

## Aplicação da análise multivariada em avaliações de divergência genética através de marcadores moleculares dominantes em plantas medicinais

CARVALHO, M.F.<sup>1\*</sup>; ALBUQUERQUE JUNIOR, C.L.<sup>1</sup>; GUIDOLIN, A.F.<sup>2</sup>; FARIAS, F.L.<sup>3</sup>

<sup>1</sup> Universidade do Sul de Santa Catarina - UNISUL, Avenida José Acácio Moreira, nº 787, Bairro Dehon - Caixa Postal: 370, CEP 88704-900 - Tubarão - SC. \*marcio.carvalho@unisul.br <sup>2</sup> Universidade do Estado de Santa Catarina, Av. Luis de Camões nº 2090, CEP 88520-000 - Lages - SC.

**RESUMO:** A estatística multivariada tem sido utilizada em estudos de divergência genética dentro de espécies vegetais medicinais. A análise da distância genética entre indivíduos é ferramenta importante no estudo de populações e para as estratégias de melhoramento da espécie. Este trabalho tem por objetivo apresentar as principais metodologias estatísticas para estudos de divergência genética quando utilizados marcadores moleculares dominantes. O agrupamento dos indivíduos segundo as distâncias genéticas calculadas resulta em dendrograma de dissimilaridade. Diferentes coeficientes de similaridade e metodologias de agrupamento apresentam características próprias, adequam-se às diferentes espécies e aos tipos de dados estudados. A utilização inadequada dos coeficientes de similaridade pode sugerir conclusões equivocadas em trabalhos de pesquisa.

**Palavras-chaves:** coeficientes de similaridade, análise de agrupamento, RAPD

**ABSTRACT:** Application of multivariate analysis for genetic divergence evaluations using dominant molecular markers in medicinal plants. The multivariate statistic has been used in genetic divergence studies concerning medicinal plants. Analysis of the genetic distance among individuals is an important tool for population studies and breeding strategies. This work aims to present the main statistical methodologies for genetic divergence studies when dominant molecular markers are used. The clustering of individuals using their calculated genetic distances results in a dissimilarity dendrogram. Different similarity coefficients and clustering methodologies present particular characteristics, adapting to different species and studied data. The inadequate utilization of similarity coefficients can lead to wrong conclusions in research works.

**Key words:** similarity coefficients, clustering analysis, RAPD

## INTRODUÇÃO

A preservação de recursos genéticos de plantas cultivadas em bancos de germoplasma (BAG) tem sido enfatizada no mundo inteiro, com o objetivo de preservar uma amostra da extensa diversidade genética de determinada espécie, preferivelmente com a classificação dos genótipos, das características e da organização genética, além de manter os acessos na sua constituição genética original, diminuindo a possibilidade de mutação, seleção ou contaminação (Karp et al., 1997). A agricultura moderna, o aumento da área urbana e a devastação de vegetações naturais têm sido os motivos principais para a criação dos BAG's em todo o mundo (Hidalgo, 1991). Atualmente, a preocupação com a perda da variabilidade e a extinção de espécies com potencial medicinal tem

gerado demanda por trabalhos de coleta, preservação e estudo sobre tais plantas. Para conservação e manejo eficiente destas espécies medicinais, a composição genética dos acessos coletados em diferentes regiões deve ser avaliada.

A investigação da variabilidade genética dos acessos de um BAG pode ser feita com o uso de técnicas de marcadores moleculares, principalmente por marcadores RAPD (Random Amplified Polymorphic DNA), AFLP (Amplified Fragment Length Polymorphism), RFLP (Restriction Fragment Length Polymorphism) e VNTR (Variable Number Tandem Repeats) (Ferreira & Grattapaglia, 1998). Várias espécies de plantas medicinais têm sofrido este tipo de avaliação, como *Maytenus ilicifolia* (Bittencourt,

2001; Mossi et al., 2007), *Chenopodium ambrosioides* (Santos & Corrêa, 2006), *Scutellaria* (Hosokawa et al., 2000), *Prunus africana* (Dawson & Powell, 1999), *Echinacea* (Nieri et al., 2003), *Epimedium* (Nakai et al., 1996), *Phyllanthus amarus* (Jain et al., 2003; Palaniappan & Marappa, 2008), *Casearia* (Silva et al., 2006), Jaborandi (Sandhu et al., 2006), *Salvia* (Lerin et al., 2007a) e *Oroxylum indicum* (Jayaram & Prasad, 2008).

Os marcadores moleculares RAPD e AFLP são denominados marcadores dominantes, devido ao polimorfismo genético detectado por estes serem de natureza binária, ou seja, detectam a presença ou ausência do segmento amplificado (Ferreira & Grattapaglia, 1998). Desta forma, estes marcadores não distinguem indivíduo homozigoto dominante de heterozigoto para um determinado locus.

Para ser realizado o estudo de divergência genética entre os acessos, ao final das análises genéticas, utiliza-se de análises estatísticas. Em primeira instância, determina-se a matriz de similaridade entre os genótipos para, posteriormente, ser aplicado uma metodologia estatística de análise de divergência. A análise de agrupamento, análise fatorial, análise discriminante e a análise do componente principal são métodos estatísticos que podem ser aplicados nos estudos de divergência genética. Entre estes métodos, a análise de agrupamento destaca-se por ser de fácil interpretação e por não necessitar de uma hipótese prévia da distribuição provável dos dados (Meyer et al., 2004).

Os métodos de classificação e ordenação da variabilidade genética entre os acessos de BAG's e das variedades lançadas pelos programas de melhoramento que estão se incrementando cada vez mais, estão ganhando cada vez mais significância. Neste âmbito, a análise multivariada apresenta-se como uma importante estratégia para classificação de germoplasma, ordenação da variabilidade de um amplo número de acessos, ou ainda para analisar as relações genéticas entre materiais melhorados (Mohammadi et al., 2003).

A estatística multivariada permite a análise simultânea das várias medidas de um determinado indivíduo que esteja sendo investigado (Hair et al., 1987), ou seja, tendo-se duas ou mais variáveis respostas (Y) de um indivíduo, aplica-se a análise multivariada. A crescente utilização desta metodologia de análise deve-se principalmente aos avanços da computação, devido à disponibilidade de vários programas aplicados a este fim, como GENES (Cruz, 2001), NTSYS (Rohlf, 1992), SAS (Sas Institute, 1992), entre outros.

O objetivo deste trabalho é demonstrar como a análise multivariada pode ser aplicada ao estudo de diversidade genética entre acessos de bancos de germoplasma ou variedades diferentes de uma mesma

espécie, utilizando-se marcadores moleculares dominantes. Para tanto, serão utilizados dados simulados do estudo de divergência genética de acessos de fáfia (*Pfaffia paniculata*) coletados no sul do Brasil.

### Simulação e análise dos dados

Utilizando-se os dados obtidos através de trabalhos com marcadores moleculares RAPD com os acessos de fáfia (denominados Baf e um número), far-se-á uma demonstração de como se procede à análise estatística destes. Supondo um estudo de sete acessos de um BAG, utilizando-se vinte marcadores RAPD. Os dados das leituras dos géis estão na Tabela 1.

A análise de agrupamento classifica  $n$  itens (podendo ser populações, indivíduos, variedades, linhas, etc.), utilizando as avaliações de  $p$  caracteres ou variáveis, a partir de medida de distância entre os itens. Esta técnica de agrupamento compreende duas etapas: na primeira, gera-se uma matriz de distâncias entre cada item, resultando em  $n(n-1)/2$  pares de distâncias. Num segundo momento, aplica-se um algoritmo de agrupamento sobre esta matriz gerada, para classificar e conectar os grupos homogêneos (Dias, 1998). O agrupamento final deve exibir alta homogeneidade dentro dos grupos e alta heterogeneidade entre os grupos (Mohammadi et al., 2003). Os resultados finais serão dados em números absolutos de distância genética e em dendrogramas de similaridade, para análise gráfica dos resultados (Ferreira & Grattapaglia, 1998).

A análise genética entre  $n$  itens, com o uso de marcadores moleculares dominantes como RAPD e o AFLP, são feitas através da leitura dos géis de eletroforese com as reações do marcador em questão. Esta leitura é feita dando-se notas para a presença ou ausência (1 e 0, por exemplo) das bandas nos géis (Dias, 1998). Estes dados, denominados dados binários, serão analisados utilizando-se coeficientes de similaridade, dois a dois, para o cálculo das distâncias genéticas. Os coeficientes de similaridade são específicos para variáveis binárias, e são utilizados em estudos de divergência genética baseados em marcadores moleculares dominantes (Meyer et al., 2004). Estes coeficientes representam a relação linear entre dois itens avaliados por um conjunto de  $p$  caracteres, sendo estes as respostas das leituras dos géis. Tendo-se as distâncias calculadas, aplicam-se os métodos de agrupamento dos itens estudados. Estes visam maximizar a similaridade dentro dos grupos e a dissimilaridade entre os grupos (Dias, 1998).

A leitura dos géis em dados binários (0 para ausência e 1 para presença da banda) é transformada em índices (a, b, c e d), quando comparamos dois itens. Supondo uma situação, onde se têm duas

populações, P e Q. O índice “a” representa a presença de uma mesma banda nas duas populações; o “b” representa a presença da banda em apenas uma das populações; o índice “c” indica a presença da banda somente na outra população que não em “b”; e “d” representa a ausência de uma mesma banda nas duas populações. Estes índices serão utilizados para

o cálculo dos coeficientes de similaridade. Existem vários coeficientes de similaridade propostos por diferentes autores (Tabela 2).

Os valores de distância genética entre dois genótipos (dissimilaridade -  $d_{ij}$ ) são calculados, geralmente, como complemento da similaridade ( $s_{ij}$ ):  $d_{ij} = 1 - s_{ij}$ . Outras funções para conversão de

**TABELA 1.** Leitura das bandas de marcadores RAPD.

Marcador	Baf 57	Baf 48	Baf 23	Baf 07	Baf 15	Baf 84	Baf 20
1	1	1	0	0	1	1	1
2	1	1	1	1	0	0	1
3	0	1	1	0	1	1	1
4	0	0	1	1	1	1	1
5	1	0	1	0	1	0	0
6	1	1	0	1	0	1	1
7	1	1	0	1	0	0	0
8	0	1	0	1	1	1	1
9	0	0	1	0	0	0	0
10	0	1	0	1	1	1	1
11	0	0	1	0	1	1	0
12	1	1	0	0	0	0	1
13	1	1	1	0	0	0	1
14	0	0	0	0	0	1	0
15	1	0	1	1	0	0	1
16	0	1	0	1	1	1	0
17	1	1	1	1	1	0	1
18	1	1	1	1	0	1	1
19	0	1	0	1	0	0	1
20	0	1	1	1	1	1	0

**TABELA 2.** Principais coeficientes de similaridade encontrados na literatura.

Coeficiente	Expressão da similaridade	Referência
Jaccard	$a/(a + b + c)$	JACCARD, 1901
Sorensen-Dice (Nei & Li)	$2a/(2a + b + c)$	DICE, 1945; SORENSEN, 1948
Russel e Rao	$a/(a + b + c + d)$	RUSSEL & Rao, 1940
Ochiai	$a / \sqrt{(a + b)(a + c)}$	OCHIAI, 1957
Coincidência Simples	$(a + d) / (a + b + c + d)$	SOKAL & MICHENER, 1958
Rogers e Tanimoto	$(a + d) / [a + 2(b + c) + d]$	ROGERS & TANIMOTO, 1960
Hamann	$[(a + d) - (b + c)] / (a + b + c + d)$	EMYGDIO et al., 2003
Kulczynski	$1 / 2 [a / (a + b) + a / (a + c)]$	EMYGDIO et al., 2003
Yule	$(ab - bc) / (ad + bc)$	EMYGDIO et al., 2003
Phi	$(ad - bc) / \sqrt{(a + b)(c + d)(a + c)(b + d)}$	EMYGDIO et al., 2003
Rogers Modificada	$[(b + c) / 2n]^{0.5}$	EMYGDIO et al., 2003

similaridade em distância existem e são válidas, como por exemplo,  $d_{ij} = -\ln(s_{ij})$  (Dias, 1998).

Analisando-se os acessos do exemplo em duplas, obtêm-se os índices **a, b, c e d**, para calcular-se o índice de similaridade e, posteriormente o índice de divergência genética (Tabela 3).

Tendo-se os índices, calculam-se os índices de similaridade e de divergência genética entre os acessos. Em primeira instância, utilizar-se-á o coeficiente de Sorensen - Dice ( $2a / (2a + b + c)$ ). Os coeficientes de similaridade entre os acessos são calculados e, logo após, calculam-se os coeficientes de dissimilaridade pela equação  $d_j = 1 - s_j$  (Tabela 4).

### Agrupamento

Após o cálculo da matriz com as distâncias

**TABELA 3.** Índices **a, b, c e d** entre acessos estudados.

Acessos/Índices	a	b	c	d
Baf57 e Baf48	8	2	6	4
Baf57 e Baf23	6	4	5	5
Baf57 e Baf07	6	4	6	4
Baf57 e Baf15	3	7	7	3
Baf57 e Baf84	3	7	8	2
Baf57 e Baf20	8	2	5	5
Baf48 e Baf23	6	8	5	1
Baf48 e Baf07	10	4	2	4
Baf48 e Baf15	7	7	3	3
Baf48 e Baf84	8	6	3	3
Baf48 e Baf20	11	3	2	4
Baf23 e Baf07	6	5	6	3
Baf23 e Baf15	6	5	4	5
Baf23 e Baf84	5	6	6	3
Baf23 e Baf20	7	4	6	3
Baf07 e Baf15	6	6	4	4
Baf07 e Baf84	7	5	4	4
Baf07 e Baf20	9	3	4	4
Baf15 e Baf84	8	2	3	7
Baf15 e Baf20	6	4	7	3
Baf84 e Baf20	7	4	6	3

**TABELA 4.** Coeficientes de dissimilaridade entre os acessos estudados.

Acessos	Baf 57	Baf 48	Baf 23	Baf 07	Baf 15	Baf 84	Baf 20
Baf 57	0,000	0,333	0,429	0,455	0,700	0,714	0,304
Baf 48	0,333	0,000	0,520	0,231	0,417	0,360	0,185
Baf 23	0,429	0,520	0,000	0,478	0,429	0,545	0,417
Baf 07	0,455	0,231	0,478	0,000	0,455	0,391	0,280
Baf 15	0,700	0,417	0,429	0,455	0,000	0,238	0,478
Baf 84	0,714	0,360	0,545	0,391	0,238	0,000	0,417
Baf 20	0,304	0,185	0,417	0,280	0,478	0,417	0,000

genéticas entre todos os itens, procede-se o agrupamento destes. Para tanto, utiliza-se de metodologias de agrupamento. O objetivo destes métodos é juntar aqueles itens com maior proximidade em grupos, resultando em um dendrograma de similaridade, que permite uma visualização mais fácil e rápida dos resultados (Dias, 1998).

Os métodos de agrupamento podem ser divididos em dois grupos, hierárquicos e não-hierárquicos, sendo o primeiro grupo mais utilizado para análise de divergência genética de espécies vegetais. Dentre os métodos não-hierárquicos, destacam-se os métodos de partição e os métodos de otimização (Dias, 1998). Estes métodos não envolvem construção de árvores ou dendrogramas. Estes métodos são raramente utilizados para análise de divergência genética intra-específica devida, principalmente, a falta de informação prévia sobre o número ótimo de grupos para alocação precisa dos indivíduos (Mohammadi et al., 2003).

As metodologias mais utilizadas em biologia são os métodos hierárquicos aglomerativos. Estes se caracterizam por séries sucessivas de fusões, ou seja, os itens são agrupados (fundidos) entre si, por subdivisões hierárquicas operadas sobre a matriz das distâncias genéticas calculada anteriormente, até resultar no dendrograma. Os métodos hierárquicos aglomerativos são o do Vizinho mais próximo, Vizinho mais distante, e das Médias das distâncias (Dias, 1998).

No método do vizinho mais próximo, identificam-se os itens mais próximos na matriz, e recalcula-se a distância deste conjunto para os outros itens  $\{d[k, (i \cup j)] = \min [d(k,i), d(k,j)]\}$ . Este cálculo é feito pegando-se a menor distância entre um dos dois itens fundidos e o outro item em questão. Por exemplo, se numa matriz fundirmos os itens 1 e 2, e a distância de um item 3 é menor para o 2 do que para o 1, a distância do grupo 1-2 para o item 3 será a distância entre 2 e 3. A partir do cálculo da nova matriz, segue-se sempre o mesmo princípio, fundindo os itens - grupos de menor distância e recalculando a matriz das distâncias.

No método do vizinho mais distante, inicia-se da mesma forma que o anterior, ou seja, fundem-se os dois itens mais próximos. A partir deste ponto,

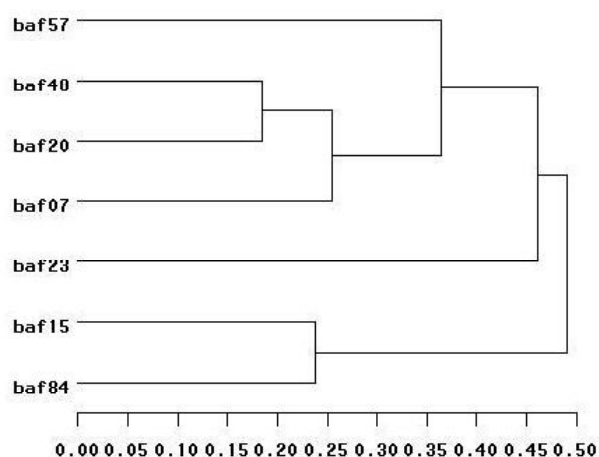
os cálculos da matriz das distâncias são feito pegando-se a maior distância entre este os itens deste grupo e outro item comparado  $\{d[k, (ij)] = \max[d(k,i), d(k,j)]\}$ . No mesmo exemplo anterior, se os itens 1 e 2 fossem fundidos, a distância para um item 3 seria a distância entre 3 e 1 (maior do que entre 2 e 3).

O método das Médias das distâncias, conhecido também como UPGMA (do inglês *unweighted pair-group method using an arithmetic average*), utiliza a média das distâncias entre os itens fundidos e um terceiro item  $\{d[k, (ij)] = p_i d(k,i) + p_j d(k,j)\}$ . Utilizando o mesmo exemplo anterior, se 1 e 2 fossem fundidos, a distância de 3 para o grupo 1-2 seria a média das distâncias de 3 para 1 e de 3 para 2. Seguem-se as fusões sempre entre os itens-grupos de menor distância genética na matriz. O método UPGMA é, dentre os métodos hierárquicos aglomerativos, o mais comumente utilizado (Mohammadi et al., 2003).

O resultado final de cada um dos métodos é um dendrograma, onde no eixo X colocam-se as distâncias, e no eixo Y os itens avaliados. O agrupamento dos itens forma uma “árvore” facilmente visualizada. Dias (1998) ressalta, porém, a necessidade da intuição e do conhecimento dos dados e algoritmo utilizado pelo analista, para uma interpretação precisa dos resultados.

Para o cálculo deste trabalho será utilizado o método das Médias das distâncias (UPGMA). Inicialmente, agrupam-se os dois acessos de menor distância genética, neste caso o Baf20 e o Baf48 (0,185). A partir deste coeficiente, calculam-se as distâncias genéticas dos outros acessos em relação ao grupo Baf20-Baf48, utilizando-se as médias das distâncias de cada acesso entre os acessos Baf 20 e Baf48. Por exemplo, a distância entre o acesso Baf57 e o grupo Baf20-Baf48 será  $d[Baf57, (Baf20Baf48)] = (0,304 + 0,333)/2 = 0,3185$ . Uma nova matriz é construída com as novas distâncias.

Novamente, agrupam-se os acessos com menor distância, neste caso o Baf5 e o Baf84 e constrói-se uma nova matriz. Este procedimento é realizado até ocorrer o agrupamento de todos os



**FIGURA 1.** Dendrograma de similaridade genética gerado pelo método UPGMA, utilizando-se o coeficiente de Sorensen - Dice.

acessos (Tabela 5). A partir deste ponto procedendo-se a esquematização do dendrograma de similaridade logo após.

Observa-se que os resultados no dendrograma (Figura 1) dispõem-se de forma facilitada para uma primeira observação. Os valores absolutos, porém, fazem-se necessários para comparações mais precisas, e para cálculos posteriores.

#### A escolha do coeficiente de similaridade e do método de agrupamento

Os coeficientes de Jaccard, Sorensen-Dice, Ochiai e Kulczynski apresentam a particularidade de não levarem em conta o índice **d**, resultante do não aparecimento de uma mesma banda nos dois indivíduos estudados. Segundo Duarte et al. (1999), esta característica é importante para marcadores RAPD, pois a ausência de banda amplificada em dois genótipos não significa, necessariamente, similaridade entre eles. O coeficiente de Sorensen-Dice, contudo, mede a proporção de bandas (alelos) compartilhadas, como resultado destes terem sido herdados de um mesmo ancestral, e representa a proporção de alelos presentes e compartilhados por dois genótipos (2a) divididos pela proporção média de bandas presentes em cada indivíduo  $(2a + b + c)$ . O coeficiente de Coincidência Simples leva em conta as concordâncias e discordâncias entre as bandas dos genótipos, dando peso igual para estes na estimativa da distância genética (Mohammadi et al., 2003).

A escolha do coeficiente a ser utilizado deve ser criteriosa, dependendo, inclusive, da espécie que se está estudando. Segundo Kosman & Leonard (2005), a escolha do coeficiente mais apropriado depende do tipo de marcador utilizado e do nível de ploidia do organismo em estudo. Aqueles coeficientes que não consideram o índice **d** (como Jaccard e Sorensen-Dice) são preferidos, já que não se sabe a

**TABELA 5.** Resultados do agrupamento segundo o método UPGMA.

Fases	Grupos		Níveis de fusão
6	Baf48	Baf20	0,1852
5	Baf15	Baf84	0,2381
4	CL6	Baf07	0,2554
3	Baf57	CL4	0,3641
2	CL3	Baf23	0,4609
1	CL2	CL5	0,4906



natureza da ausência da banda. O coeficiente de Jaccard tem sido largamente utilizado em estudos de biologia, para populações dentro de uma espécie, onde as concordâncias são freqüentes. O coeficiente de Sorensen-Dice é adequado para comparar itens muito diversos, onde a probabilidade de detecção dos índices *a* (presença comum de uma mesma banda) é menor (Dias, 1998).

Os coeficientes de Jaccard e de Sorensen-Dice são os mais utilizados para estudos de variabilidade genética em espécies medicinais. Bittencourt (2001) e Mossi et al. (2007), em estudo de variabilidade genética em populações naturais de *Maytenus ilicifolia*, Wadt (2001) em estudo em populações naturais de pimenta longa (*Piper hispidinervum* C.DC.), Echeverrigaray & Agostini (2006), em estudo da correlação genética entre cultivares comerciais e acessos brasileiros de *Salvia officinalis*, Palaniappan & Marappa (2008) em estudo entre acessos de *Phyllanthus amarus*, Gaia et al. (2004), estudando similaridade em populações naturais de *Piper aduncum*, Lerin et al. (2007a, 2007b), em estudo de diversidade genética de populações de *Salvia* sp., *Cymbopogon* sp., *Stipa tenacissima* e *Andropogon schoenanthus*, Silva et al. (2006), em estudo de variabilidade genética entre variedades de *Caesaria sylvestris*, utilizaram o coeficiente de Jaccard; Santos & Corrêa (2006), em estudos com *Chenopodium ambrosioides*, Moura et al. (2005) em estudos de diversidade genética em jaborandi (*Pilocarpus microphyllus* Stapf.), Ângelo et al., (2006) estudando a diversidade genética entre acessos de *Croton cajucara* utilizaram o coeficiente de Sorensen-Dice.

Trabalhos foram feitos para se comparar os coeficientes de similaridade em estudos em espécies não medicinais, que podem auxiliar na escolha destes. Duarte et al. (1999), comparando diferentes coeficientes de similaridade em estudo com feijão, determinaram que o coeficiente de Sorensen-Dice apresenta-se como o mais adequado para estudo de divergência genética para esta espécie, quando utilizados marcadores RAPD. Já Emygdio et al. (2003) constataram que os marcadores de Russel e Rao não se adequam para esta espécie e para marcadores dominantes, enquanto os coeficientes de Sorensen-Dice, Ochiai e Kulczynski são os mais adequados para o estudo de divergência genética em feijão por meio de marcadores RAPD. Meyer et al. (2004), utilizando marcadores RAPD e AFLP em milho para comparação de coeficientes de similaridade demonstram que, para esta situação, pode-se utilizar os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, já que os resultados para estes coeficientes apresentaram pouca variação. Os mesmos autores salientam, ainda, que este resultado vem corroborar a maior utilização do índice de Jaccard nas análises de divergência genética, apesar de não

ser o mais indicado para todas as espécies.

A escolha dos métodos de agrupamento também deve ser criteriosa. Os diferentes métodos podem produzir diferentes resultados aos mesmos dados. Dudley (1994) afirma que o método UPGMA, para melhoramento genético, é superior às metodologias do vizinho mais próximo e do vizinho mais distante. Ao se avaliar os métodos de agrupamento, devem-se levar em conta dois aspectos: a estabilidade do agrupamento formado e o grau de concordância entre o dendrograma e a matriz das distâncias. A primeira refere-se à possibilidade de se separar os dados em subgrupos e analisá-los separadamente, resultando no mesmo agrupamento quando analisado conjuntamente ou ainda a não distorção do agrupamento ao se excluir algum dado. O grau de concordância entre a matriz e o dendrograma é dado pelo coeficiente de correlação cofenética, calculado através da correlação entre a matriz inicial e a matriz dos valores cofenéticos entre os itens.

Alguns métodos estatísticos de análises, por exemplo, a análise de correlação cofenética (Rohlf & Sokal, 1961) associadas ao agrupamento, pode ser empregada para aumentar a confiabilidade dos agrupamentos gerados, bem como levar a interpretações menos dúbias dos agrupamentos formados. O coeficiente de correlação cofenética estima o grau de concordância entre o dendrograma formado pelo método de agrupamento e a matriz das distâncias. Calcula-se esta se correlacionando os dados da matriz cofenética C com os dados da matriz de distâncias D, pelo coeficiente de correlação de Pearson. A matriz C constitui-se das distâncias entre os pares de itens que representam os níveis de fusão verificados no dendrograma. Quando o coeficiente de correlação for acima de 0,80, tem-se um bom ajuste entre as duas matrizes (Dias, 1998).

O método UPGMA tem sido o mais utilizado para caracterização de diversidade genética em plantas medicinais (Jayaram & Prasad, 2008; Gaia et al., 2004; Lerin et al., 2007; Silva et al., 2006; Sandhu et al., 2006; Bellon et al., 2007; Bertoni et al., 2007; Mossi et al., 2007; Palaniappan & Marappa, 2008). Santos & Corrêa (2006) utilizaram o método do vizinho mais próximo para o agrupamento de acessos de *Chenopodium ambrosioides*. Arriel et al. (2006) compararam os métodos hierárquicos (UPGMA, vizinho mais próximo e vizinho mais distante) no agrupamento de acessos de gergelim usando marcadores RAPD, onde foi demonstrado o melhor ajuste das distâncias originais e estimadas pelo método UPGMA, apresentando um coeficiente de correlação cofenética de 0,89.

A utilização de métodos complementares para a melhor visualização da dispersão dos acessos tem sido uma alternativa aplicada por alguns autores. Santos & Corrêa (2006), no estudo da diversidade

genética de *Chenopodium ambrosioides* utilizaram o método do vizinho mais próximo juntamente com análises de dispersão bidimensional. Bellon et al. (2007), no estudo da variabilidade genética entre acessos selvagens e comerciais de *Passiflora edulis* utilizaram o método UPGMA para o agrupamento, e o método das coordenadas principais para análise de dispersão. Arriel et al. (2006) concluíram que o método UPGMA deve ser complementado com o método de otimização de Tocher para a análise de agrupamento. Relatam, ainda, que, caso o número de genótipos analisados seja elevado, deve-se utilizar a dispersão gráfica originadas pela análise das coordenadas principais para melhor distinção dos agrupamentos formados.

A escolha dos parâmetros estatísticos para análise dos resultados de um experimento deve obedecer a critérios para que os resultados apresentados sejam confiáveis. Cada coeficiente de similaridade possui características próprias que devem ser levadas em conta, juntamente com a espécie estudada. Estudos de comparação entre estes coeficientes dão aos pesquisadores as diretrizes para se conduzir de maneira coerente suas análises. De forma geral, os coeficientes de Jaccard e de Sorensen-Dice têm tido uma maior aceitação e têm sido os mais utilizados para a análise de divergência genética em plantas medicinais. Entretanto, selecionar o coeficiente a ser utilizado não pode depender apenas da utilização prévia em trabalhos anteriormente publicados para uma determinada espécie (Kosman & Leonard, 2005). Trabalhos prévios de eficiência dos coeficientes de similaridade para a espécie em estudo podem dar maior garantia na escolha.

Em relação às metodologias de agrupamento, Dias (1998) cita que a diferença entre os algoritmos de agrupamento (vizinho mais próximo, vizinho mais distante e UPGMA) está nas informações sobre diferentes aspectos, nas respostas a diferentes questões e na geração de diferentes soluções que proporcionam. Mohammadi et al. (2003) ressalta que poucos estudos foram realizados para determinar as vantagens e desvantagens de cada um. De uma forma geral, o método UPGMA é relatado como superior para o agrupamento no estudo de divergência genética (Dias, 1998; Arriel, 2006). Para maior confiabilidade, os trabalhos desta natureza em plantas medicinais devem conter uma justificativa para a escolha do coeficiente de similaridade e do método de agrupamento utilizados (Kosman & Leonard 2005), além de parâmetros estatísticos como o coeficiente de correlação de fenética.

## REFERÊNCIA

ÂNGELO, P.C.S. et al. Genetic diversity in sacaca (*Croton cajucara* Benth.) accessed by RAPD markers. **Revista Brasileira de Plantas Medicinais**, v.8, n.esp., p.18-22, 2006.

ARRIEL, N.H.C. et al. Técnicas multivariadas na determinação da diversidade genética em gergelim usando marcadores RAPD. **Pesquisa Agropecuária Brasileira**, v.41, n.5, p.801-9, 2006.

BELLON, G. et al. Variabilidade genética de acessos silvestres e comerciais de *Passiflora edulis* Sims. Com base em marcadores RAPD. **Revista Brasileira de Fruticultura**, v.29, n.1, p.124-7, 2007.

BERTONI, B.W. et al. Genetic variability in natural populations of *Zeyheria montana* Mart. from the Brazilian cerrado. **Scientia Agricola**, v.64, n.4, p.409-15, 2007.

BITTENCOURT, J.V.M. Variabilidade genética em populações naturais de *Maytenus ilicifolia* por meio de marcadores RAPD. **Scientia Agraria**, v.2, p.1-2, 2001.

CRUZ, C.D. **Programa Genes**: versão Windows; aplicativo computacional em genética e estatística. Viçosa: Ed. UFV, 2001. 648p.

DAWSON, I.K.; POWELL, W. Genetic variation in the Afromontane tree *Prunus africana*, an endangered medicinal species. **Molecular Ecology**, v.8, n.1, p.151-6, 1999.

DIAS, L.A.S. Análises multidimensionais. In: ALFENAS, L.C. (Ed.). **Eletroforese de isoenzimas e proteínas afins**: fundamentos e aplicações em plantas e microrganismos. Viçosa: Editora UFV, 1998. p.405-75.

DICE, L.R. Measures of the amount of ecologic association between species. **Ecology**, v.26, p.297-302, 1945.

DUARTE, J.M.; SANTOS, J.B.; MELO, L.C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics and Molecular Biology**, v.22, n.3, p.427-32, 1999.

DUDLEY, J.W. Comparison of genetic distance estimators using molecular marker data. In: SIMPOSIUM ANALYSIS OF MOLECULAR MARKER DATA, 1994, Oregon. **Proceedings...Oregon**: American Society for Horticultural Science/Crop Science Society American, 1994. p.3-7.

ECHEVERRIGARAY, S.; AGOSTINI, G. Genetic relationships between commercial cultivars and Brazilian accessions of *Salvia officinalis* L. based on RAPD markers. **Revista Brasileira de Plantas Medicinais**, v.8, n.esp., p.13-7, 2006.

EMYGDIO, B.M. et al. Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores RAPD. **Pesquisa Agropecuária Brasileira**, v.38, n.2, p.243-50, 2003.

FERREIRA, M.E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética**. 3.ed. Brasília: EMBRAPA-CENARGEM, 1998. 220p.

GAIA, J.M.D. et al. Similaridade genética de populações naturais de pimenta-de-macaco por análise RAPD. **Horticultura Brasileira**, v.22, n.4, p.686-9, 2004.

HAIR, J.F.; ANDERSON, R.E.; TATHAM, R.L. **Multivariate data analysis with readings**. 2.ed. New York: Macmillan, 1987. 449p.

HIDALGO, R. CIAT's World *Phaseolus* Collection. In: VAN SCHOONHOVEN, A.; VOYSEST, O. (Eds.). **Common Beans**: research for crop improvement. Wallingford: CAB International, 1991. 980p.

HOSOKAWA, K. et al. Discrimination among three species of medicinal *Scutellaria* plants using RAPD Markers. **Planta Medica**, v.66, p.270-2, 2000.

JACCARD, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. **Bulletin de la**

**Société Vaudoise des Sciences Naturelles**, v.37, p.547-79, 1901.

JAYARAM, K.; PRASAD, M.N.V. Genetic diversity in *Oroxylum indicum* (L.) Vent. (Bignoniaceae), a vulnerable medicinal plant by random amplified polymorphic DNA marker. **African Journal of Biotechnology**, v.7, n.3, p.254-62, 2008.

JAIN, N. et al. Molecular diversity in *Phyllanthus amarus* assessed through RAPD analysis. **Current Science**, v.85, n.10, p.1454-8, 2003.

KARP, A. **Molecular tolls in plant genetic resources conservation: a guide to the technologies**. Rome: International Plant Genetic Resources Institute, 2005. 126p.

KOSMAN E.; LEONARD, J. Similarity coefficients for molecular markers in studies of genetic relationship between individuals for haploid, diploid, and polyploid species. **Molecular Ecology**, v.14, p.415-24, 2005.

LERIN, L.A. et al. Avaliação da diversidade genética e comparação com o perfil químico dos óleos essenciais em *Salvia* spp. **Revista Brasileira de Biociências**, v.5, supl.1, p.369-71, 2007a.

LERIN, L.A. et al. Determinação de variabilidade genética em *Cymbopogon* sp., *Stipa tenacissima* L. e *Andropogon schoenanthus* L. usando Marcadores RAPD. **Revista Brasileira de Biociências**, v.5, supl.1, p.375-7, 2007b.

MEYER, A.S. et al. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (*Zea mays* L.). **Genetics and Molecular Biology**, v.27, n.1, p.83-91, 2004.

MOHAMMADI, S.A.; PRASANNA, B.M. Analysis of genetic diversity in crop plants – Salient statistical tools and considerations. **Crop Science**, v.43, n.4, p.1235-48, 2003.

MOSSI, A.J. et al. Intra and inter populational genetic variability in *Maytenus ilicifolia* Mart. ex Reiss. 1861, through RAPD markers. **Brazilian Journal of Biology**, v.67, n.4, p.957-61, 2007.

MOURA, E.F. et al. Genetic diversity in a jaborandi (*Pilocarpus microphyllus* Stapf.) germplasm bank assessed by RAPD markers. **Revista Brasileira de Plantas Mediciniais**, v.7, n.3, p.1-8, 2005.

NAKAI, R. et al. Genetic characterization of *Epimedium* species using random amplified polymorphic DNA (RAPD) and PCR-restriction fragment length polymorphism (RFLP) diagnosis. **Biological & Pharmaceutical Bulletin**, v.19, n.1, p.67-70, 1996.

NIERI, P. et al. Genetic characterization of the three medicinal *Echinacea* species using RAPD analysis. **Planta Medica**, v.69, p.685-6, 2003.

OCHIAI, A. Zoogeographic studies on the soleoid fishes found in Japan and its neighboring regions. **Bulletin of the Japanese Society of Scientific Fisheries**, v.22, p.526-30, 1957.

PALANIAPPAN, M.; MARAPPA, M. RAPD and ISSR analyses reveal low levels of genetic diversity in *Phyllanthus amarus*. **Acta Horticulturae**, v.765, p.179-88, 2008.

ROGERS, D.J.; TANIMOTO, T.T. A computer program for classifying plants. **Science**, v.132, p.1115-8, 1960.

ROHLF, F.J. **NTSYS-pc. Program numerical taxonomy and multivariate analysis system. Version 1.70**. New York: Exeter Software, 1992. 217p.

ROHLF, F.J.; SOKAL, R.R. Comparing numerical taxonomic studies. **Systematic Zoology**, v.30, p.459-90, 1981.

RUSSEL, P.F.; RAO, T.R. On habitat and association of species of anopheline larvae in south-eastern Madras. **Journal of the Malaria Institute of India**, v.3, p.153-78, 1940.

SANDHU, S.S. et al. Pilocarpine content and molecular diversity in Jaborandi. **Scientia Agricola**, v.63, n.5, p.478-82, 2006.

SANTOS, S.G.; CORRÊA, R.X. Diversidade genética de *Chenopodium ambrosioides* da região cacaueira da Bahia com base em marcadores RAPD. **Pesquisa Agropecuária Brasileira**, v.41, n.1, p.161-4, 2006.

SAS INSTITUTE INC. **Statistical analysis system. Release 6.08**. Cary, NC, 1992.

SILVA, M.A.S. et al. Phytochemical and genetic variability of *Casearia sylvestris* Sw. from São Paulo State Atlantic Forest and Cerrado populations. **Revista Brasileira de Plantas Mediciniais**, v.8, n.esp., p.159-66, 2006.

SOKAL, R.R.; MICHENER, C.D. A statistical method for evaluating systematic relationships. **University of Kansas Science Bulletin**, v.38, p.1409-38, 1958.

SORENSEN, T. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. Kongelige Danske Videnskabernes Selskabs. **Biologiske Skrifter**, v.5, p.1-34, 1948.

WADT, L.H.O. **Estrutura genética de populações naturais de pimenta longa (*Piper hispidinervum* C.D.C.) visando seu uso e conservação**. 2001. 95p. Tese (Doutorado - Área de Concentração em Genética e Melhoramento de Plantas) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.