticultura Brasileira**, v. 33, n. 3, p. 290–298, 2015.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. Visconde do Rio Branco, MG: Suprema. 620 p. 2011.

CRUZ, C. D. Genes Software – extended and integrated with the R, Matlab and Selegen. **Revista Acta Scientiarum Agronomy**, Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec. 2016.

EQUIPE ESTATCAMP (2014). **Software Action**. Estatcamp- Consultoria em estatística e qualidade, São Carlos - SP, Brasil. URL <http://www.portaaction.com.br/> IPGRI. **Descritores para Capsicum (Capsicum spp.)**. IPGRI. Roma 51 p.1995.

JOLLIFFE. I.T. Discarding variables in a principal components analysis. I: Artificial data. **Applied statistics**. v. 21, p.160-173, 1972.

JOLLIFFE, I. T. Discarding variables in a principal component analysis. II: Real data. **Applied statistics**, p.21-31, 1973.

KHATTREE, R.; NAIK, D. N. **Applied multivariate statistical with SAS Software**.

2. ed., a co publication of Cary, NC: SAS Institute Inc. and New York, John Wiley & Sons, 338p. 2000.

MCKAY, R. J. & CAMPBELL, N. A. Variable selection techniques in discriminant analysis. I. Description. **British Journal of Mathematical and Statistical Psychology**, v.35, p. 1-29, 1982.

MIANO, D. W., LABONTE, D. R., CLARK, C.A. Identification of molecular markers associated with sweet potato resistance to sweet potato virus disease in Kenya **Euphytica**, v160: p15–24, 2008.

PESSOA, A. M. S.; RÊGO, E. R.; BARROSO, P. A.; RÊGO, M. M. Genetic diversity and importance of morpho-agronomic traits in a segregating F2 population of ornamental pepper. **Acta Horticulturae**, v. 1087, p.195-200, 2015.

RÊGO, E. R.; FINGER, F. L.; NASCIMENTO, N. F. F.; ARAÚJO, E. R.; SAPUCAY, M. J.L. C. Genética e melhoramento de pimentas. In: RÊGO, E. R.; FINGER, F. L.; RÊGO, M. M. (Orgs.). **Produção, genética e melhoramento de pimentas (Capsicum spp.)**. 1. ed. Recife: Imprima, p.117-136. 2011a.

SINGH, Y. P.; KUMAR, A.; CHAUHAN, B. P. S. Genetic divergence in pearl millet. **The Indian Journal of Genetic and Plant Breeding**, 41:1, 186-190, 1981.

SOUZA, M. S. **Análises discriminantes não paramétricas aplicadas ao estudo da diversidade genética baseado em dados fenotípicos quantitativos**. Manaus, UFAM. Dissertação mestrado. 78 p. 2017.

SOUZA, L. R. de; SCOSSA, F.; CHAVES, I. S.; KLEESSEN, S.; SALVADOR, L. F. D.; MILAGRE, J. C.; FINGER, F.L.; BHERING, L. L.; SULPICE, R.; ARAÚJO, W. L.;

NIKOLOSKI, Z.; FERNIE, A. R.; NESI, A. N. **Exploring natural variation of photosynthetic, primary metabolism and growth parameters in a large panel of *Capsicum chinense* accessions**. *Planta*, Heidelberg, v. 242, n.3, p.677-691, 2015.

SUDRÉ CP; CRUZ CD; RODRIGUES R; RIVA EM; AMARAL JÚNIOR AT; SILVA DJH; ZHANG, N.; XU, Y; AKASH, M.; MCCOUCH, S.; OARD, J. H. Identification of candidate markers associated with agronomic traits in rice using discriminant analysis, **Theor Appl Genet** 110: 721–729, 2005.