Consistência de agrupamentos de acessos de alho via análise discriminante

Anderson Rodrigo da Silva[1]
Paulo Roberto Cecon[1]
Mário Puiatti[2]
Moysés Nascimento[1]
Fabyano Fonseca e Silva[1]

1 Introdução

A escolha de um particular algoritmo de agrupamento exige o conhecimento de suas propriedades, aliado aos objetivos da pesquisa (Bussab et al., 1990). Na literatura estão disponíveis muitos algoritmos para se realizar a análise de agrupamento, os quais se distinguem pelo tipo de resultado e pelas diferentes formas de definir a proximidade entre as entidades. No entanto, a abrangência dos estudos, de informações, de métodos e de material biológico tem levado a certa dificuldade em escolher e aplicar corretamente as metodologias disponíveis e interpretar, convenientemente, o significado dos resultados das análises biométricas (Cruz et al., 2011).

Estudo sobre comparações entre métodos de agrupamento de otimização e hierárquicos foi realizado por Cargnelutti Filho et al. 2008, em divergência genética em feijão. Utilizando dados reais de cultivares de milho, Cargnelutti Filho & Guadagnin (2011) avaliaram a consistência do padrão de agrupamento de quatro métodos hierárquicos por meio do coeficiente de correlação cofenética. Também se baseando na correlação cofenética, Cargnelutti Filho et al. (2010) realizaram estudo sobre a consistência do padrão de agrupamento de cultivares de feijão a partir da combinação de oito medidas de dissimilaridade e oito métodos de agrupamento. Não foram encontradas na literatura, comparações entre os resultados da análise discriminante de Fisher a partir de agrupamentos de acessos de alho via métodos de otimização de Tocher e Tocher modificado (Vasconcelos et al., 2007), e hierárquicos de Ward e ligação média, com base em dados reais.

Este trabalho teve o objetivo de avaliar, quanto a consistência do agrupamento de acessos de alho, dois métodos hierárquicos (ligação média e Ward) e dois métodos de otimização (Tocher, Tocher modificado), por meio da análise discriminante de Fisher.

[1] DET – UFV. e-mail: anderson.agro@hotmail.com
[2] DFT – UFV.
Agradecimentos à FAPEMIG e a CAPES pelo apoio financeiro.

2 Material e métodos

Foram utilizados dados de diâmetro do bulbo, em mm, comprimento do bulbo, em mm, peso médio do bulbo, em g, número de bulbilhos por bulbo e produtividade, em $t.ha^{-1}$, em experimento com 89 acessos de alho registrados no Banco de Germoplasmas de Hortaliças da Universidade Federal de Viçosa (BGH/UFV). O experimento foi realizado em área experimental pertencente ao setor de olericultura do Departamento de Fitotecnia da UFV, município de Viçosa, Zona da Mata de Minas Gerais. O delineamento experimental foi o de blocos completos casualizados com quatro repetições.

As variâncias e covariâncias residuais entre as variáveis foram estimadas por ocasião da realização de análise de variância univariada. Foram calculadas as médias de cada acesso para cada variável analisada. Com base nessas estimativas, a distância generalizada de Mahalanobis foi adotada como medida de dissimilaridade entre os acessos para realizar os agrupamentos com os seguintes métodos: Tocher, Tocher modificado, algoritmo de Ward e ligação média (UPGMA). O número de grupos, no caso dos métodos hierárquicos, foi determinado pela aplicação do método de Mojena (1977), baseado no tamanho relativo dos níveis de fusão do dendrograma.

Para avaliar a consistência dos agrupamentos foi aplicada a análise discriminante de Fisher, a partir de combinações lineares das variáveis originais, às partições obtidas com cada método de agrupamento. Considerou-se como mais consistente o método de agrupamento que apresentou menor taxa de erro aparente (TEA) da análise discriminante, isto é, menor número total (todos os grupos) de classificações incorretas. As análises estatísticas foram realizadas utilizando-se do software R (R Development Core Team, 2011) e GENES versão 2009.7.9 (Cruz, 2006).

3 Resultados e discussões

Pelo método de agrupamento de Tocher baseado na distância generalizada de Mahalanobis os 89 acessos de alho foram divididos em 15 grupos, onde 7 destes apresentaram apenas 1 acesso cada. Pelo resultado da análise discriminante de Fisher para os grupos obtidos com o agrupamento de Tocher foram detectadas 24 classificações erradas, ou seja, a taxa de erro aparente foi de 26,96%, indicando baixa consistência do agrupamento (Tabela 1).

Tabela 1. Resumo dos resultados obtidos com os métodos de agrupamento aplicados a dados de 89 acessos de alho.

Método de agrupamento	N	CCC	TEA (%)
Tocher	15	-	26,96
Tocher modificado	7	-	19,10
Ligação média	6	0,76	6,74
Ward	5	0,56	4,49

N: número de grupos; CCC: coeficiente de correlação cofenética; TEA: taxa de erro aparente da análise discriminante.

De acordo com o método de Tocher modificado os 89 acessos foram separados em 7 grupos. Pelo resultado da análise discriminante de Fisher foram detectadas 17 classificações incorretas, ou seja, 19,10% foi a taxa de erro aparente. Embora a taxa de erro tenha reduzido em sete pontos percentuais em relação ao método original de Tocher, verifica-se que ainda é baixa a consistência do agrupamento (Tabela 1).

Utilizando o método de Mojena (1977) foi possível identificar um ponto de corte no dendrograma que corresponde a 28,2% da distância máxima observada nos níveis de fusão obtidos com o método da ligação média, com a formação de 6 grupos de acessos. A análise discriminante permitiu identificar seis classificações incorretas, resultando numa taxa de erro aparente igual a 6,74%. Em comparação com os métodos de otimização, observa-se que o método da ligação média apresentou redução significativa da taxa de erro aparente, com diferença de 20,22% em relação ao método de Tocher e 12,36% em relação ao método de Tocher modificado, apesar de a conformação dos grupos ter sido coincidente com aquela obtida com o método de Tocher modificado (Tabela 1).

Com o algoritmo de Ward foram obtidos 5 grupos de acessos, de acordo com o ponto de corte do método de Mojena (1977) na distância 33,91, que corresponde a 16,8% da distância máxima do dendrograma (Figura 1). É interessante destacar que com o algoritmo de Ward nenhum grupo foi constituído por apenas 1 acesso, como ocorreu com os métodos de otimização e com o método da ligação média. Isso ocorre devido ao fato de o método de Ward tender a formar grupos com o mesmo número de indivíduos, tendo como base os princípios da análise de variância (Mingoti, 2005). A taxa de erro aparente (4,49%) foi a menor dentre todas obtidas com os métodos de agrupamento neste estudo (Tabela 1). À respeito, Kuiper & Fisher (1975) concluíram que o método de Ward agrupa tão bem quanto a análise discriminante de Fisher quando existe igual número de indivíduos nos grupos e quando os dados seguem distribuição normal multivariada.

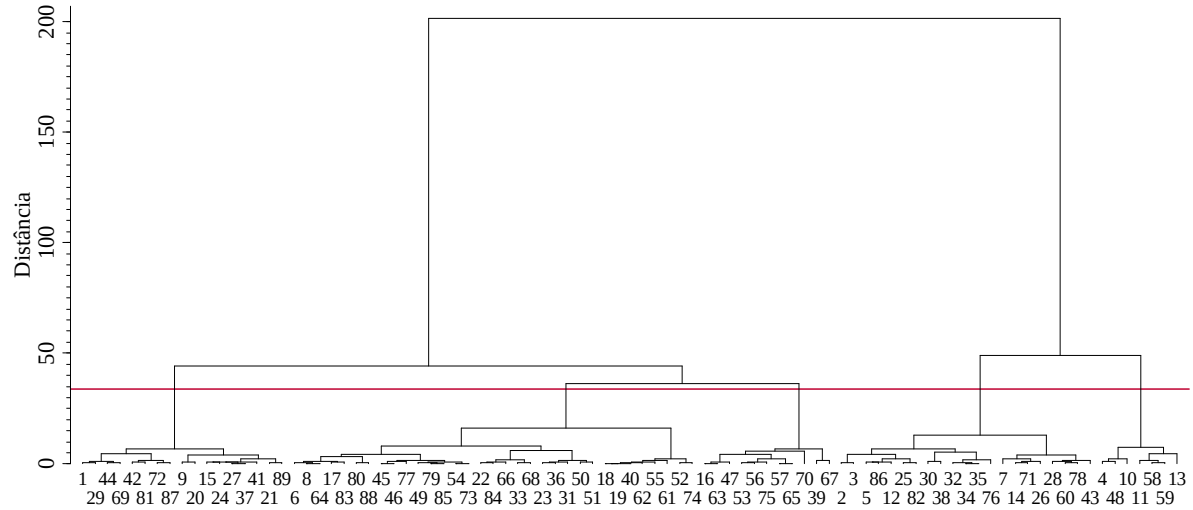

Figura 1. Dendrograma obtido com o algoritmo de Ward com base na distância generalizada de Mahalanobis entre acessos de alho.

Ainda na tabela 1 é possível observar que a correlação cofenética obtida com o método da ligação média (0,76) foi superior a do algoritmo de Ward (0,56). Isso pode ser explicado pelo fato de a medida utilizada no método de Ward depender do quadrado da distância euclidiana, fazendo com que os valores da matriz cofenética sejam maiores que os observados no método da ligação média, embora Rohlf (1982) alerte para o fato de que uma correlação cofenética próxima de 0,9 não garante que o dendrograma tenha sido capaz de sintetizar a relação fenética dos indivíduos.

De acordo com o estudo de consistência do padrão de agrupamento via coeficiente de correlação cofenética realizado por Cargnegutti Filho & Guadagnin (2011) com dados reais de milho, independente do número de cultivares, do número de variáveis e da medida de dissimilaridade, a consistência dos métodos aumenta na seguinte ordem: Ward, ligação completa, ligação simples e ligação média entre grupo. Estes resultados são concordantes com os encontrados no presente estudo, se for considerado apenas a correlação cofenética. Entretanto, pelo critério da taxa de erro aparente da análise discriminante verifica-se que o método de Ward é mais consistente que o método da ligação média.

É possível observar que a TEA média dos métodos hierárquicos é cerca de 17% inferior àquela observada com os métodos de otimização. Notar ainda que a TEA está diretamente ligada ao número de grupos obtidos no agrupamento. No entanto, estudos mais aprofundados devem ser realizados para que se conheça o comportamento desta variável (TEA), em cada método de agrupamento.

4 Conclusões

A análise discriminante de Fisher aplicada aos grupos dos métodos hierárquicos permitiu observar as menores taxas de erro aparente, sendo, portanto, os métodos mais consistentes para este estudo de divergência genética de acessos de alho.

A consistência do agrupamento aumenta na seguinte ordem dos métodos: Tocher, Tocher modificado, ligação média e Ward.

5 Referências

[1] BUSSAB, W. O.; MIAZAKI, E. S.; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 105 p.

[2] CRUZ, C. D. **Programa Genes**: Biometria. Viçosa: Editora UFV. 382 p. 2006

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

[4] CARGNELUTTI FILHO, A.; RIBEIRO, N. D.; REIS, R. C. P.; SOUZA, J. R.; JOST, E. Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. **Ciência Rural**. v. 38, n. , p. 2138-2145, 2008.

[5] CARGNELUTTI FILHO, A.; RIBEIRO, N. D.; BURIN, C. Consistência do padrão de agrupamento de cultivares de feijão conforme medidas de dissimilaridade e métodos de agrupamento. **Pesquisa Agropecuária Brasileira**. v. 45, n. 3, p. 236-243, 2010.

[6] CARGNELUTTI FILHO, A.; GUADAGNIN, J. P. Consistência do padrão de agrupamento de cultivares de milho. **Ciência Rural**. v. 41, n. 9, p. 1503-1508, 2011.

[7] KUIPER, F. K.; FISHER, L. A. Monte Carlo comparison of six clustering procedures. **Biometrics**. v. 31, n. 3, p.777-783, 1975.

[8] MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2005. 297p.

[9] MOJENA, R. Hierárquical grouping method and stopping rules: an evaluation. **Computer Journal**. v. 20, n. 4, p. 359-363, 1977.

[10] R DEVELOPMENT CORE TEAM (2011). **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

[11] ROHLF, F. J. Consensus índices for comparing classifications. **Mathematical Bioscience**. v. 59, n. 1, p. 131-144, 1982.

[12] VASCONCELOS, E. S.; CRUZ, C. D.; BHERING, L. L.; RESENDE JÚNIOR, M. F. R. Método alternativo para análise de agrupamento. **Pesquisa Agropecuária Brasileira.** v. 42, n. 10, p. 1421-1428, 2007.