COMPARAÇÃO DE COEFICIENTES DE SIMILARIDADE USADOS EM ANÁLISES DE AGRUPAMENTO COM DADOS DE MARCADORES MOLECULARES DOMINANTES

ANDRÉIA DA SILVA MEYER

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para obtenção do título de Mestre em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA
Estado de São Paulo – Brasil
Janeiro - 2002

COMPARAÇÃO DE COEFICIENTES DE SIMILARIDADE USADOS EM ANÁLISES DE AGRUPAMENTO COM DADOS DE MARCADORES MOLECULARES DOMINANTES

ANDRÉIA DA SILVA MEYER

Licenciada em Matemática

Orientador: Prof. Dr. **Antonio Augusto Franco Garcia**

Dissertação apresentada à Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, para obtenção do título de Mestre em Agronomia, Área de Concentração: Estatística e Experimentação Agronômica.

PIRACICABA

Estado de São Paulo – Brasil

Janeiro - 2002

Dados Internacionais de Catalogação na Publicação (CIP)
DIVISÃO DE BIBLIOTECA E DOCUMENTAÇÃO - ESALQ/USP

Meyer, Andréia da Silva
Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes / Andréia da Silva Meyer. - - Piracicaba, 2002.
106 p.

Dissertação (mestrado) - - Escola Superior de Agricultura Luiz de Queiroz, 2002.
Bibliografia.

1. Análise de conglomerados 2. Estatística aplicada 3. Genética molecular vegetal 4. Milho I. Título

CDD 519.53

A minha família, em especial aos meus pais Gerson e Gersina, por sempre me incentivarem a seguir em busca dos meus ideais.

DEDICO

AGRADECIMENTOS

A Deus, pela força para a realização desse trabalho.

Ao meu orientador Professor Doutor Antonio Augusto Franco Garcia, pela dedicação, disponibilidade e orientação na execução deste trabalho; pela amizade e apoio durante todo o curso e principalmente pela confiança em mim depositada.

Aos Professores Doutores Isaías Olívio Geraldi, Roseli Aparecida Leandro e Sônia Maria de Stefano Piedade, pelas sugestões na elaboração da dissertação.

Aos Professores Doutores Cláudio Lopes de Souza Jr. e Anete Pereira de Souza, pela gentil cessão dos dados usados no trabalho.

As minhas irmãs que tanto amo, Claudia e Luciane, aos meus cunhados, André e Nicéio, e aos meus sobrinhos, André, Vinícius e Vítor, por todo apoio, carinho e incentivo

À minha querida prima Paula, por ter me incentivado a ingressar no curso; por ter estado sempre ao meu lado dedicando todo o seu carinho e atenção e principalmente por poder ter contado sempre com seu ombro amigo.

As minhas colegas de moradia, Glaucy e Graciela, por todo apoio nas horas difíceis e também pelos ótimos momentos vivenciados juntas.

Aos amigos Cristiano, Denise, Ezequiel, Geneville, Janaína, Mariana, Osmir e Ramiro, pela ótima convivência durante o curso.

A André Jalles Monteiro, pelo apoio, carinho e paciência nas horas difíceis e principalmente pelos ótimos momentos vivenciados juntos.

Aos professores e funcionários do Departamento de Ciência Exatas da ESALQ - USP, especialmente as secretárias Luciane, Solange e Rosa, pelo carinho, respeito e amizade.

À Professora Doutora Maria Cristina Stolf Nogueira, pelo apoio e incentivo no início do curso.

Aos funcionários da biblioteca da ESALQ, pela ajuda na elaboração ficha catalográfica da dissertação.

À Escola Superior de Agricultura “Luiz de Queiroz”, pela oportunidade de realização do meu mestrado.

À Coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior – CAPES e ao Conselho Nacional de Desenvolvimento Científico- CNPq, pela concessão de bolsa de estudos.

A todos que de alguma forma contribuíram para a realização deste trabalho.

SUMÁRIO

COMPARAÇÃO DE COEFICIENTES DE SIMILARIDADE USADOS EM ANÁLISES DE AGRUPAMENTO COM DADOS DE MARCADORES MOLECULARES DOMINANTES

Autora: ANDRÉIA DA SILVA MEYER
Orientador: Prof. Dr. ANTONIO AUGUSTO FRANCO GARCIA

RESUMO

Estudos de divergência genética e relações filogenéticas entre espécies vegetais de importância agronômica têm merecido atenção cada vez maior com o recente advento dos marcadores moleculares. Nesses trabalhos, os pesquisadores têm interesse em agrupar os indivíduos semelhantes de forma que as maiores diferenças ocorram entre os grupos formados. Métodos estatísticos de análise, tais como análise de agrupamentos, análise de fatores e análise de componentes principais auxiliam nesse tipo de estudo. Contudo, antes de se empregar algum desses métodos, deve ser obtida uma matriz de similaridade entre os genótipos, sendo que diversos coeficientes são propostos na literatura para esse fim. O presente trabalho teve como objetivo avaliar se diferentes coeficientes de similaridade influenciam os resultados das análises de agrupamentos, feitas a partir de dados provenientes de análises com marcadores moleculares dominantes. Foram utilizados dados de 18 linhagens de milho provenientes de duas diferentes populações, BR-105 e BR-106, as quais foram analisadas por marcadores dos tipos AFLP e RAPD. Foram considerados para comparação os

coeficientes de Jaccard, Sorensen-Dice, Anderberg, Ochiai, Simple Matching, Rogers e Tanimoto, Ochiai II e Russel e Rao, para os quais foram obtidas as matrizes de similaridade. Essas matrizes foram comparadas utilizando as correlações de Pearson e Spearman, análise de agrupamentos com construção de dendrogramas, correlações, distorção e estresse entre as matrizes de similaridade e as matrizes cofenéticas, índices de consenso entre os dendrogramas, grupos obtidos com o método de otimização de Tocher e com a projeção no plano bidimensional das matrizes de similaridade. Os resultados mostraram que para praticamente todas metodologias usadas, para ambos marcadores, os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai mostraram resultados semelhantes entre si, o que foi atribuído ao fato deles apresentarem como propriedade comum a desconsideração da ausência conjunta de bandas. Isso também foi observado para os coeficientes de Simple Matching, Rogers e Tanimoto e Ochiai II, que também não apresentaram entre si grandes alterações nos resultados, possivelmente devido ao fato de todos considerarem a ausência conjunta. O coeficiente de Russel e Rao apresentou resultados muito diferentes dos demais coeficientes, em função dele excluir a ausência conjunta do numerador e incluí-la no denominador, não sendo recomendado seu uso. Devido ao fato da ausência conjunta não significar necessariamente que as regiões do DNA são idênticas, sugere-se a escolha dentre os coeficientes que desconsideram a ausência conjunta.

COMPARISON OF SIMILARITY COEFFICIENTS USED IN CLUSTER ANALYSIS WITH DOMINANT MARKERS DATA

Author: Andréia da Silva Meyer

Adviser: Prof. Dr. Antonio Augusto Franco Garcia

SUMMARY

With the recent advent of the molecular markers, studies of divergence and phylogenetic relationships between and within vegetable species of agricultural interest have been received greater attention. In these studies, the aim is to group similar individuals looking for bigger differences among the groups. Statistical methods of analysis such as cluster analysis, factor analysis and principal components analysis can be used in this kind of study. However, before to employ some method, the similarity matrix between genotypes must be obtained using one of the several coefficients proposed in the concerning literature. The aim of this study was to evaluate if different similarity coefficients can influence the results of cluster analysis with dominant markers. Data from 18 inbred lines of maize from two different populations, BR-105 and BR-106, were analyzed by AFLP and RAPD markers and eight similarity coefficients (Jaccard, Sorensen-Dice, Anderberg, Ochiai, Simple-matching, Rogers and Tanimoto, Ochiai II and Russel and Rao) were obtained. The similarity matrices were compared by Pearson´s and Spearman´s correlations, cluster analysis (with dendrograms, correlations,

distortion and stress between the similarity and cofenetical matrices, consensus fork index between all pairs of dendrograms), Tocher´s optimization procedure and with the projection in two-dimensional space of the similarity matrices. The results showed that for almost all of the methodologies and both markers, the coefficients of Jaccard, Sorensen-Dice, Anderberg and Ochiai, gave similar results, due to the fact that all of them excludes negative co-occurences. It was also observed that the Simple Matching, Rogers and Tanimoto, and Ochiai II, probably due to the fact of all including the negative co-occurences. The Russel and Rao coefficient presented results very different from the others, because it excludes the negative co-occurences in the numerator and include it in the denominator of its expression, which is a reason for not recommending it. Due the fact of the negative co-occurences does not mean, necessarily, that the regions of the DNA are identical, it is suggested to choose one those coefficients that do not include it.

1 INTRODUÇÃO

Estudos de divergência genética e relações filogenéticas entre espécies vegetais de importância agronômica têm merecido atenção cada vez maior, principalmente com o advento recente dos chamados marcadores moleculares (Duarte et al., 1999). Nesses trabalhos, os pesquisadores têm interesse em agrupar os indivíduos semelhantes, de forma que as maiores diferenças ocorram entre os grupos formados.

Os marcadores moleculares fornecem informações diretamente sobre o material genético (DNA) dos indivíduos (ou genótipos) em estudo, com o emprego de métodos bioquímicos que identificam diferenças diretamente no DNA. Tais diferenças são, então, codificadas na forma de uma matriz retangular, cujas colunas identificam os genótipos em questão, que podem ser linhagens, clones, híbridos etc. As linhas, por sua vez, indicam as regiões do DNA em que foram avaliadas as diferenças nas características de interesse (Ferreira & Grattapaglia, 1996).

Os marcadores revelam, para uma dada região do DNA, uma marca ou banda que permite comparar os indivíduos em estudo quanto à sua presença ou ausência. As bandas reveladas são codificadas pelo número 1 e as não reveladas pelo número 0. Dessa forma, os dados aos quais serão aplicados métodos estatísticos provêm dessa matriz de uns e zeros.

Métodos estatísticos de análise, tais como análise de agrupamentos, análise de fatores, análise discriminante, análise de componentes principais, podem ser aplicados para auxiliar nesse tipo de estudo. Dentre eles, destaca-se a análise de agrupamentos por não exigir pressuposição inicial quanto à distribuição de probabilidades dos dados e por ser de fácil interpretação. Essa técnica é muito usada,

principalmente pelos pesquisadores na área de melhoramento genético, em estudos de divergência, além de estudos evolutivos.

Contudo, antes de se empregar algum desses métodos, deve ser obtida uma matriz de similaridades (ou distância) entre os genótipos. Essas distâncias podem ser calculadas de diversas formas, sendo que diferentes propostas são encontradas atualmente na literatura (Sokal & Sneath, 1963; Sneath & Sokal, 1973; Johnson & Wichern, 1988; Weir, 1996).

Esses coeficientes de similaridade são específicos para variáveis dicotômicas e seu uso é sugerido para estudos de divergência baseados em marcadores moleculares (Duarte et al., 1999). De forma geral, eles baseiam-se em comparações entre a ocorrência de bandas em comum (indicados por uns em comum na matriz de dados) e bandas diferentes (indicados por um e zero ou zero e um) entre cada par de genótipos. Alguns coeficientes consideram também a ocorrência de zeros em comum. Seus valores comumente variam entre 0 e 1 (Skroch et al., 1992).

Considerando que os resultados dos agrupamentos podem ser influenciados pela escolha do coeficiente de similaridade (Gower & Legendre, 1986; Jackson et al., 1989; Duarte et al., 1999), estes coeficientes precisam ser melhor entendidos, de forma que os mais eficientes em cada situação específica possam ser empregados. Outro aspecto a ser considerado é que, usualmente, os artigos publicados não justificam a escolha dos coeficientes empregados, mostrando a necessidade de estudos a esse respeito (Duarte et al., 1999). Além disso, são raros os artigos que comparam os coeficientes para análises de agrupamento, principalmente usando dados de diferentes marcadores moleculares.

O objetivo deste trabalho foi pesquisar a influência da escolha de diferentes coeficientes de similaridades sobre a subseqüente análise de agrupamentos, feita a partir de dados provenientes de análise com marcadores moleculares dominantes, para 18 linhagens de milho.

2 REVISÃO DE LITERATURA

2.1 Marcadores Moleculares

Diversas técnicas de biologia molecular estão disponíveis, atualmente, para detecção de variabilidade genética no DNA, isto é, para detecção de polimorfismo genético. O DNA é o material genético dos seres vivos e, assim, as diferenças encontradas podem ser usadas para estudos refinados de divergência entre indivíduos.

Estas técnicas permitem a obtenção de um número virtualmente ilimitado de marcadores moleculares, cobrindo todo o genoma (DNA total) do organismo. Tais marcadores podem ser utilizados para as mais diversas aplicações, tanto em estudos de genética básica como no melhoramento de plantas (Ferreira & Grattapaglia, 1996).

Um marcador de DNA é tipicamente uma pequena região do DNA mostrando polimorfismo (ou diferença) entre indivíduos (Liu, 1998). Dois enfoques metodológicos são usados para detectar tais diferenças: hibridação e amplificação. No primeiro, uma pequena seqüência é marcada radioativamente e ligada ao DNA dos indivíduos, usando princípios de pareamento das bases do DNA. Nesse caso, enquadram-se, por exemplo, os marcadores do tipo RFLP ("restriction fragment length polymorphism").

Os marcadores baseados em amplificação têm como princípio o emprego de reações para aumento de seqüências específicas, sendo os principais desse tipo o SSR (ou microsatélites), o RAPD ("random amplified polymorphism DNA") e o AFLP ("amplified fragment length polymorphism").

Uma característica fundamental nos marcadores RAPD e AFLP é o fato de eles se comportarem usualmente como marcadores dominantes, e os marcadores RFLP e SSR, como codominantes. Isso implica em diferentes abordagens na escolha dos coeficientes de similaridade, uma vez que os dados obtidos são de natureza diferente.

Os resultados obtidos com o emprego dos marcadores moleculares fornecem uma matriz de valores binários, em que as colunas dessa matriz representam os diferentes genótipos e as linhas representam as diferentes características a serem comparadas. No caso dos dominantes, cada linha corresponde à identificação de uma marca (ou banda) numa região específica do DNA. A presença da marca é identificada pelo número 1 e a ausência, por 0.

Pode ser obtida, por exemplo, a seguinte matriz:

$$ {}_{p}M_{n} = \begin{bmatrix} 1 & 1 & 0 & \dots & 1 \\ 1 & 0 & 1 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 1 & \dots & 1 \end{bmatrix}, $$

em que p e n indicam o número de marcas e de genótipos, respectivamente. Quando dois genótipos são comparados podem ocorrer as seguintes situações, para cada marca p: i) ambos possuem a mesma marca, ou seja, valor 1 (código 1 e 1); ii) um genótipo possui a característica e o outro não (código 1 e 0, ou 0 e 1); iii) ambos não possuem tal característica (código 0 e 0).

2.2 Análises de Agrupamento

As técnicas de agrupamentos têm por objetivo agrupar indivíduos em classes. Portanto, dado um conjunto de n indivíduos, todos avaliados para p variáveis, tais indivíduos devem ser agrupados em classes, de forma que os mais semelhantes permaneçam na mesma classe. De forma geral, o número de classes não é conhecido inicialmente. Porém, quando essas técnicas geram grupos não esperados, isso pode sugerir que as relações entre os objetos precisam ser melhor estudadas (Manly, 1994).

É necessária, para a utilização das técnicas de análise de agrupamentos, decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição. Tais decisões podem envolver conteúdos metodológicos ou de caráter técnico.

Inicia-se o processo definindo-se os indivíduos e os objetivos desejados para a aplicação da análise, além dos critérios que irão definir as semelhanças entre eles. Obtidos esses dados, eles são dispostos na forma de uma matriz, em que as colunas representam os indivíduos de interesse e as linhas representam as variáveis .

Pode ocorrer que as variáveis consideradas não sejam medidas na mesma unidade. Assim, é possível que elas não contribuam igualmente para a similaridade entre os indivíduos, ou ainda, que tenham alguma influência arbitrária nos grupos formados. Para solucionar tais problemas é usualmente empregada a padronização (ou estandardização) dos dados (Sneath & Sokal, 1963; Johnson & Wichern, 1988; Bussab et al., 1990):

$$Z_{ik} = \frac{x_{ik} - \overline{x}_k}{s_k},$$

em que Z_{ik} é o valor estandardizado para o indivíduo i $(i = 1,\ldots,n)$ e a variável k $(k = 1,\ldots, p)$; $\overline{x}_k$ e s_k denotam a média e o desvio padrão das x_{ik} observações originais, respectivamente, para a variável k. No caso dos marcadores moleculares, essa etapa não é necessária.

Obtida a matriz de dados padronizados, o próximo passo é a escolha de uma medida que quantifique o quanto dois indivíduos são parecidos. Tais medidas são denominadas coeficientes de similaridade e são elas que vão gerar a matriz *D* de similaridade:

$$D_{ij} = \begin{bmatrix} d_{12} & d_{13} & \ldots & d_{1n} \\ & d_{23} & \ldots & d_{2n} \\ & & \ldots & \ldots \\ & & & d_{n-1,\,n} \end{bmatrix},$$

em que d_{ij} representa a distância entre os indivíduos i e j $(i = 1,\ldots,n-1);(j = 2,\ldots,n)$.

Tais coeficientes podem ser divididos em duas categorias: medidas de similaridade e medidas de dissimilaridade. Para a primeira categoria, quanto maior o valor observado, mais parecidos são os indivíduos; para a segunda, quanto maior o valor observado, menos parecidos são os indivíduos. Para cada tipo de variável (quantitativas, qualitativas ordinais, qualitativas nominais e mistas), são definidos diferentes coeficientes de similaridade (Sneath & Sokal, 1963; Johnson & Wichern, 1988; Bussab et al., 1990).

De um modo geral, os coeficientes de similaridade são criados com o intuito de moldar situações especiais de interesse do pesquisador. Por esse motivo, dispõe-se de uma série bem ampla de tais medidas. Um levantamento e uma análise das propriedades desses coeficientes ajudam a identificar alguns princípios gerais e encontrar algum coeficiente que melhor se ajuste aos interesses de uma pesquisa em particular (Bussab et al., 1990). Contudo, essa escolha não é trivial, e os trabalhos, comumente, não justificam nenhuma escolha em particular, notadamente para dados de marcadores moleculares.

Os marcadores moleculares geram variáveis qualitativas binárias, caracterizadas pela presença ou ausência da marca (banda) após as análises laboratoriais. Dunn & Everitt (1980) designaram por zeros e uns os estados do caráter para caracteres binários, que serão submetidos posteriormente à análise. Como definido anteriormente, o número 1 indica a presença da marca e 0, a ausência. Genótipos que possuam maior coincidência de presenças e/ou ausências das marcas serão considerados mais similares entre si .

De acordo com Carlini-Garcia (1998), para cada par de genótipos (i, j) estudado, as medidas de similaridade/dissimilaridade para caracteres binários, baseiam-se em tabelas 2 x 2, como apresentado na Tabela 1:

Tabela 1. Quantificação da similaridade/dissimilaridade entre os genótipos *i* e *j*.

		Genótipo *i*		
		1	0	
Genótipo *j*	1	a	b	$a+b$
	0	c	d	$c+d$
		$a+c$	$b+d$	$a+b+c+d$

Nessa situação, a é o número de marcas para os quais ambos os genótipos tiveram código 1 (presença), b é o número de códigos 0 para o genótipo i e 1 para o genótipo j, c é o número de códigos 1 para o genótipo i e 0 para o genótipo j, d é o número de marcas para os quais ambos os genótipos tiveram código 0 (ausência), $a+b+c+d$ é o número total de caracteres (ou marcas) binários estudados. O número de coincidências é representado por $a+d$ e o número de diferenças é representado por $b+c$ (Dunn & Everitt, 1980). Essas indicações serão usadas em diversas partes do presente trabalho.

De um modo geral, as medidas de similaridade e de dissimilaridade são interrelacionadas e facilmente transformáveis entre si (Clifford & Stephenson, 1975; Bussab et al., 1990). Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura. Segundo Clifford & Stephenson (1975) , tais coeficientes podem ser divididos em cinco diferentes classes, apresentadas a seguir.

i) Coeficientes de Similaridade

Disponíveis para dados binários, baseiam-se na comparação entre o número de atributos comuns para um par de objetos e o número total de atributos envolvidos. Tais coeficientes podem ser facilmente convertidos para coeficientes de dissimilaridade:

se a similaridade for denominada s, a medida de dissimilaridade será o seu complementar $(1-s)$. Os coeficientes de similaridades podem ser divididos em dois grupos: os que consideram a ausência conjunta e os que não consideram a ausência conjunta (Carlini-Garcia, 1998). Alguns coeficientes de similaridade que consideram a ausência conjunta são apresentados na Tabela , ressaltando que ela é indicada pela letra d nas expressões.

Tabela 2. Coeficientes de similaridade que consideram a ausência conjunta.

Coeficientes	**Fórmula***	**Intervalo de Ocorrência**
Simple Matching (1958)	$\frac{a+d}{a+b+c+d}$	[0, 1]
Russel e Rao (1940)	$\frac{a}{a+b+c+d}$	[0, 1]
Rogers e Tanimoto (1960)	$\frac{a+d}{a+d+2(b+c)}$	[0, 1]
Hamann (1961)	$\frac{(a+d)-(b+c)}{a+b+c+d}$	[−1, 1]
Ochiai II (1957)	$\frac{ad}{\sqrt{(a+d)(a+c)(d+b)(d+c)}}$	[0, 1]
Sokal e Sneath (1963)	$\frac{2(a+d)}{2(a+d)+b+c}$	[0, 1]

*O significado de a, b, c e d é apresentado na Tabela 1.

Estes coeficientes têm propriedades semelhantes devido ao fato de considerarem a ausência conjunta. Porém, variam em relação à importância dada à ausência e à presença conjunta, bem como às não-coincidências (Carlini-Garcia, 1998). Todos variam no mesmo intervalo, exceto o coeficiente de Hamann.

Alguns coeficientes de similaridade que desconsideram a ausência conjunta são apresentados na Tabela 3.

Tabela 3. Coeficientes de similaridade que desconsideram a ausência conjunta

Coeficientes	**Fórmula***	**Intervalo de Ocorrência**
Jaccard (1908)	$\frac{a}{a+b+c}$	$[0, 1]$
Anderberg (1973)	$\frac{a}{a+2(b+c)}$	$[0, 1]$
Czekanowsky (1913)	$\frac{2a}{2a+b+c}$	$[0, 1]$
Kulczynski I (1927)	$\frac{a}{b+c}$	$[0, +\infty)$
Kulczynski II (1927)	$\frac{a}{2}(\frac{1}{a+b}+\frac{1}{a+c})$	$[0, 1]$
Sorensen-Dice (1945)	$\frac{2a}{2a+b+c}$	$[0, 1]$
Ochiai (1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	$[0, 1]$

* O significado de a, b, c e d é apresentado na Tabela 1.

Segundo Clifford & Stephenson (1975), a escolha dos coeficientes, que estão restritos ao intervalo $[0, 1]$ é mais adequada, pois índices que tendem ao infinito são sensíveis a pequenas mudanças, especialmente em a. Alguns desse coeficientes têm princípios fáceis de entender, como por exemplo, o coeficiente de Jaccard, que compara o número de presenças de bandas comuns e o número total de bandas envolvidas, excluindo o número de ausências conjuntas. Porém, outros não são de interpretação simples, como é o caso do coeficiente de Kulczynski II, por exemplo.

ii) Coeficientes de Associação

Tais coeficientes mostram como os pares de indivíduos estão associados. Geralmente variam de -1, quando a mudança em uma variável é acompanhada por mudança de igual magnitude na outra, porém em sentido contrário, a $+1$ quando a mudança em uma variável é acompanhada por mudança de igual magnitude na outra (Clifford & Stephernson, 1975; Carlini-Garcia, 1998). Alguns deles são mostrados na Tabela 4.

Tabela 4. Coeficientes de associação contidos no intervalo $[-1,+1]$.

Coeficientes	**Fórmula**	**Respectivos Coeficientes de Dissimilaridade**
Yule	$\frac{ad-bc}{ad+bc}$	$\frac{\left(1-\frac{ad-bc}{ad+bc}\right)}{2}$
Pearson	$\frac{(ad-bc)}{(a+b)(c+d)(a+c)(b+d)}$	$\frac{\left(1-\frac{(ad-bc)}{(a+b)(c+d)(a+c)(b+d)}\right)}{2}$
McConnaughy	$\frac{a^2-bc}{(a+b)(a+c)}$	$\frac{\left(1-\frac{a^2-bc}{(a+b)(a+c)}\right)}{2}$

Fonte: Carlini-Garcia (1998)

Há também coeficientes de associação que variam no intervalo $[0,+\infty)$, sendo mostrados na Tabela 5.

Tabela 5. Coeficientes de associação contidos no intervalo $[0, +\infty)$.

Coeficientes	Fórmula
X^2	$X^2 = \dfrac{(ad-bc)(a+b+c+d)}{(a+b)(c+d)(a+c)(b+d)}$
M.S.C. ("mean square contingency")	$M.S.C. = \dfrac{X^2}{a+b+c+d}$

Fonte: Carlini-Garcia (1998)

iii) Distância Euclidiana

Considerada como uma medida de dissimilaridade, a distância euclidiana é interpretada como a distância entre dois indivíduos, cujas posições são determinadas em relação às suas coordenadas, definidas com referência a um grupo de eixos cartesianos, os quais possuem ângulos retos entre si (Clifford & Stephenson, 1975). Considerando apenas duas variáveis, sua fórmula é dada por:

$$d_{ij} = [(x_{ia} - x_{ja})^2 + (x_{ib} - x_{jb})^2]^{\frac{1}{2}},$$

em que x_{ia} e x_{ib} representam, respectivamente, as características a e b do indivíduo i e x_{ja} e x_{jb} representam, respectivamente, as características a e b do indivíduo j.

Generalizando, para p variáveis, tem-se:

$$d_{ij} = \left[\sum_{k=1}^{p} (x_{ik} - x_{jk})^2 \right]^{\frac{1}{2}},$$

em que x_{ik} e x_{jk} representam, respectivamente, os indivíduos i e j para a k-ésima variável $(k = 1, \ldots, p)$.

A distância euclidiana como medida de dissimilaridade pode ser aplicada para dados binários, sendo, contudo, mais comum nesse caso o uso da distância euclidiana quadrada (D^2), pois esta tem a vantagem da diferença entre os indivíduos ser superior ao número de atributos em que eles diferem, com a restrição de que não haja

valores faltantes. Assim, todos os valores entre os indivíduos são baseados no mesmo número de comparações. Uma outra opção é o uso da distância euclidiana média (Bussab et al., 1990).

iv) Conteúdo de Informações ou Medidas de Diversidade

São medidas de dissimilaridade somente entre indivíduos. Tais medidas são utilizadas para gerar classificações nas quais grupos são unidos de forma a minimizar a diversidade intragrupo a cada passo. O intervalo de variação das medidas de informação ou diversidade pode ser de zero a valores muito grandes e não possuem medidas de similaridade correspondentes (Clifford & Stephenson, 1975; Carlini-Garcia, 1998).

v) Medidas de Similaridade Dependentes da Probabilidade Estimada

Em tais medidas, os objetos a serem classificados são considerados como amostras de uma população maior e podem ser estimadas as probabilidades de se obterem, por acaso, pares de objetos tão similares quanto àqueles sob observação. A probabilidade é estimada com a determinação prévia do valor X^2 que será baseado no grau de associação entre os objetos. Tais probabilidades podem ser usadas como medida de dissimilaridade e seu complemento como medidas de similaridade (Clifford & Stephenson, 1975; Carlini-Garcia, 1998).

2.3 Classificação das técnicas de agrupamento

Vários são os tipos de técnicas de agrupamento encontradas na literatura. Manly (1994) destacou duas abordagens particulares, destacando, a primeira delas: i) técnicas que produzem dendrogramas, em que o primeiro passo é calcular as medidas de dissimilaridade (ou similaridade) entre todos os pares possíveis de indivíduos e, assim, formar os grupos por processos aglomerativos ou divisivos; ii) técnicas que envolvem

partições, em que os indivíduos podem se mover fora e dentro dos grupos em diferentes estágios da análise.

Inicialmente, na aplicação dessas técnicas, são estabelecidos centros de grupos, arbitrariamente, e cada indivíduo é agrupado em relação ao centro mais próximo. Novos centros são calculados e cada indivíduo se move para o grupo cujo centro seja mais próximo de si. O processo continua de forma iterativa até encontrar a estabilidade nos grupos. Descrições detalhadas desses métodos são fornecidas por Anderberg (1973), Sneath & Sokal (1973), Clifford & Stephenson (1975), Morrison (1976), Mardia et al. (1979), Chatfield & Collins (1980), Dunn & Everitt (1980), Bussab et al. (1990), Johnson & Wichern (1992), Manly (1994), Totti (1998).

Segundo Sneath & Sokal (1973), tais métodos podem ser classificados em diversas categorias contrastantes, apresentadas a seguir.

a) Métodos aglomerativos e divisivos

Nos métodos aglomerativos, consideram-se, inicialmente, t indivíduos, que serão agrupados de forma sucessiva, baseando-se em sua proximidade, em g grupos $(g < t)$. No processo divisivo, o procedimento é oposto, ou seja, inicialmente todos os indivíduos estão num mesmo grupo, que se divide em um ou mais subgrupos, os quais se subdividem sucessivamente até o final do processo. Nos processos divisivos, se um indivíduo for alocado de maneira inadequada, sua posição não será posteriormente corrigida.

b) Métodos hierárquicos e métodos não hierárquicos

Um método de agrupamento é considerado hierárquico se ele consiste numa seqüência de $(w+1)$ agrupamentos $(G_0, G_1, \ldots, G_w)$, em que G_0 é a partição disjunta de todos os n indivíduos e G_w é a partição conjunta. O número de partes k_i na partição G_i

deve obedecer à regra $k_i \geq k_{i+1}$, em que k_{i+1} é o número de partes do grupo G_{i+1}. Os métodos são considerados não hierárquicos quando o indivíduo não tem ordem parcial.

c) Métodos sem sobreposição e métodos com sobreposição

Nos métodos sem sobreposição, os grupos, em dado nível hierárquico, são mutuamente exclusivos. A classificação hierárquica é dada quando a sobreposição está relacionada com a hierarquia.

d) Métodos seqüenciais e métodos simultâneos

Nos métodos seqüenciais, é aplicada, ao grupo de indivíduos estudados, uma seqüência recorrente. Nos métodos simultâneos, um único procedimento não recorrente é aplicado ao grupo inteiro de indivíduos.

e) Critérios locais e critérios globais

Em geral, muitos métodos seqüenciais aglomerativos estimam a similaridade entre objetos dentro de um grupo de forma confiável. Contudo, quanto maior forem os grupos considerados, menor será esta confiabilidade. Portanto, as confiabilidades locais e globais podem diferir em vários métodos de agrupamento.

f) Soluções diretas e soluções iterativas

São considerados métodos de agrupamentos diretos aqueles em que o algoritmo, para a construção da classificação, é feito de modo direto e a solução obtida é considerada como ótima em algum sentido. Os procedimentos iterativos visam à otimização local, global ou ambas.

g) Agrupamentos ponderados e agrupamentos não ponderados

Os métodos ponderados dão pesos aos ramos agrupados durante o processo de agrupamento de um método aglomerativo seqüencial. A ponderação também é usada quando se considera que algumas dimensões são mais importantes do que as outras.

h) Agrupamentos adaptativos e agrupamentos não adaptativos

Nos métodos de agrupamentos não adaptativos, o algoritmo é aplicado de forma iterativa ou direta, para encontrar uma solução em que o método de agrupamento é programado para interagir com todos os pontos no espaço A-dimensional para formar o grupo.

2.4 Métodos de agrupamento seqüenciais, aglomerativos, hierárquicos, sem sobreposição

Nestes métodos, conhecidos como SAHN ("sequencial, agglomerative, hierarquic, nonoverlapping clustering methods"), em cada passo do agrupamento há a necessidade de recalcular o coeficiente de similaridade (ou dissimilaridade) entre os grupos estabelecidos e os possíveis candidatos a futuras admissões no grupo. Além disso, reconsidera-se também o critério de admissão de novos membros aos grupos já estabelecidos (Sneath & Sokal, 1973). Tais métodos, são amplamente usados para estudos de divergência genética, objeto do presente trabalho, e assim serão detalhados nesse item.

Segundo Sneath & Sokal (1973), o critério de admissão de novos indivíduos a um grupo, para qualquer método SAHN, pode ser apresentado, de forma geral, como a seguir.

Seja U uma medida de dissimilaridade e seja L um indivíduo pertencente a qualquer grupo que não J, K. Sendo J denominado um grupo ou indivíduo, esse juntar-se-á a K, se e somente se $U_{JK} < U_{JL}$ e $U_{JK} < U_{KL}$. Isso significa que J e K

formam o par mais próximo de indivíduos ou de grupos. Freqüentemente, ocorre a seguinte situação: $U_{JK} < U_{JL}$, mas $U_{JK} = U_{KM} < U_{KL}$ (empates), sendo M outro indivíduo ou grupo. Nestes casos, as decisões são tomadas arbitrariamente.

A seguir, são apresentados diversos métodos de agrupamento que fazem parte dos métodos SAHN, e têm muita aplicação prática. Vale ressaltar que a escolha de um método de agrupamento depende do material e dos objetivos em questão, pois métodos diferentes podem conduzir a resultados bem distintos. Não há método considerado como o melhor, mas alguns são mais indicados para determinadas situações do que os outros (Kaufman & Rosseeuw, 1990).

2.4.1 Método da Ligação Simples ou do Vizinho mais Próximo

Neste método, também denominado "Single Linkage Clustering", as conecções entre objetos e grupos ou entre grupos são feitas por ligações simples entre pares de objetos, ou seja, a distância entre os grupos é definida como sendo aquela entre os objetos mais parecidos entre esses grupos. Este método leva a grupos longos se comparados aos grupos formados por outros métodos de agrupamento SAHN. Os dendrogramas (gráficos em formas de árvores) resultantes deste procedimento são geralmente pouco informativos, devido à informação dos indivíduos intermediários que não são evidentes (Carlini-Garcia, 1998). De acordo com Sneath & Sokal (1973), agrupamentos pelo método de ligação simples podem ser obtidos tanto pelo procedimento aglomerativo quanto divisivo.

2.4.2 Método da Ligação Completa ou do Vizinho mais Longe

Segundo Bussab et al. (1990), neste método, também denominado "Complete Linkage Clustering", a similaridade entre dois grupos é definida como aquela apresentada pelos indivíduos de cada grupo que menos se parecem, ou seja, formam-se

todos os pares com um membro de cada grupo, e a similaridade entre os grupos é definida pelo par que menos se parece. Este método, geralmente, leva a grupos compactos e discretos, tendo os seus valores de similaridade relativamente pequenos.

2.4.3 Métodos de Ligação Média

Estes métodos também denominados como "Average Linkage Clustering", são uma ponderação entre os métodos de ligação simples e completa. É usada a similaridade média dos indivíduos ou grupo que pretende se unir a um grupo já existente. Há vários tipos de métodos, pois há vários tipos de médias, sendo que quatro são mais comuns, provenientes da combinação de dois critérios alternativos: agrupamento em função da média aritmética *versus* agrupamento com base no centróide, podendo ser ou não ponderados em ambos casos.

Nos métodos de agrupamento com base na média aritmética, os coeficientes de similaridade (ou dissimilaridade) médios entre o indivíduo que se pretende agrupar e os indivíduos do grupo já existente são calculados. O método do centróide busca o centróide dos indivíduos para construir grupos e medir a dissimilaridade relativa a esse ponto entre qualquer indivíduo ou grupo candidato. Os métodos ponderados pretendem dar pesos iguais a todos o ramos do dendrograma, sendo que o número de indivíduos que compõem cada ramo não são considerados (Bussab et al., 1990).

Sneath & Sokal (1973) descrevem as quatro combinações possíveis para esses critérios descritos:

a) UPGMA (“Unweighted Pair-Group Method Using Arithmetic Averages”)

Este algoritmo não considera a estrutura de subdivisão do grupo, dando pesos iguais a cada indivíduo do grupo e calcula a similaridade média de um indivíduo que pretende se juntar ao grupo já existente. Nos estudos de divergência genética, esse algoritmo é o mais comum, como mostra o trabalho de Duarte et al. (1999), por exemplo.

b) WPGMA (“Weighted Pair-Group Method Using Arithmetic Averages”)

Difere do anterior apenas por atribuir aos indivíduos, admitidos mais recentemente a um grupo, peso igual aos dos demais indivíduos já pertencentes ao grupo (Romesburg, 1984). As matrizes sucessivas de agrupamentos podem ser calculadas em função da matriz de dados originais, atribuindo pesos previamente definidos pelo pesquisador, pela fórmula

$$\sum w_{jk} U_{jk} ,$$

em que w_{jk} é um peso para a distância entre os indivíduos j e k, e U_{jk} é uma medida de dissimilaridade entre j e k.

c) UPGMC (“Unweighted Pair-Group Centroid Method”)

Calcula o centróide dos indivíduos que se unirão para formar os grupos, e a partir desse centróide, a distância. Não produz resultados monotônicos, em função da desigualdade ultramétrica não ser encontrada.

d) WPGMC (“Weighted Pair-Group Centroid Method”)

Neste método, o indivíduo mais recentemente admitido ao grupo tem o mesmo peso dos outros já pertencentes a ele. Tem características similares ao UPGMC, diferindo pelos pesos diferentes dados aos últimos indivíduos admitidos ao grupo, que alteram as classificações.

2.5 Outros métodos de agrupamento

Em situações práticas, além dos métodos SAHN, outras alternativas de análise são possíveis de ser empregadas. Destacam-se dentre elas o método de

otimização de Tocher e a projeção das distâncias no plano bidimensional, que também não fazem pressuposição sobre a matriz de similaridade, apresentados a seguir.

2.5.1 Métodos de otimização

Os métodos de otimização mantêm o princípio de estabelecer grupos de forma que exista homogeneidade dentro dos grupos e heterogeneidade entre os grupos. Eles realizam a partição do conjunto de indivíduos em subgrupos não-vazios e mutuamente exclusivos por meio da maximização ou minimização de uma medida de similaridade preestabelecida (Cruz, 2001).

Entre os métodos de otimização, o mais comumente empregado nas pesquisas genéticas, é o método de otimização de Tocher (Cruz & Regazzi, 1997), como mostra o trabalho de Duarte et al. (1999). Esse método requer a obtenção da matriz de dissimilaridades, sobre a qual será identificado o par de indivíduos mais similares. Estes indivíduos formarão o grupo inicial e a partir daí avalia-se a possibilidade de inclusão de novos indivíduos ao grupo, adotando o critério de que a média das medidas de dissimilaridade dentro de cada grupo seja menor que as distâncias médias entre quaisquer grupos.

A entrada de um indivíduo em um grupo aumenta o valor médio das distâncias dentro do grupo. Assim, pode-se usar o critério da inclusão de um indivíduo no grupo, por meio da comparação entre o acréscimo no valor médio da distância dentro do grupo e um nível máximo permitido, que pode ser estabelecido arbitrariamente ou, como comumente tem sido usado, empregando o valor máximo da medida de dissimilaridade encontrada no conjunto das menores distâncias envolvendo cada indivíduo (Cruz & Regazzi, 1997). Esse método traz como vantagem a fácil interpretação dos grupos formados. Além disso, como o critério de inclusão pode ser arbitrário, esse método traz versatilidade, embora nem sempre seja fácil definir esse critério.

2.5.2 Projeção das similaridades no plano bidimensional

Proposto por Cruz e Viana (1994), neste método as medidas de dissimilaridade são transformadas em escores relativos a duas variáveis X e Y, que serão representados em um gráfico de dispersão, refletindo no espaço bidimensional, as distâncias originalmente obtidas a partir do espaço p-dimensional, sendo p o número de variáveis utilizadas para a obtenção das distâncias (Cruz, 2001).

Para tanto, é necessário inicialmente estabelecer uma ordem decrescente de dissimilaridade entre os pares de indivíduos, sendo que aquele que apresentar o maior valor de dissimilaridade é considerado o primeiro par.

Este método estima coordenadas para cada indivíduo, a partir da matriz de dissimilaridades, por procedimentos estatísticos que minimizam as diferenças entre as distâncias originais e as distâncias obtidas no espaço bidimensional. Permite assim, avaliar a eficiência da projeção obtida, considerando por exemplo, a correlação entre as distâncias originais e as distâncias obtidas pela representação gráfica da dispersão bidimensional, o grau de distorção e o valor do estresse (Cruz, 2001). Essa avaliação da eficiência não é exclusiva para método, sendo também possível realizá-la, por exemplo, para os dendrogramas.

3 MATERIAL E MÉTODOS

3.1 Material

No presente trabalho, foram utilizadas 18 linhagens desenvolvidas por programa de melhoramento de milho do Departamento de Genética – ESALQ/USP, pelo professor Dr. Cláudio Lopes de Souza Jr. As linhagens são provenientes das populações de milho BR-105 e BR-106, desenvolvidas pelo ao Centro Nacional de Milho e Sorgo (Embrapa Milho e Sorgo). Ambas apresentam ciclo precoce e baixa altura da planta (Pinto, 2000). A origem das linhagens utilizadas nesse trabalho é apresentada na Tabela 6, sendo que 8 delas se originam da população BR-105 e 10 de BR-106.

Tabela 6. Origem das 18 linhagens usadas no presente trabalho.

Origem	**Linhagens (código de uso)**									
BR - 105	1	2	3	4	5	6	7	8		
BR - 106	9	10	11	12	13	14	15	16	17	18

Como observado, as linhagens pertencem a duas diferentes populações, sendo então esperado *a priori* que essas linhagens formem dois grupos quando aplicada alguma técnica de agrupamento. Isso pode ser justificado em função da origem diversa dessas duas populações, que reflete na semelhança genética que elas apresentam (Pinto, 2000).

3.2 Métodos

3.2.1 Marcadores moleculares

A amplificação para o marcador RAPD foi efetuada como descrita por Williams et al. (1990). Somente bandas polimórficas foram usadas para a construção da matriz de valores binários, representando a ausência e a presença de bandas por 0 e 1, respectivamente. Cada banda foi considerada como um loco.

O marcador AFLP foi analisado como descrito por Vos et al. (1995). Vinte combinações de primers (iniciadores) foram usadas e a ausência e a presença de bandas foram codificadas por 0 e 1, respectivamente. Cada banda foi considerada como um loco.

3.2.2 Coeficientes de similaridade

Foram obtidas estimativas de similaridade genética (sg_{ij}) entre cada par de linhagens (i, j), para os dois marcadores, por oito coeficientes de similaridade: Jaccard (Jaccard, 1901), Sorensen-Dice (Dice, 1945), Anderberg (Anderberg, 1973), Ochiai (Ochiai, 1957), Simple Matching (Sokal & Michener, 1958), Rogers e Tanimoto (Rogers & Tanimoto, 1960), Ochiai II (Ochiai, 1957) e Russel e Rao (Russel & Rao, 1940), apresentados na Tabela 7.

Tabela 7. Coeficientes de similaridade usados entre as 18 linhagens de milho, para os marcadores AFLP e RAPD.

Coeficientes	**Fórmula***	**Intervalo de Ocorrência**
Jaccard (1901)	$\frac{a}{a+b+c}$	[0, 1]
Sorensen-Dice (1945)	$\frac{2a}{2a+b+c}$	[0, 1]
Anderberg (1973)	$\frac{a}{a+2(b+c)}$	[0, 1]
Ochiai (1957)	$\frac{a}{\sqrt{(a+b)(a+c)}}$	[0, 1]
Simple Matching (1958)	$\frac{a+d}{a+b+c+d}$	[0, 1]
Rogers e Tanimoto (1960)	$\frac{a+d}{a+d+2(b+c)}$	[0, 1]
Ochiai II (1957)	$\frac{ad}{\sqrt{(a+d)(a+c)(d+b)(d+c)}}$	[0, 1]
Russel e Rao (1940)	$\frac{a}{a+b+c+d}$	[0, 1]

* O significado de a, b, c e d é apresentado na Tabela 1.

Esses coeficientes foram escolhidos por serem os mais usados na prática (Duarte et al., 1999), motivo pelo qual estão implementados no NTSYS (Rohlf, 1992) que é um programa computacional muito usado. Além disso, variam num intervalo de fácil interpretação $[0, 1]$. Para melhor compreensão de como é feito o cálculo de tais coeficientes, há no Apêndice 1 um exemplo de seu emprego.

As similaridades obtidas a partir desses coeficientes foram transformadas em medidas de dissimilaridade (dg_{ij}) pela seguinte equação:

$$dg_{ij} = 1 - sg_{ij}\,.$$

Os valores de dg_{ij} foram organizados em matrizes de dimensão 18x18 (número de linhagens), sendo que todas elas satisfazem às pressuposições para a transformação de similaridades em dissimilaridades descritas por Jonhson & Wichern (1988). Os coeficientes de similaridade foram calculados com o programa computacional SAS (Sas Institute, 1992), usando programa apresentado por Victória et al. (2001).

3.2.3 Correlações entre as distâncias

Com o objetivo de verificar qual o grau de relacionamento entre os coeficientes de similaridade, foram calculados os coeficientes de correlação linear de Pearson (r) entre as matrizes geradas pelos oitos coeficientes de similaridade, para os dois marcadores. A expressão usada foi (Iemma, 1992):

$$r = \frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(x_{ij}-\bar{x})(y_{ij}-\bar{y})}{\sqrt{[\sum_{i=1}^{n-1}\sum_{j=2}^{n}(x_{ij}-\bar{x})^2][\sum_{i=1}^{n-1}\sum_{j=2}^{n}(y_{ij}-\bar{y})^2]}}\text{, em que}$$

x_{ij} : medida de similaridade entre os indivíduos i e j, para o coeficiente x;

y_{ij} : medida de similaridade entre os indivíduos i e j, para o coeficiente y;

$$\bar{x} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=2}^{n}x_{ij}\,;$$

$$\bar{y} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=2}^{n}y_{ij}\,.$$

O coeficiente de correlação de Pearson avalia se há relação linear entre as variáveis. No contexto do presente trabalho, para estudar se há alterações na classificação relativa com base nas distâncias genéticas, ou seja, para verificar se os diferentes coeficientes classificam diferentemente as linhagens quanto à divergência, os coeficientes foram também comparados pelo coeficiente de correlação ordinal ou

coeficiente de correlação de Spearman (r_s). Isso é importante, pois para os geneticistas planejarem um cruzamento, é necessária a informação de quais linhagens são mais divergentes. Portanto, se houver mudanças nos postos para os diferentes coeficientes, isso implica que as linhagens classificadas como mais divergentes não repetem isso para todos os coeficientes, gerando dúvidas para o pesquisador na sua escolha.

Sua obtenção foi baseada na ordem (ou posto) dos dados (Iemma, 1992):

$$r_s = 1 - \frac{6\sum d_i^{\,2}}{n(n^2 - 1)}, \text{ em que}$$

n: número de pares de linhagens;

d_i: diferença entre os postos das linhagens.

Para visualização das diferenças entre os coeficientes, foram construídos gráficos de dispersão entre todos eles.

Há no Apêndice 2 um exemplo mostrando como foram calculadas as correlações de Spearman.

3.2.4 Dendrogramas

Foram construídos 8 dendrogramas, para cada marcador, a partir do complemento das medidas de similaridade, para verificar se há alterações nos grupos formados quando são usados diferentes coeficientes de similaridades. Os dendrogramas foram construídos segundo o método das médias aritméticas de grupos não ponderados (UPGMA), que é uma técnica SAHN (Sneath & Sokal, 1973).

Seguindo o recomendado pelo método UPGMA, considerou-se inicialmente as 18 linhagens, que foram então agrupadas de forma sucessiva, baseando-se em sua proximidade. Foram formados assim, pares de indivíduos mais próximos. A cada grupo formado foi necessário reconstruir a matriz de similaridade, definindo a distância entre o novo grupo com cada um dos demais, usando para tanto a média entre os valores individuais de dissimilaridade dos indivíduos de um dos grupos com cada um dos outros

grupos. Esse processo se repetiu até a reunião de todos os indivíduos em um único grupo.

Os passos dos agrupamentos foram então representados graficamente nos chamados dendrogramas, sob a forma de diagramas de árvore nos quais o eixo das abcissas representa os níveis em que os indivíduos foram agrupados e o eixo das ordenadas representa os indivíduos. Os dendrogramas foram construídos usando o programa computacional Statistica (1999), sendo que cada linhagem foi considerada um indivíduo. Os diferentes dendrogramas obtidos foram então comparados usando inspeção visual, de forma análoga ao que foi apresentado por Duarte et al. (1999). Há um exemplo de construção de dendrograma no Apêndice 3.

3.2.4.1 Correlação Cofenética

A correlação cofenética mede o grau de ajuste entre a matriz de similaridade original (matriz S) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Tal correlação foi calculada usando (Bussab et al., 1990):

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})(s_{ij}-\bar{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-\bar{s})^2}}, \text{ em que}$$

c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade;

$$\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}c_{ij};$$

$$\overline{s} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} \, .$$

Nota-se que essa correlação equivale à correlação de Pearson entre a matriz de similaridade original e aquela obtida após a construção do dendrograma. Assim, quanto mais próxima de 1, menor será a distorção provocada pelo agrupamento dos indivíduos com o método UPGMA.

As matrizes cofenéticas de todos dendrogramas foram obtidas com uso do programa NTSYS-PC (Rohlf, 1992), e as correlações entre as matrizes de similaridade originais e as matrizes cofenéticas foram obtidas com o programa computacional Statistica (1999).

Há no Apêndice 4 um exemplo de como essas correlações foram calculadas.

3.2.4.2 Distorção entre a matriz de similaridade e matriz cofenética

O grau da distorção (1 - α) foi calculado por (Kruskal, 1964):

$$\alpha = \frac{\sum_{i=1}^{n-1} \sum_{j=2}^{n} c_{ij}}{\sum_{i=1}^{n-1} \sum_{j=2}^{n} s_{ij}}, \text{ em que}$$

c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade.

Esse parâmetro mede a distorção entre a matriz original e aquela obtida após a construção do dendrograma. Há um exemplo de cálculo no Apêndice 4.

3.2.4.3 Estresse entre a matriz de similaridade e matriz cofenética

O valor do estresse (S), foi calculado por (Kruskal, 1964):

$$S = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(s_{ij} - c_{ij})^2}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} s_{ij}}}$$, em que

c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade;

Esta representação estatística do estresse (soma de quadrados de resíduos padronizados), proposta por Kruskal (1964), é um parâmetro que determina a precisão de ajuste da projeção gráfica. Neste caso, foi usada para determinar a precisão do ajuste obtido com a da projeção da matriz de similaridade no dendrograma. O estresse foi classificado de acordo com os critérios apresentados na Tabela 8 (Kruskal, 1964).

Tabela 8: Classificação do estresse *(S)*.

Nível de estresse (%)	Ajuste
40	Insatisfatório
20	Regular
10	bom
5	Excelente
0	Perfeito

Há no Apêndice 4 um exemplo de seu cálculo.

3.2.4.4 Índice de Consenso

Os dendrogramas foram também comparados usando o Índice de Consenso CI_C (Rohlf, 1982). Esse índice forneceu uma estimativa relativa das similaridades dos dendrogramas, ou seja, informa quantos passos de agrupamentos comuns ocorreram na

construção de dois dendrogramas obtidos a partir de diferentes coeficientes. Tal índice foi obtido pela divisão do número de passos comuns entre dois dendrogramas, excluindo o último passo do agrupamento, em que todos as linhagens foram reunidas em um único grupo, pelo número máximo possível de passos de agrupamento, que é $n-2$ para dendrogramas resolvidos integralmente (n corresponde ao número de linhagens) (Rohlf, 1982).

O índice CI_C é de fácil interpretação, pois é uma proporção entre o número de subconjuntos comuns entre dois dendrogramas e o número total de subconjuntos formado excluindo o conjunto formado por todos as linhagens. Seu valor varia portanto, de 0 (nenhum consenso) a 1 (dendrogramas idênticos). Os índices CI_C entre os dendrogramas foram obtidos usando o programa computacional NTSYS (Rohlf, 1992). Há no Apêndice 4 um exemplo de seu cálculo.

3.2.5 Método de Otimização de Tocher

Foram estabelecidos também os grupos pelo procedimento de otimização de Tocher (Rao, 1952), para os dois marcadores (AFLP e RAPD) e todos os coeficientes. Tal procedimento consistiu em realizar a partição do conjunto dos n indivíduos em subgrupos não-vazios e mutuamente exclusivos, por meio da maximização ou minimização de uma medida preestabelecida (Cruz & Ragazzi, 1994). No caso, foi considerado como limite da distância intergrupo o maior valor do conjunto de menores distâncias envolvendo cada linhagem estudada.

A partir de cada matriz de dissimilaridade, foi identificado o par de indivíduos mais próximos, o qual formou o grupo inicial. Uma vez formado o grupo, calcularam-se as medidas de dissimilaridade entre este grupo e as demais linhagens. O critério de inclusão de novos objetos em um grupo baseou-se no acréscimo da distância média do grupo, pelo cálculo da distância entre a linhagem k e o grupo ij, por meio de:

$$d_{(ij)k} = d_{ik} + d_{jk}\text{, em que}$$

d_{ik}: medida de dissimilaridade entre os indivíduos i e k (no caso, complemento das medidas de similaridade);

d_{jk}: medida de dissimilaridade entre os indivíduos j e k (no caso, complemento das medidas de similaridade);

$d_{(ij)k}$: medida de dissimilaridade entre o grupo ij e o indivíduo k, obtida a parti de d_{ik} e d_{jk}.

Foi usado o seguinte critério para avaliar a possibilidade de inclusão de uma linhagem em um grupo:

$$\text{Se } \frac{d_{(grupo)k}}{n} \leq \alpha \Rightarrow \text{ inclui-se a linhagem } k \text{ ao grupo;}$$

$$\text{Se } \frac{d_{(grupo)k}}{n} > \alpha \Rightarrow \text{ a linhagem } k \text{ não deve ser incluída ao grupo, em que;}$$

α: maior valor do conjunto de menores distâncias envolvendo cada linhagem estudada;

n: número de linhagens que constitui o grupo já formado anteriormente.

O estabelecimento dos grupos pelo método de otimização de Tocher foi obtido usando o programa Genes (Cruz, 1997). No Apêndice 5, há um exemplo de como esses grupos são formados.

3.2.6 Projeção das dissimilaridades no plano bidimensional

Uma outra abordagem usada para verificar a influência da escolha dos coeficientes nas análises de agrupamentos foi a de Cruz e Viana (1994), que consiste em realizar a projeção da matriz de dissimilaridade num espaço bidimensional.

Para estabelecer a coordenadas para cada uma das linhagens, para os dois marcadores e oito coeficientes, inicialmente foi estabelecida uma ordem decrescente de dissimilaridade entre estes. O par de linhagens que apresentou o maior valor de dissimilaridade indicou como tomar os dois primeiros indivíduos para iniciar o processo. Os próximos indivíduos mais divergentes foram aqueles que apresentaram maiores valores de $d_{(ij)k}$, dado por:

$$d_{(ij)k} = d_{ik} + d_{jk}\text{, em que}$$

$d_{(ij)k}$: medida de dissimilaridade entre os indivíduos do grupo ij e o indivíduo k. Este critério foi usado sucessivamente até que se estabelecesse toda a ordem.

As coordenadas para os dois primeiros indivíduos i e j foram formados como $(0,0)$ e $(d_{ij},0)$ respectivamente. As coordenadas para o terceiro indivíduo foram obtidas por:

$$x_k = \frac{d_{jk}^{\,2} - d_{ik}^{\,2} - d_{ij}^{\,2}}{-2d_{ij}} \text{ e}$$

$$y_k = (d_{ik}^{\,2} - x_k^{\,2})^{\frac{1}{2}}\text{, em que}$$

x_k e y_k: abscissa e ordenada, respectivamente, para o indivíduo k.

As coordenadas para o quarto e demais indivíduos sucessivamente, foram calculadas por:

$$C = P^{-1}Q;$$

sendo

$$C = [k\ x_l\ y_l]\text{, em que}$$

k: constante;

x_l: abscissa do indivíduo l;

y_l: ordenada do indivíduo l;

$$P = \begin{bmatrix} m & -2\sum_m x_m & -2\sum_m y_m \\ -2\sum_m x_m & 4\sum_m x_m^{\,2} & 4\sum_m x_m y_m \\ -2\sum_m y_m & 4\sum_m x_m y_m & 4\sum_m y_m^{\,2} \end{bmatrix};$$

$$Q = \begin{bmatrix} \sum d^2_{lm} - \sum (x^2_m + y^2_m) \\ -2\sum x_m d^2_{lm} + 2\sum x_m (x^2_m + y^2_m) \\ -2\sum y_m d^2_{lm} + 2\sum y_m (x^2_m + y^2_m) \end{bmatrix};$$

x_m : abscissa do indivíduo m $(m = i, j, k, ...n)$;

y_m : ordenada do indivíduo m $(m = i, j, k, ...n)$.

Os coeficientes de similaridade, para os dois marcadores, foram comparados quanto à eficiência dessa projeção obtida, de forma análoga ao que foi feito para os dendrogramas. Para tanto, considerou-se (Cruz & Viana, 1994):

a) Correlação entre as distâncias originais e as distâncias obtidas pela representação gráfica da dispersão bidimensional, dada por:

$$r = \frac{\sum_{i=1}^{n-1} \sum_{j=2}^{n} (dg_{ij} - \overline{dg})(do_{ij} - \overline{do})}{\sqrt{[\sum_{i=1}^{n-1} \sum_{j=2}^{n} (dg_{ij} - \overline{dg})^2][\sum_{i=1}^{n-1} \sum_{j=2}^{n} (do_{ij} - \overline{do})^2]}}$$, em que

dg_{ij} : distâncias gráficas entre as linhagens i e j , no espaço bidimensional;

do_{ij} : distâncias originais entre as linhagens i e j , no espaço *n*–dimensional;

$\overline{dg}$: média das distâncias gráficas;

$\overline{do}$; média das distâncias originais.

b) Grau da distorção (1 - α), dado por (Kruskal, 1964):

$$\alpha = \frac{\sum_{i<j} dg_{ij}}{\sum_{i<j} do_{ij}}$$, sendo

dg_{ij} : distâncias gráficas entre as linhagens i e j , no espaço bidimensional;

do_{ij} : distâncias originais entre as linhagens i e j , no espaço *n*–dimensional.

c) Valor do estresse *(S)*, dado por (Kruskal, 1964):

$$S = \sqrt{\frac{\sum_{i<j} (do_{ij} - dg_{ij})^2}{\sum_{i<j} do^2_{ij}}}$$, em que

dg_{ij} : distâncias gráficas entre as linhagens i e j, no espaço bidimensional;

do_{ij} : distâncias originais entre as linhagens i e j, no espaço *n*–dimensional.

O estresse foi classificado de acordo com os critérios apresentados na Tabela 8, já apresentada anteriormente (Kruskal, 1964).

As coordenadas para as projeções no plano bidimensional, os cálculos das correlações entre as distâncias originais e as distâncias estimadas, o grau de distorção das projeções e o valor do estresse foram obtidos com o auxílio do programa Genes (Cruz, 1997), e os gráficos das projeções foram feitos pelo programa computacional Statistica (1999).

No Apêndice 6, é apresentado um exemplo de como esses cálculos foram obtidos.

4 RESULTADOS E DISCUSSÃO

4.1 Marcadores Moleculares

Os resultados dos marcadores AFLP e RAPD, aplicados as 18 linhagens, usados para obtenção dos coeficientes de similaridade, são apresentados na Tabela 9. O número de bandas polimórficas obtido foi considerado suficiente para estudos de divergência, como no presente caso (Garcia, et al., 2002).

Tabela 9. Número bandas polimórficas e alelos por loco, obtidos pelos marcadores AFLP e RAPD para 18 linhagens de milho das populações BR-105 e BR-106.

	Marcador	
Parâmetros	**AFLP**	**RAPD**
Número de bandas polimórficas	774	262
Número médio de alelos por locos	2,0	2,0

4.2 Coeficientes de similaridade

Os coeficientes de similaridade entre as linhagens, com os dois marcadores, usando os critérios de Jaccard, Sorensen-Dice, Anderberg, Ochiai, Simple Matching, Rogers e Tanimoto, Ochiai II e Russel e Rao são apresentados nas Tabelas 10 a 17, no anexo.

Fica evidente que uma comparação dos coeficientes com base nesses dados brutos é muito difícil de ser feito, em função do volume elevado de informações.

Os menores e os maiores valores, as médias e a amplitude para os oito coeficientes de similaridade, entre as 18 linhagens de milho, para os marcadores AFLP e RAPD, estão apresentados na Tabela 18. Para o marcador AFLP, todos os coeficientes, exceto Russel e Rao, apresentaram o maior valor de similaridade para as linhagens 11 e 12. Para este último coeficiente, este valor foi observado para entre as linhagens 17 e 18. Os menores valores de similaridade foram apresentado pelo mesmo par de linhagem (8 e 12) para os coeficientes de Jaccard, Sorensen-Dice, Anderberg, Ochiai e Ochiai II. Para os coeficientes Simple Matching e de Rogers e Tanimoto, o par que apresentou o menor valor de similaridade foi composto pelas linhagens 4 e 12, sendo que para Russel e Rao, isto ocorreu entre as linhagens 8 e 9. Inclusive a matriz de similaridades obtida a partir do coeficiente de Russel e Rao pode apresentar valores na diagonal diferentes de 1, o que não é observado para os demais coeficientes. Isso se deve ao fato desse coeficiente considerar a ausência conjunta no denominador e não no numerador (resultados não apresentados).

Para o marcador RAPD, o maior valor de similaridade foi apresentado pelo par composto das linhagens 17 e 18, para todos os coeficientes. O menor valor de similaridade foi apresentado pelo par formado pelas linhagens 1 e 5 para os coeficientes de Jaccard, Sorensen-Dice, Anderberg, Ochiai e Russel e Rao; pelo par de linhagens 2 e 13 para os coeficientes de Simple Matching e Rogers e Tanimoto, e pelas linhagens 13 e 16 para Ochiai II.

Observando as médias dos valores de similaridade entre as linhagens, nota-se que há diferenças. Para alguns coeficientes, esses valores apresentam-se maiores do que para outros, como para os coeficientes Simple Matching e de Sorensen-Dice, para os dois marcadores. Contudo, não existe um padrão para comparar se os valores obtidos por estes coeficientes é baixo ou alto (Bussab et al., 1990). Além disso, os coeficientes dão diferentes pesos aos valores das ocorrências e ausências conjuntas, bem como as diferenças encontradas, o que influencia diretamente nos valores encontrados. Contudo, o que realmente é importante observar é se os diferentes coeficientes mantêm as mesmas

ordens de semelhanças, quando classificam as linhagens. A amplitude de variação das similaridades foi praticamente as mesmas para todos os coeficientes, exceto para Russel e Rao, com RAPD.

Tabela 18. Menores valores, maiores valores, médias e amplitude para os coeficientes de similaridade de Jaccard, Sorensen-Dice, Anderberg, Ochiai, Simple Matching, Rogers e Tanimoto, Ochiai II e Russel e Rao, entre 18 linhagens de milho das populações BR-105 e BR-106, para os marcadores AFLP e RAPD.

Coeficientes		**Mínimo**	**linhagens**	**Máximo**	**linhagens**	**Amplitude**	**Média**
Jaccard	**AFLP**	0,2984	8 e 12	0,9148	11 e 12	0,6164	0,3930
	RAPD	0,2327	1 e 5	0,8864	17 e 18	0,6537	0,4194
Sorensen-Dice	**AFLP**	0,4597	8 e 12	0,9555	11 e 12	0,4958	0,5643
	RAPD	0,3776	1 e 5	0,9398	17 e 18	0,5622	0,5909
Anderberg	**AFLP**	0,1754	8 e 12	0,8430	11 e 12	0,6676	0,2446
	RAPD	0,1317	1 e 5	0,7959	17 e 18	0,6642	0,2653
Ochiai	**AFLP**	0,4605	8 e 12	0,9555	11 e 12	0,4950	0,5676
	RAPD	0,3800	1 e 5	0,9398	17 e 18	0,5599	0,5924
Simple Matching	**AFLP**	0,5388	4 e 12	0,9599	11 e 12	0,4211	0,6123
	RAPD	0,5038	2 e 13	0,9427	17 e 18	0,4389	0,6450
Rogers e Tanimoto	**AFLP**	0,3687	4 e 12	0,9230	11 e 12	0,5543	0,4412
	RAPD	0,3367	2 e 13	0,8917	17 e 18	0,5550	0,4761
Ochiai II	**AFLP**	0,2773	8 e 12	0,9207	11 e 12	0,6434	0,3646
	RAPD	0,2257	13 e 16	0,8886	17 e 18	0,6629	0,4058
Russel e Rao	**AFLP**	0,1886	8 e 9	0,4315	17 e 18	0,6544	0,2545
	RAPD	0,1412	1 e 5	0,4466	17 e 18	0,3053	0,2557

Nota-se novamente que é difícil avaliar o comportamento dos coeficientes diretamente nas matrizes de similaridades ou usando estatísticas como a média, o que por si já justifica o uso das técnicas de agrupamentos, principalmente para verificar quais linhagens são mais parecidas entre si.

4.3 Correlações

A partir dos gráficos de dispersão entre os coeficiente para o marcador AFLP (Figura 1) , observa-se que os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, mostraram uma relação linear entre si. O mesmo aconteceu com os coeficientes Simple Matching, Rogers e Tanimoto e Ochiai II que, em relação aos coeficientes anteriores, apresentaram-se bastante diferentes, mas foram semelhantes entre si. O coeficiente de Russel e Rao foi o que apresentou as maiores diferenças em relação a todos os outros coeficientes. O mesmo ocorreu quando usado os dados do marcador RAPD (Figura 2), também com Russel e Rao diferindo dos demais.

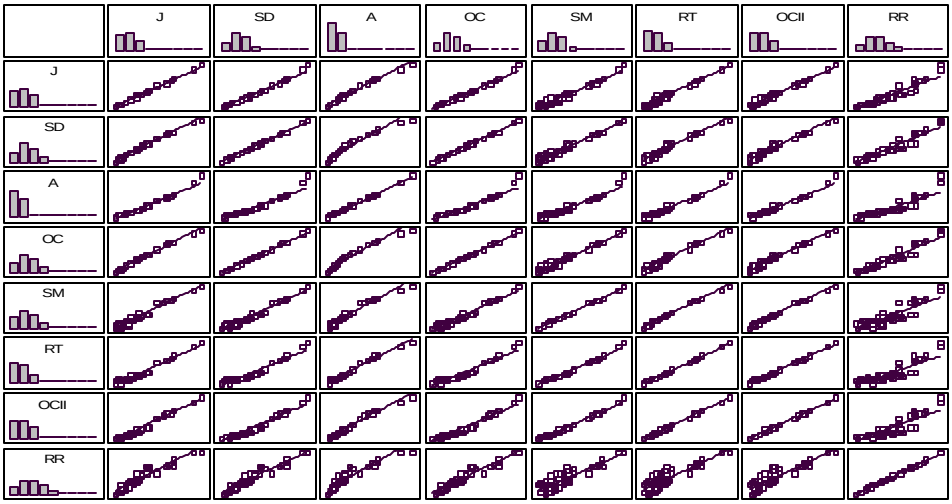

Figura 1 - Gráfico da dispersão entre os coeficientes de Jaccard (J), Sorensen-Dice (SD), Anderberg (A), Ochiai (O), Simple Matching (SM), Rogers e Tanimoto (RT), Ochiai II (OII) e Russel e Rao, calculados a partir do marcador AFLP, entre 18 linhagens de milho das populações BR-105 e BR-106.

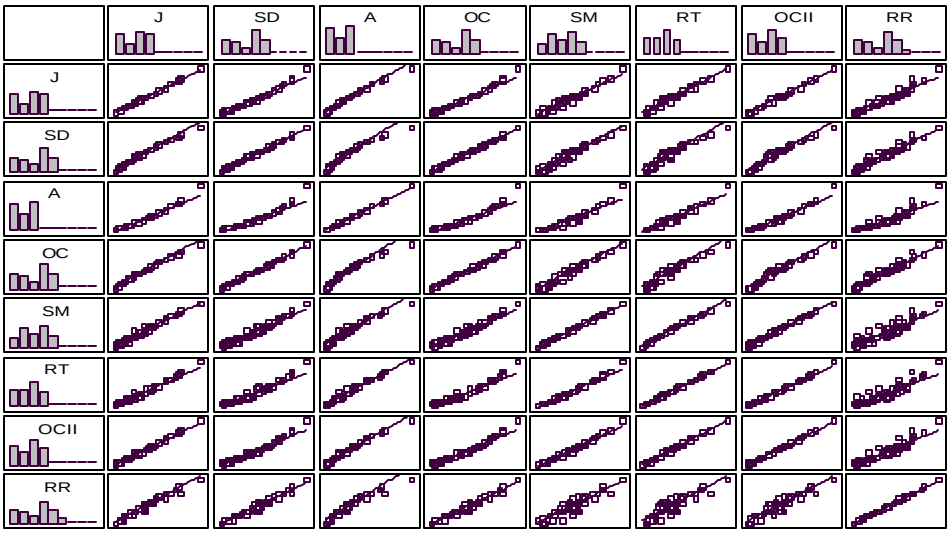

Figura 2 - Gráfico da dispersão entre os coeficientes de Jaccard (J), Sorensen-Dice (SD), Anderberg (A), Ochiai (O), Simple Matching (SM), Rogers e Tanimoto (RT), Ochiai II (OII) e Russel e Rao, calculados a partir do marcador RAPD, entre 18 linhagens de milho das populações BR-105 e BR-106.

Para ambos marcadores, os coeficientes de correlação de Pearson, que medem justamente essa tendência de associação linear entre os coeficientes, foram todos próximos de 1, deixando evidente que as medidas são linearmente relacionadas, confirmando os diagramas de dispersão (Tabela 19). Porém, observa-se que para o coeficiente de Russel e Rao, os valores da correlação com os demais são levemente inferiores, principalmente para os coeficientes de Simple Matching, Rogers e Tanimoto e Ochiai II.

Contudo, não se observou clara tendência, para os dois marcadores, de maior semelhança entre os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, e entre Simple Matching, Rogers e Tanimoto e Ochiai II, como observados nos diagramas dispersão. Isso mostra ainda que o coeficiente de correlação de Pearson não é um bom

parâmetro para comparar as medidas de similaridade, por não detectar pequenas diferenças.

Tabela 19. Coeficiente de correlação de Pearson entre os coeficientes de similaridades para dados do marcador AFLP (acima da diagonal) e marcador RAPD (abaixo da diagonal)* (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

	J	**SD**	**A**	**O**	**SM**	**RT**	**OII**	**RR**
J	-	0,9939	0,9914	0,9934	0,9721	0,9759	0,9874	0,9419
SD	0,9928	-	0,9715	0,9999	0,9673	0,9615	0,9770	0,9567
A	0,9895	0,9655	-	0,9705	0,9611	0,9763	0,9831	0,9115
O	0,9930	0,9999	0,9659	-	0,9653	0,9592	0,9754	0,9588
SM	0,9736	0,9677	0,9613	0,9672	-	0,9955	0,9948	0,8549
RT	0,9723	0,9547	0,9741	0,9546	0,9949	-	0,9976	0,8493
OII	0,9912	0,9798	0,9852	0,9800	0,9928	0,9939	-	0,8809
RR	0,9686	0,9759	0,9420	0,9761	0,8985	0,8865	0,9288	-

* Todos os valores são significativamente diferentes de zero ($P < 0,05$).

Os valores dos coeficientes de correlação de Sperman entre os oito coeficientes de similaridades, para ambos marcadores, também foram todas altas, mostrando que há uma forte relação entre eles, com poucas inversões nos postos (Tabela 20). Os coeficientes de Jaccard, Sorensen-Dice e Anderberg, apresentaram os valores da correlação igual a 1 entre si, indicando que não há nenhuma mudança nos postos para tais coeficientes, ou seja, eles classificam a similaridade entre as linhagens na mesma ordem. O coeficiente de Ochiai apresentou valores da correlação praticamente igual a 1 em relação a esses coeficientes. Portanto, os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, mostraram-se altamente relacionados. A mesma situação ocorreu entre os coeficientes de Simple Matching e Rogers e Tanimoto que também

apresentaram o valor da correlação igual ou próximo de 1. Da mesma forma que anteriormente, esses dois grupos de coeficientes diferiram ligeiramente entre si.

O coeficiente de Russel e Rao, por sua vez, obteve valores de correlação levemente inferiores com os demais coeficientes, principalmente em relação aos coeficientes de Simple Matching, Rogers e Tanimoto e Ochiai II.

Novamente, há indícios que os coeficientes podem ser divididos em três gupos. O primeiro grupo seria composto pelos coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, e o segundo, pelos coeficientes de Simple Matching, Rogers e Tanimoto e Ochiai II. O coeficiente de Russel e Rao parece não se enquadrar em nenhum desses grupos.

Verifica-se contudo que a correlação de Spearman também não permite uma clara distinção entre os coeficientes.

Tabela 20. Coeficiente de correlação de Spearman entre os coeficientes de similaridades para dados do marcador AFLP (acima da diagonal) e marcador RAPD (abaixo da diagonal)* (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficiente	**J**	**SD**	**A**	**O**	**SM**	**RT**	**OII**	**RR**
J	-	1,0000	1,0000	0,9996	0,9103	0,9103	0,9590	0,9463
SD	1,0000	-	1,0000	0,9996	0,9103	0,9103	0,9590	0,9464
A	1,0000	1,0000	-	0,9996	0,9103	0,9103	0,9590	0,9463
O	0,9997	0,9997	0,9997	-	0,9067	0,9067	0,9568	0,9490
SM	0,9531	0,9531	0,9531	0,9518	-	1,0000	0,9871	0,7443
RT	0,9531	0,9531	0,9531	0,9518	1,0000	-	0,9871	0,7443
OII	0,9804	0,9804	0,9804	0,9798	0,9906	0,9906	-	0,8279
RR	0,9581	0,9581	0,9581	0,9590	0,8500	0,8500	0,8962	-

* Todos os valores são significativamente diferentes de zero ($P < 0,05$).

Analisando a natureza desses coeficientes. A partir de suas expressões (Tabela 7), pode-se perceber que os coeficientes que tem propriedades comuns, mostraram-se relacionados. Como observa-se para os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, eles têm a propriedade comum de não considerarem a ausência conjunta. Os coeficientes de Simple Matching, Rogers e Tanimoto e Ochiai II, por sua vez, consideram a ausência conjunta. O coeficiente de Russel e Rao é um coeficiente que considera a ausência conjunta no denominador mas não no numerador, o que talvez explique comportamento diverso dos demais.

4.4 Dendrogramas

4.4.1 Inspeção visual

Uma inspeção visual dos dendrogramas pode ser feita com base nas Figuras 3 e 4. De forma geral, os dendrogramas apresentaram estruturas de agrupamento similares. Para os dendrogramas obtidos a partir do marcador AFLP, observa-se que eles foram capazes de separar as linhagens das populações BR-105 e BR-106 nos dois grupos que eram previamente conhecidos, conforme apresentado em Material e Métodos (linhagens 1 a 8, BR-105; linhagens 9 a 18, BR-106). A exceção refere-se à linhagem 16, pertencente originalmente à população BR-106, que agrupou-se com um maior nível de similaridade com as linhagens da população BR-105. Trabalhos recentes, contudo, mostram que a linhagem 16 apresenta-se mais semelhante ao grupo contrastante (BR-105) (Benchimol et al., 2000; Barbosa, 2001; Pinto et al., 2001), confirmando o que foi observado nos dendrogramas e mostrando portanto que esse método de análise fornece resultados esperados *a priori*.

Embora a estrutura geral dos agrupamentos seja bastante parecida, pode-se observar que há pequenas alterações nos níveis em que as linhagens são agrupadas, ou seja, as linhagens que estão dentro de um mesmo grupo podem ser agrupadas em outra ordem, quando se mudam os coeficientes. Entretanto, isso causa poucos problemas práticos.

Ainda para o marcador AFLP, observa-se que os dendrogramas para os coeficientes de Jaccard, Sorensen-Dice, Anderberg, Ochiai, Simple Matching, Rogers e Tanimoto e Ochiai II apresentam praticamente as mesmas estruturas de agrupamentos. As maiores diferenças encontradas no dendrograma foram obtidas para o coeficiente de Russel e Rao, em que a linhagem 9 apresenta-se distante das demais linhagens, além do fato da linhagem 10 estar no grupo junto com as linhagens d1 a 8.

Embora não exista um critério objetivo para determinar um ponto de corte no dendrograma, ou seja, para determinar quais os grupos foram formados, nota-se que, de forma geral, as estruturas de agrupamentos dos dendrogramas para o marcador RAPD, apresentam diferenças em relação as estruturas apresentadas pelos dendrogramas a partir do marcador AFLP. Com esse marcador, todos os dendrogramas exceto aquele obtido usando o coeficiente de Russel e Rao, mostram a formação de três grupos: o primeiro, formado pelos linhagens 1, 10, 12 e 13; o segundo, formado pelas linhagens 2, 3, 4, 5, 6, 7, 8 e 16; o terceiro, pelas linhagens 9, 11, 14, 15, 17 e 18. Esses resultados são coerentes com o que foi apresentado por Lanza et al. (1997). Isso deixa evidente que diferentes marcadores influenciam na formação da estrutura de agrupamento. Contudo, para cada marcador separadamente, todos os coeficientes mostraram resultados muito semelhantes, exceto para o coeficiente de Russel e Rao.

A mesma situação já observada nos dendrogramas obtidos a partir dos oito coeficientes e marcador AFLP, foi portanto também apresentada para o marcador RAPD, ou seja, as estruturas de agrupamento mostradas para os coeficientes de Jaccard, Sorensen-Dice, Anderberg, Ochiai, Simple Matching, Rogers e Tanimoto e Ochiai II, foram novamente praticamente as mesmas, sendo que o coeficiente de Russel e Rao apresenta estruturas de agrupamentos muito diferentes em relação aos demais coeficientes. Para esse último, as linhagens 2, 4 e 5, pertencentes originalmente à população BR-105, mostraram-se próximas as linhagens 11, 14, 15, 17 e 18, originalmente pertencentes a população BR-106. Por sua vez, essas últimas linhagens apresentaram-se distantes das linhagens 10, 12 e 13 também pertencentes à população BR-106. A linhagem 1 mostrou-se muito distante das demais linhagens, o que não foi observado para os demais coeficientes.

Os resultados obtidos com o marcador AFLP foram mais coerentes com a natureza biológica dos dados do que os resultados observados para o marcador RAPD, concordando com os resultados encontrados por Garcia et al. (2001), de que o marcador AFLP apresenta-se mais eficiente em relação ao marcador RAPD para classificação das linhagens.

É importante destacar que o fato desse tipo de análise não apresentar um critério objetivo para identificação dos grupos dificulta muito a interpretação dos resultados. No presente caso, isso só foi possível porque os grupos eram conhecidos *a priori*, o que nem sempre ocorre na prática.

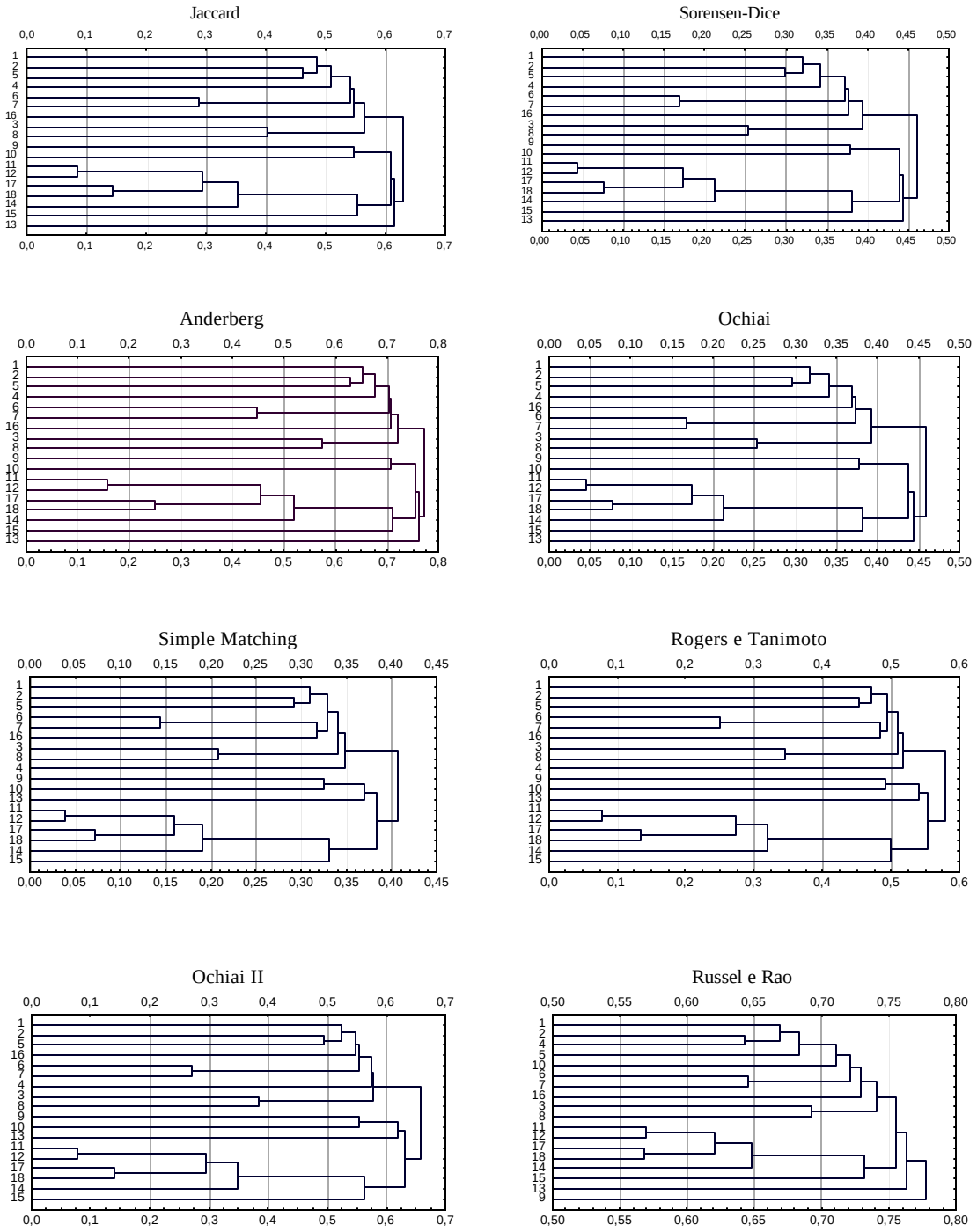

Figura 3 - Dendrogramas para 18 linhagens de milho das populações BR-105 e BR-106, obtidos a partir das matrizes do complemento dos coeficientes de similaridade. Marcador molecular AFLP, método UPGMA.

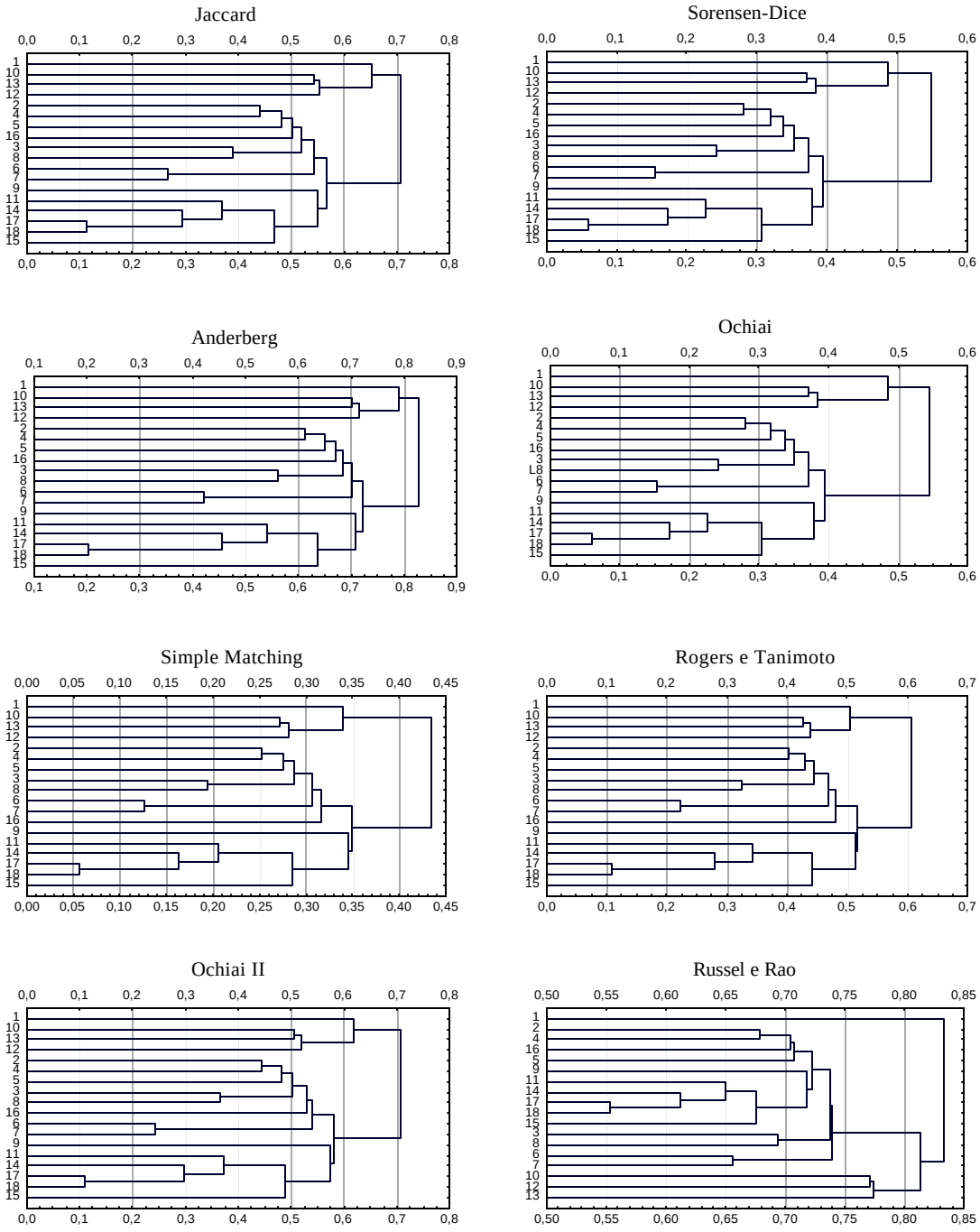

Figura 4 - Dendrogramas para 18 linhagens de milho das populações BR-105 e BR-106, obtidos a partir das matrizes do complemento dos coeficientes de similaridade. Marcador molecular RAPD, método UPGMA.

4.4.2 Correlações cofenéticas

Os valores das correlações cofenéticas (Tabela 20) foram todas de magnitude elevada, para ambos marcadores, exceto o Russel e Rao para o marcador AFLP (r = 0,8630). Isso mostra que há uma boa representação das matrizes de similaridade na forma de dendrogramas e que isso independe do coeficiente usado. Porém, os valores foram ligeiramente superiores para o marcador AFLP, exceto para os coeficientes de Jaccard e de Russel e Rao.

Para o marcador AFLP, observa-se que o coeficiente que apresenta o maior valor da correlação cofenética é o de Rogers e Tanimoto, e o que apresenta o menor valor é o coeficiente de Russel e Rao. Para o marcador RAPD, o coeficiente que apresenta o maior valor da correlação cofenética é o coeficiente de Anderberg e o que apresenta o menor valor é o coeficiente de Sorensen-Dice. Observa-se que as diferenças nos valores das correlações são muito pequenas. Assim, da mesma forma que os coeficientes de correlação de Pearson e Spearman, a correlação cofenética não permite fazer uma clara distinção entre os coeficientes, quanto aos dendrogramas obtidos.

Tabela 20. Correlações de Pearson entre as matrizes cofenéticas e as matrizes de similaridade para 18 linhagens de milho das populações BR-105 e BR-106 (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficientes	J	SD	A	O	SM	RT	OII	RR
AFLP	0,9488	0,9272	0,9671	0,9258	0,9597	0,9712	0,9682	0,8634
RAPD	0,9519	0,8904	0,9624	0,9406	0,9413	0,9507	0,9498	0,9512

4.4.3 Distorção entre a matriz de similaridades e a matriz cofenética

Conforme mostra a Tabela 21, para ambos marcadores, o grau distorção foi igual a zero para todos os coeficientes, com exceção para o marcador RAPD e coeficiente de similaridade de Sorensen-Dice, que apresentou 0,59 (%) de distorção. Isso confirma o que já tinha sido observado na correlação cofénetica, de que há uma boa representação das matrizes de similaridade na forma de dendrogramas e que isso independe do coeficiente usado e do marcador.

Tabela 21. Grau de distorção (%) entre a distância original e a obtida através dos dendrogramas, para os marcadores moleculares AFLP e RAPD (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficientes	J	SD	A	O	SM	RT	OII	RR
AFLP	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00
RAPD	0,00	0,59	0,00	0,00	0,00	0,00	0,00	0,00

4.4.4 Estresse entre a matriz de similaridade e a matriz cofenética

Os níveis de estresses apresentados para os oito coeficientes (Tabela 22), para ambos marcadores, foram de baixa magnitude. Para o marcador AFLP, o nível de estresse variou de 3,21% para o coeficiente Simple Matching a 8,78% para o coeficiente de Russel e Rao. Para o marcador RAPD, o nível de estresse variou de 3,99% para o coeficiente de Simple Matching a 9,34% para o coeficiente de Anderberg. Os valores do estresse sempre foram maiores para o marcador AFLP, com exceção do coeficiente de Russel e Rao.

De acordo com a classificação de Kruskal (1964) (Tabela 8), para o marcador AFLP, os valores do estresse foram considerados excelentes para os coeficientes de Simple Matching e Rogers e Tanimoto, e bom para os demais coeficientes. Para o marcador RAPD, os valores de estresse foi considerado excelente para Simple Matching, e bom para os demais coeficientes. Isso reitera o fato da representação das matrizes de similaridade em dendrogramas ter sido eficiente

Tabela 22. Valor de estresse (%) obtido para a representação dos dendrogramas para os marcadores moleculares AFLP e RAPD (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficientes	J	SD	A	O	SM	RT	OII	RR
AFLP	7,29	5,50	8,50	5,55	3,21	4,45	6,41	8,78
RAPD	8,06	8,93	9,34	6,29	3,99	5,61	8,30	7,21

4.4.5 Índice de consenso

A comparação dos dendrogramas gerados, usando o valor do índice de consenso CI_c, permite um refinamento do que tinha sido observado através da inspeção visual (Tabela 23). Diferentes coeficientes provocam mudanças nas estruturas dos agrupamentos. A amplitude desse índice vai de 0 a 1, sendo os coeficientes considerados idênticos quando o valor de CI_c entre eles for igual a 1.

Para o marcador AFLP, os dendrogramas obtidos pelos coeficientes de Jaccard, Sorensen-Dice e Anderberg foram idênticos (CI_c = 1,0000) o mesmo acontecendo para os coeficientes de Simple Matching e Rogers e Tanimoto. O coeficiente de Ochiai origina dendrograma com estrutura mais parecida com as dos coeficientes de Jaccard, Sorensen-Dice e Anderberg (CI_c = 0,9375), do que com as dos

coeficientes de Simple Matching, e Rogers e Tanimoto (CI_c = 0,7500). Assim, os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai podem ser considerados semelhantes entre si. O coeficiente de Ochiai II, por sua vez, origina dendrograma mais parecido com os originados a partir dos coeficientes Simple Matching e Rogers e Tanimoto (CI_c = 0,8750), do que em relação aos dendrogramas obtidos com Jaccard, Sorensen-Dice, Anderberg e Ochiai (CI_c = 0,8125). Dessa forma, pode-se considerar que os coeficientes de Simple-Matching, Rogers e Tanimoto e Ochiai II são mais parecidos entre si que com relação aos demais. O coeficiente de Russel e Rao apresenta valores baixos de coincidências nas estruturas dos agrupamentos em relação aos demais coeficientes, chegando a valores muito baixos com os coeficientes Simple-Matching, Rogers e Tanimoto e Ochiai II (CI_c = 0,4375). Isso reitera o fato dele originar dendrograma diferente dos demais.

Para o marcador RAPD, os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, apresentaram estruturas de agrupamentos idênticas (CI_c = 1,0000). O mesmo acontece para os coeficientes Simple Matching e Rogers e Tanimoto, sendo que o coeficiente de Ochiai II apresenta o mesmo valor do índice (CI_c = 0,9375) com os demais coeficientes, com exceção do coeficiente de Russel e Rao, que novamente apresenta valores baixos de coincidências entre as estruturas de agrupamentos. Então, houve praticamente as mesmas coincidências que para o marcador AFLP, exceto pelo fato de Ochiai II não estar mais próximo de Simple Matching e Rogers e Tanimoto que dos demais.

Nota-se que os resultados obtidos com o índice de consenso entre os dendrogramas, para os dois marcadores, permitem um maior detalhamento do que foi observado com a inspeção visual. Assim, nota-se que os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai apresentam tendência de possuírem estruturas de agrupamento muito semelhantes, o que também ocorreu para os coeficentes Simple Matching, Rogers e Tanimoto e Ochiai II. O coeficiente de Russel e Rao mostrou-se diferente dos outros.

Tabela 23. Índice de consenso entre os dendrogramas gerados pelos complementos dos coeficientes de similaridade, entre 18 linhagens de milho das populações BR-105 e BR-106. Acima da diagonal, marcador AFLP ; abaixo da diagonal, marcador RAPD (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficientes	**J**	**SD**	**A**	**O**	**SM**	**RT**	**OII**	**RR**
J	-	1,0000	1,0000	0,9375	0,7500	0,7500	0,8125	0,5000
SD	1,0000	-	1,0000	0,9375	0,7500	0,7500	0,8125	0,5000
A	1,0000	1,0000	-	0,9375	0,7500	0,7500	0,8125	0,5000
O	1,0000	1,0000	1,0000	-	0,7500	0,7500	0,8125	0,5000
SM	0,8750	0,8750	0,8750	0,8750	-	1,0000	0,8750	0,4375
RT	0,8750	0,8750	0,8750	0,8750	1,0000	-	0,8750	0,4375
OII	0,9375	0,9375	0,9375	0,9375	0,9375	0,9375	-	0,4375
RR	0,6875	0,6875	0,6875	0,6875	0,6250	0,6250	0,6250	-

4.5 Método de otimização

Outro método de agrupamentos usado, o método de otimização de Tocher, citado por Rao (1952), tem seus resultados apresentados nas Tabelas 24 e 25 Em tal método, o número de grupos formados variou de 4 a 7, para o marcador AFLP, e de 3 a 5, para o marcador RAPD.

Os resultados desse método, usando o marcador AFLP, concordam com o observado pelos dendrogramas, considerando o índice de consenso, ou seja, os coeficientes se dividem em três conjuntos. O primeiro, formado pelos coeficientes Jaccard, Sorensen-Dice, Anderberg e Ochiai, em que o número e a estrutura dos grupos formados foi exatamente a mesma. Para estes coeficientes os grupos formados foram compostos pelas seguintes linhagens : grupo I: 1, 2, 10, 11, 12, 14, 15, 17 e 18; grupo II: 3, 4, 5, 6, 7, 8 e 16; grupo II: 9 e grupo IV: 13. O segundo, formado pelos coeficientes de

Simple Matching, Rogers e Tanimoto e Ochiai II, sendo que os dois primeiros apresentaram os mesmos grupos, formados pelas seguintes linhagens: grupo I: 11, 12, 14, 15, 17 e 18; grupo II 1, 2, 3, 5, 6, 7, 8 e 16; grupo III: 4; grupo IV: 9 e 10 e grupo V: 13. O coeficiente de Ochia II apresentou apenas uma pequena diferença em relação aos anteriores, pois a linhagem 4 pertence ao grupo II, sendo que para os dois coeficientes anteriores ela não se agrupou as demais linhagens. Novamente, o coeficiente de Russel e Rao formou grupos muito diferentes em relação as demais, sendo o grupo I composto pelas linhagens 1, 2, 4, 5, 10, 11, 12, 14, 17 e 18; o grupo II pelas linhagens 6 e 7; o grupo III pelas linhagens 3 e 8; o grupo IV pela linhagem 9; o grupo V pela linhagem 13; o grupo VI pela linhagem 15; o grupo VII pela linhagem 16. Nota-se claramente que o coeficiente de Russel e Rao formou o maior número de grupos.

Observa-se que os resultados obtidos para o marcador AFLP, com o método de otimização de Tocher, não foram exatamente os mesmos que aqueles observados nos dendrogramas. Para os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, o grupo I incluiu as linhagens 1 e 2, que deveria estar no grupo II. Isso não ocorreu no dendrograma, além disso, as linhagens 9 e 13 não se agruparam. Para os coeficientes Simple Matching, Rogers e Tanimoto e Ochiai II isso não ocorreu com as linhagens 1 e 2, mas houve tendência da linhagem 4 não se agrupar (exceto para Ochiai II). O Coeficiente de Ochiai II foi portanto, quem apresentou para o método de otimização os resultados mais semelhantes ao dendrograma. Novamente, Russel e Rao foi diferente dos demais.

Para o marcador RAPD, observa-se que os grupos formados a partir dos coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, foram exatamente os mesmos. Contudo, os grupos formados para esse marcador foram diferentes dos formados pelo marcador AFLP, o que é justificado pelas diferentes propriedades dos marcadores (Garcia et al., 2001). Para tais coeficientes, os grupos formados foram: grupo I: linhagens 2, 3, 4, 5, 6, 7, 8, 9, 11, 14, 15, 16, 17 e 18; grupo II: linhagens 10, 12 e 13; grupo III: linhagem 1. Esse resultado foi coerente com os dendrogramas para esses coeficientes, pois nota-se que, dependendo do ponto de corte escolhido para determinar os grupos, as linhagens 2, 3, 4, 5, 6, 7, 8, 9, 11, 14, 15, 16, 17 e 18, podem ser

consideradas como pertencentes ao mesmo grupo, e a linhagem 1 apresenta-se a distante em relação as demais, sendo a última a ser agrupada (Figura 4). As linhagens 10, 12 e 13 mostram-se como um grupo isolado das demais.

Os coeficientes de Simple Matching e Rogers e Tanimoto, para o marcador RAPD, também formaram grupos iguais para o método de otimização de Tocher. Os grupos formados foram: grupo I: linhagens 11, 14, 15, 17 e 18; grupo II: linhagens 2, 3, 4, 5, 6, 7, 8 e 16; grupo III: linhagens 10, 12 e 13; grupo IV: linhagem 1; grupo V: linhagem 9. Também para esses coeficientes, os grupos formados mostram-se bastante coerentes com os dendrogramas, dependendo de onde se considerar o ponto de corte. O coeficiente de Ochiai II, por sua vez, apresentou resultados parecidos com os obtidos com os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, exceto para a linhagem 9, que não se agrupou.

O coeficiente de Russel e Rao formou apenas três grupos, sendo que o primeiro agrupou quase todas as linhagens com a exceção das linhagens 1 e 13 que não se agruparam. Esses grupos foram ligeiramente diferentes dos apresentados no dendrograma, pois nele as linhagens 10, 12 e 13, apresentam-se próximas.

De forma geral, houve novamente a classificação dos coeficientes em três categorias, uma delas composta pelos coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, outra por Simple Matching, Rogers e Tanimoto, e Ochiai II e finalmente, uma por Russel e Rao. A exceção foi observada para ochiai II com o marcador RAPD, mas isso pode ser explicado por problemas inerentes a esse marcador (Garcia et al., 2001).

Tabela 24. Agrupamentos das 18 linhagens de milho das populações BR-105 e BR-106 pelo método de otimização de Tocher, considerando marcador molecular AFLP (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Grupos	Coeficientes de similaridade							
	J	**SD**	**A**	**O**	**SM**	**RT**	**OII**	**RR**
I	1 2 10 11 12 14 15 17 18	1 2 10 11 12 14 15 17 18	1 2 10 11 12 14 15 17 18	1 2 10 11 12 14 15 17 18	11 12 14 15 18 17	11 12 14 15 18 17	11 12 14 15 18 17	1 2 4 5 10 11 12 14 17 18
II	3 4 5 6 7 8 16	3 4 5 6 7 8 16	3 4 5 6 7 8 16	3 4 5 6 7 8 16	1 2 3 5 6 7 8 16	1 2 3 5 6 7 8 16	1 2 3 4 5 6 7 8 16	6 7
III	9	9	9	9	4	4	9 10	3 8
IV	13	13	13	13	9 10	9 10	13	9
V					13	13		13
VI								15
VII								16

Tabela 25. Agrupamentos das 18 linhagens de milho das populações BR-105 e BR-106 pelo método de otimização de Tocher, considerando marcador molecular RAPD (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Grupos	**Coeficientes de similaridade**							
	J	**SD**	**A**	**O**	**SM**	**RT**	**OII**	**RR**
I	2 3 4 5 6 7 8 9 14 11 15 16 17 18	2 3 4 5 6 7 8 9 14 11 15 16 17 18	2 3 4 5 6 7 8 9 14 11 15 16 17 18	2 3 4 5 6 7 8 9 14 11 15 16 17 18	11 14 15 17 18	11 14 15 17 18	2 3 4 5 6 7 8 11 14 15 16 17 18	2 3 4 5 6 7 8 9 10 11 12 14 15 16 17 18
II	10 12 13	10 12 13	10 12 13	10 12 13	2 3 4 5 6 7 8 16	2 3 4 5 6 7 8 16	10 12 13	1
III	1	1	1	1	10 13 12	10 13 12	1	13
IV					1	1	9	
V					9	9		

Verifica-se ainda que este tipo de análise não necessariamente forma os mesmos grupos que os dendrogramas. Contudo, não há nenhuma informação sobre a similaridade das linhagens dentro de cada grupo, nem sobre a similaridade entre os grupos. Isso pode ser considerado uma desvantagem desse método.

4.6 Projeção das similaridades no plano bidimensional

Nas figuras 5 e 6, são apresentados os gráficos obtidos com a projeção bidimensional das matrizes de similaridades, para os oito coeficientes e dois marcadores

Observa-se que os diferentes coeficientes de similaridades pouco alteram a projeção das distâncias no plano bidimensional. Para o marcador AFLP, pode-se observar para os coeficientes de Jaccard, Sorensen-Dice, Ochiai e Ochiai II que as linhagens 8 e 9, posicionaram-se distantes em relação as demais. Para tais coeficientes, observa-se a existência de dois grupos, o primeiro composto pelas linhagens 11, 12, 14, 17 e 18 e o segundo pelas linhagens 1, 2, 3, 4, 5, 6, 7,10, 13, 15 e 16. Contudo, essa distinção não é clara, principalmente para as linhagens 10 e 15.

Para os coeficientes Simple Matching e Rogers e Tanimoto, as linhagens 4 e 9 mostram-se distante das demais linhagens e também para estes, dois grupos pode-se ser observados. O primeiro grupo composto pelas linhagens 11, 12, 14, 17 e 18 e o segundo pelas linhagens 1, 2, 3, 5, 6, 7, 8, 10, 13, 15 e 16. Assim, nota-se que a única diferença com os coeficientes mencionados anteriormente deve-se à separação da linhagem 4, e não da 8. Também nesse caso a distinção dos grupos não foi tão clara. Para o coeficiente de Russel e Rao, com exceção das linhagens 8, 9 e 12, as demais apresentam-se bastante próximas, não sendo distinguidos grupos.

O Coeficiente de Ochiai II, da mesma forma que no método de Tocher para o marcador RAPD, mostrou-se mais semelhante aos coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai, o que mostra que a separação desses coeficientes em três grupos não ocorre necessariamente em todas as situações.

Para o marcador RAPD, nota-se que para todos coeficientes não foi possível identificar com clareza a formação dos grupos. As linhagens das duas populações mostram-se bastante parecidas, havendo separação das linhagens 1, 5, 10, 12 e 13 das demais para os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai. No Coeficiente de Anderberg, porém, as linhagens 10 e 12 se aproximaram das demais. Para os coeficientes Simple Matching, Rogers e Tanimoto e Ochiai II, as linhagens que se afastaram das demais foram as 15, 16 e 13. As linhagens 1, 10 e 12 afastaram-se um

pouco das outras. O coeficiente de Russel e Rao separou as linhagens 1, 6 e 13, novamente diferindo dessas duas caregorias.

A eficiência da projeção bidimensional foi avaliada por três parâmetros, sendo esses o grau de distorção, a correlação entre a distância original e a estimada e o nível de estresse. Tais resultados são apresentados nas Tabelas 26 e 27.

Para o marcador AFLP, a distorção variou de 44,52% para o coeficiente de Ochiai a 67,28% para o coeficiente de Russel e Rao. O nível do estresse variou de 49,56% a 71,56%, para os mesmos coeficientes e a correlação variou de 0,6216 para o coeficiente de Russel e Rao a 0,6933 para o coeficiente de Ochiai.

Para o marcador RAPD, a distorção variou de 49,02% para o coeficiente de Simple Matching a 71,70% para o coeficiente de Russel e Rao, o nível do estresse variou de 52,12% para o coeficiente de Sorensen-Dice a 75,57% para o coeficiente de Russel e Rao. A correlação variou de 0,5766 para o coeficiente de Russel e Rao. 0,8253 para o coeficiente de Sorensen-Dice. As correlações para esse marcador foram levemente mais altas que as observadas para o marcador AFLP.

De acordo com a classificação de Kruskal (1964) (Tabela 8), os valores do estresse foi insatisfatório para todos os coeficientes para os dois marcadores, sugerindo assim que esse método de projeção bidimensional não é adequado para esse conjunto de dados, ou seja, as projeções não foram eficientes para representar as matrizes de similaridade. Desse modo, a comparação dos coeficientes nessa situação deve ser feita com ressalvas. Também para os dois marcadores, o grau de distorção foi elevado e as correlações foram baixas em todas as situações, corroborando essa última afirmativa. Entretanto, o coeficiente de Russel e Rao mostrou resultados notadamente piores que os demais.

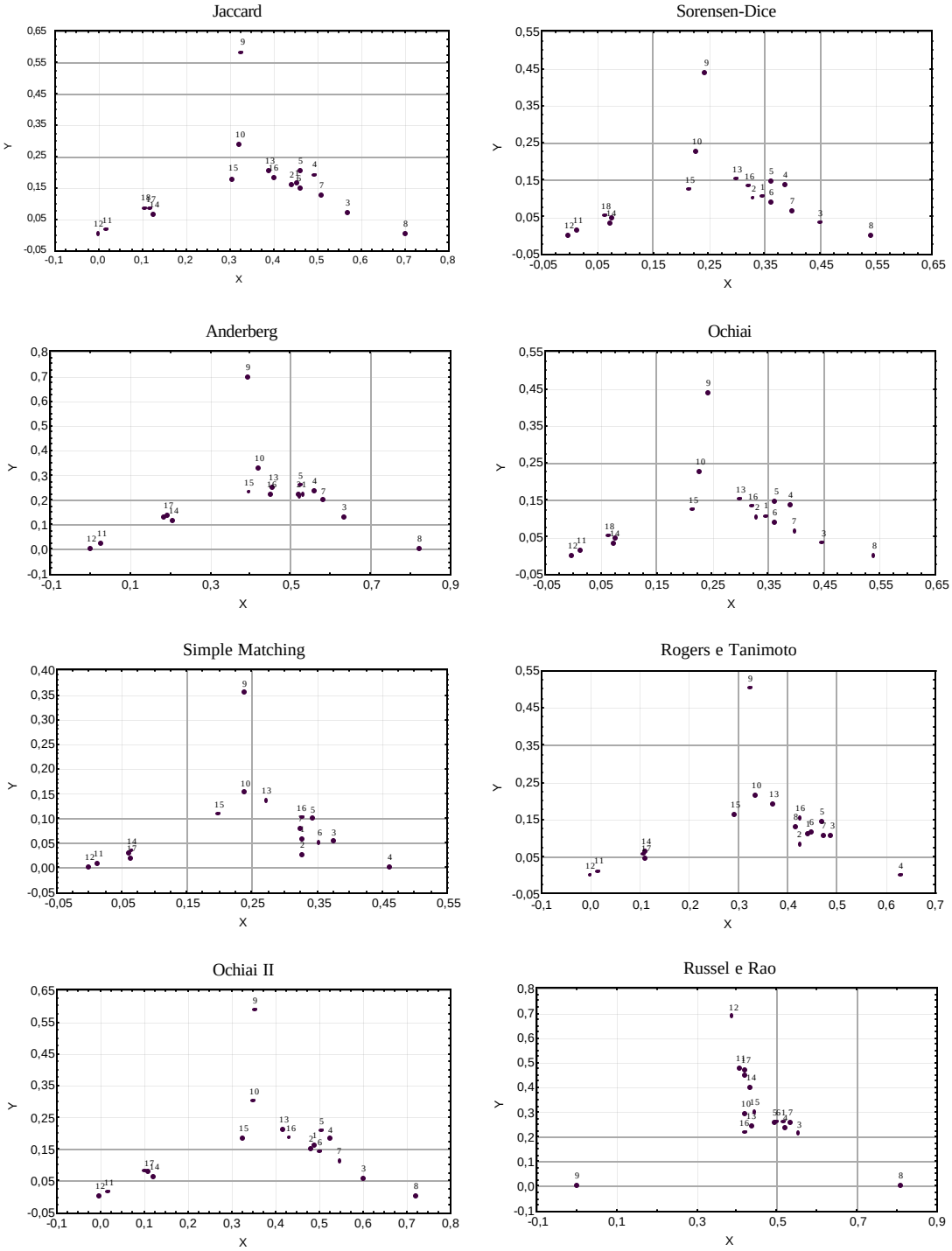

Figura 5 - Projeção do complemento dos coeficientes de similaridades entre 18 linhagens de milho das populações BR 105 e BR 106 no plano bidimensional. Marcador molecular AFLP.

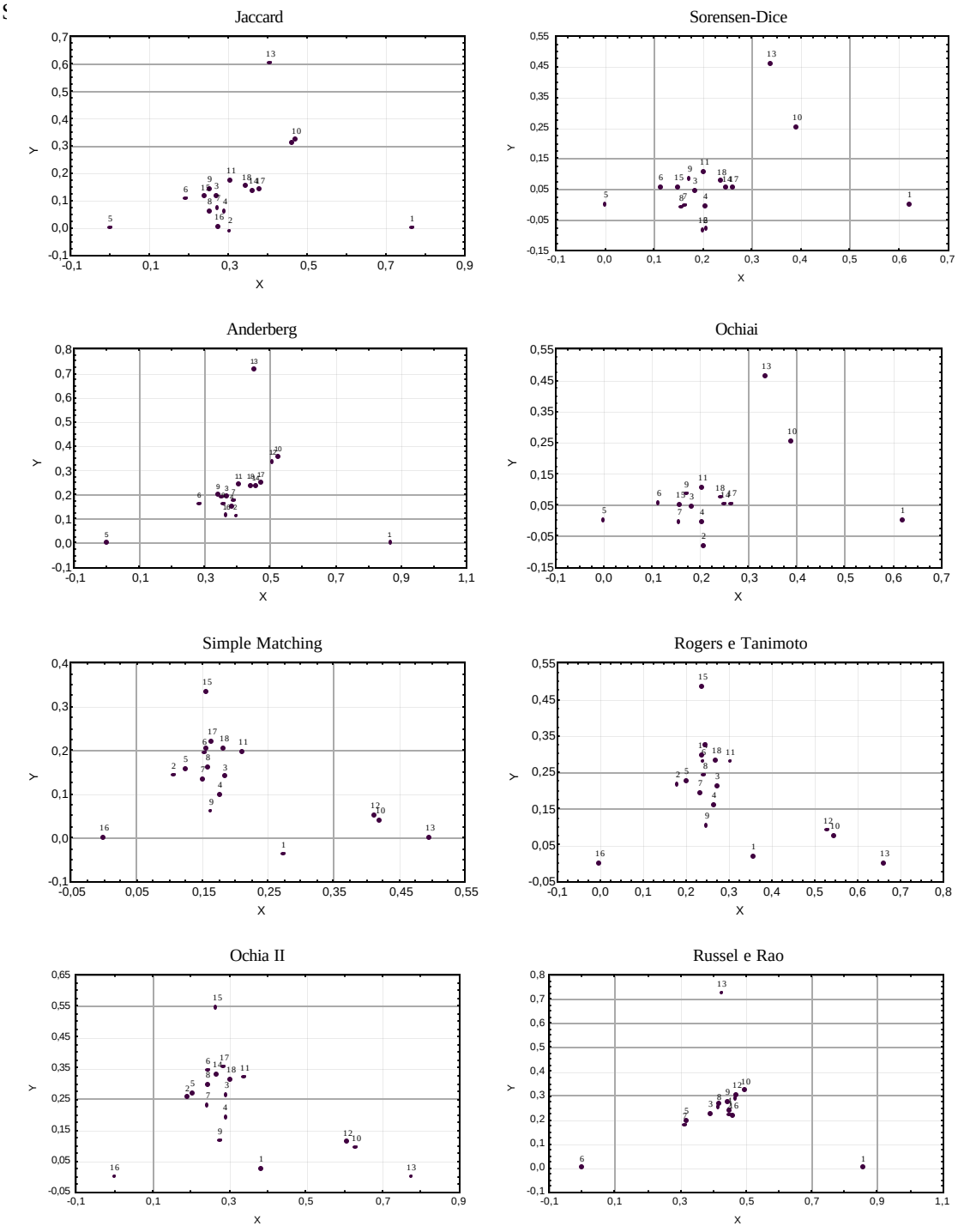

Figura 6 - Projeção do complemento dos coeficientes de similaridades entre 18 linhagens de milho das populações BR 105 e BR 106 no plano bidimensional. Marcador molecular RAPD.

Tabela 26. Grau de distorção, correlação entre as distâncias originais e estimadas (r) e valor do estresse, obtido pela projeção das distâncias no espaço bidimensional. Marcador molecular AFLP (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficientes	Distorção (%)	r	Estresse (%)
J	50,60	0,6039	55,49
SD	44,62	0,6919	49,66
A	57,21	0,4756	61,97
O	44,52	0,6933	49,56
SM	46,87	0,6694	52,12
RT	52,10	0,5853	57,22
OII	49,40	0,6216	54,37
RR	67,28	0,5047	71,35

Tabela 27. Grau de distorção, correlação entre as distâncias originais e estimadas (r) e valor do estresse, obtido pela projeção das distâncias no espaço bidimensional. Marcador molecular RAPD (J: coeficiente de Jaccard; SD: coeficiente de Sorensen-Dice; A: coeficiente de Anderberg; O: coeficiente de Ochiai; SM: coeficiente Simple Matching; RT: coeficiente de Rogers e Tanimoto; OII: coeficiente de Ochiai II; RR: coeficiente de Russel e Rao).

Coeficientes	Distorção (%)	r	Estresse (%)
J	59,21	0,7149	61,90
SD	49,59	0,8253	52,12
A	67,37	0,5463	70,70
O	49,57	0,8241	52,15
SM	49,02	0,7430	52,93
RT	55,46	0,6446	59,17
OII	55,03	0,6622	58,66
RR	71,70	0,5766	75,57

4.7 Considerações finais

Apesar do presente trabalho não ter como objetivo comparar os métodos de análise e os marcadores moleculares, algumas considerações podem ser feitas. Depreende-se, com base em tudo que foi apresentado, que, de forma geral, não se deve usar coeficientes de correlação para comparar coeficientes de similaridade, uma vez que, as correlações normalmente são de elevada magnitude, não havendo necessariamente a formação dos mesmos grupos em análises subseqüentes.

Dentre todas as análises de agrupamento empregadas (dendrogramas, método de otimização de Tocher e projeção da matriz de similaridade no plano bidimensional), a que apresentou resultados mais próximos da natureza dos dados foi a que fornece dendrogramas, pois ela refletiu a constituição dos grupos esperados *a priori*, sendo que a inclusão da linhagem 16 no grupo contrastante (BR-105), já tinha sido reiterada anteriormente (Barbosa, 2001). Porém, essa coerência nos agrupamentos foi maior quando considerados os dados provenientes do marcador AFLP. Além disso, com base nas correlações cofenéticas, distorções e estresses, nota-se que os dendrogramas permitiram uma representação das matrizes de similaridade superior aos demais métodos. Vale salientar que a comparação dos dendrogramas entre si é mais evidente quando se emprega o índice de consenso.

A principal dificuldade para interpretar os resultados da análise de agrupamentos com construção de dendrogramas deve-se ao fato de não haver um critério objetivo para identificar os grupos formados. Em diversos trabalhos com linhagens de milho adaptados ao clima tropical, a formação desses grupos é mais evidente, em função da ocorrência de menor variabilidade dentro deles. Contudo, para germoplasma tropical, como é o presente caso, há muita variabilidade também dentro dos grupos, dificultando sua identificação.

O método de otimização de Tocher, embora de fácil implementação, não fornece informação sobre a similaridade das linhagens dentro dos grupos e tampouco entre os grupos. Seus resultados são razoavelmente semelhantes aos obtidos com os

dendrogramas e embora seja permitido alterar o critério para considerar a inclusão de linhagens nos grupos formados, nota-se que isso não é feito de forma trivial.

A projeção das matrizes de similaridade no plano bidimensional não mostrou bons resultados, o que impede seu emprego para comparar os coeficientes de forma efetiva. Algumas alternativas, como a análise de componentes principais e análise de coordenadas principais, devem ser consideradas em trabalhos futuros.

Mesmo com marcadores de natureza diferente, e com análises também com princípios teóricos distintos, algumas tendências gerais foram observadas. Os oito coeficientes comparados podem ser separados em três categorias: a primeira delas, composta pelos coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai; a segunda, pelos coeficientes Simple Matching, Rogers e Tanimoto e Ochiai II. A terceira categoria é composta somente pelo coeficiente de Russel e Rao, que sempre mostrou resultado diferente dos demais, pouco concordando com as evidências biológicas a respeito do agrupamento das linhagens. De forma geral, a escolha de um dos coeficientes dentro de uma dessas categorias não causa alterações nas análises e na sua interpretação, ou seja, obtém-se quase sempre os mesmos grupos.

É interessante notar que os coeficientes dentro de cada categoria possuem princípios comuns. Os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai não consideram as ausências em comum das bandas, sendo que os coeficientes Simple Matching, Rogers e Tanimoto e Ochiai II incluem-nas nas suas expressões. O coeficiente de Russel eRao considera as ausências conjuntas comuns no denominador de sua expressão e não no numerador, sendo portanto um coeficiente misto. Isso possivelmente explique ao surgimento dessas três categorias.

De acordo com os resultados obtidos, o coeficiente de Russel e Rao deve ser usado apenas em situações específicas, em que seja realmente de interesse do pesquisador utilizá-lo, tendo uma justificativa para tal procedimento. Além dos grupos formados serem diferentes, a diagonal da matriz de similaridades não apresenta necessariamente valores iguais a 1, o que não deveria acontecer, pois a similaridade de uma linhagem com ela mesma deveria sempre ser igual a 1. Sokal & Sneath (1963) e Duarte et al. (1999) discutem que a utilização desse coeficiente é questionável.

Em função dos coeficientes que têm propriedades semelhantes apresentarem resultados muito parecidos, a decisão do pesquisador deve ser baseada no fato de considerar ou não as ausências conjuntas de bandas (0, 0) como medida de coincidência.

No caso dos marcadores moleculares dominantes do tipo RAPD aplicados a cultivares de feijão, Duarte et al. (1999) encontraram maior eficiência na projeção no espaço bidimensional para o coeficiente de Sorensen-Dice, motivo pelo qual eles sugerem sua utilização. Na presente situação, isso não ocorreu, impedindo que uma recomendação geral seja feita. Porém, em função das propriedades bioquímicas dos marcadores dominantes, não há garantias que as regiões do DNA em que são reveladas as bandas entre duas linhagens (0, 0) sejam de fato idênticas. Assim, parece razoável considerar que os coeficientes que desconsideram a ausência conjunta tenham maior justificativa para serem usados. Desse modo, poder-se-ia escolher qualquer coeficiente dentre Jaccard, Sorensen-Dice, Anderberg e Ochiai para obter bons resultados. Isso mostra que é coerente o fato do coeficiente de Jaccard ser o mais comumente empregado nos artigos da literatura, possivelmente devido a sua fácil interpretação (é uma taxa entre o número de coincidências e o número total de bandas, sem considerar as ausências conjuntas).

5 CONCLUSÕES

a) Praticamente em todas as situações estudadas, os coeficientes de Jaccard, Sorensen-Dice, Anderberg e Ochiai mostraram os mesmos resultados entre si. Isso foi atribuído ao fato deles apresentarem como propriedade comum a desconsideração das ausências conjuntas de bandas.

b) Isso também ocorreu para os coeficientes Simple Matching, de Rogers e Tanimoto e de Ochiai II, que também não apresentaram alterações sensíveis nos resultados obtidos entre si. Possivelmente, isso foi devido ao fato de todos eles considerarem as ausências conjuntas nas suas expressões.

c) O coeficiente de Russel e Rao apresentou resultados muito diferentes dos demais coeficientes, possivelmente por excluir as ausências conjuntas de bandas no numerador de sua expressão, incluindo-as no denominador. Seu uso não é recomendado.

d) Em função das ausências conjuntas de bandas não significarem necessariamente que as regiões do DNA são idênticas, sugere-se o emprego de algum coeficiente dentre os que desconsideram a ausência conjunta, ou seja, entre Jaccard, Sorensen-Dice, Anderberg ou Ochiai.

ANEXOS

Tabela 10. Valores dos coeficientes de similaridade entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir de dados do marcador AFLP (acima da diagonal: coeficiente de Jaccard; abaixo da diagonal: coeficiente de Sorensen-Dice).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,5164	0,4542	0,4794	0,5150	0,4614	0,4872	0,4416	0,3737	0,4375	0,4043	0,3942	0,3812	0,3964	0,4102	0,4637	0,4322	0,4244
2	0,6811	-	0,4752	0,5267	0,5397	0,4910	0,4831	0,4433	0,3901	0,4540	0,4318	0,4219	0,3917	0,4192	0,3925	0,4818	0,4637	0,4511
3	0,6246	0,6443	-	0,4510	0,4283	0,4273	0,4698	0,5965	0,3292	0,3747	0,3353	0,3307	0,3713	0,3482	0,3354	0,4289	0,3640	0,3392
4	0,6481	0,6900	0,6217	-	0,4645	0,4433	0,4240	0,4387	0,3642	0,4407	0,3606	0,3438	0,3960	0,3578	0,3777	0,4333	0,3849	0,3846
5	0,6798	0,7011	0,5997	0,6343	-	0,4512	0,4243	0,3846	0,3663	0,4382	0,3517	0,3418	0,3714	0,3704	0,3693	0,4501	0,3855	0,3661
6	0,6314	0,6586	0,5988	0,6143	0,6218	-	0,7117	0,4200	0,3333	0,3764	0,3662	0,3506	0,3612	0,3689	0,3591	0,4494	0,3842	0,3783
7	0,6552	0,6515	0,6393	0,5955	0,5958	0,8316	-	0,4434	0,3424	0,3877	0,3642	0,3540	0,3591	0,3668	0,3627	0,4439	0,3713	0,3653
8	0,6127	0,6143	0,7473	0,6098	0,5556	0,5915	0,6144	-	0,3087	0,3304	0,3108	0,2984	0,3807	0,3231	0,3404	0,3775	0,3294	0,3175
9	0,5441	0,5612	0,4953	0,5339	0,5362	0,5000	0,5102	0,4717	-	0,4518	0,3449	0,3320	0,3799	0,3472	0,3658	0,3441	0,3526	0,3516
10	0,6087	0,6245	0,5452	0,6118	0,6094	0,5469	0,5587	0,4967	0,6224	-	0,4217	0,4165	0,4000	0,4200	0,4295	0,4178	0,4651	0,4580
11	0,5758	0,6031	0,5022	0,5301	0,5203	0,5361	0,5339	0,4742	0,5129	0,5933	-	0,9148	0,3694	0,6594	0,4661	0,3253	0,7218	0,7213
12	0,5655	0,5935	0,4970	0,5116	0,5095	0,5192	0,5229	0,4597	0,4985	0,5881	0,9555	-	0,3563	0,6570	0,4325	0,3234	0,6910	0,6859
13	0,5520	0,5629	0,5415	0,5674	0,5416	0,5307	0,5284	0,5515	0,5506	0,5714	0,5395	0,5254	-	0,3693	0,4163	0,3467	0,3988	0,3930
14	0,5678	0,5907	0,5165	0,5270	0,5406	0,5389	0,5367	0,4884	0,5154	0,5915	0,7948	0,7930	0,5394	-	0,4389	0,3327	0,6404	0,6383
15	0,5818	0,5637	0,5023	0,5483	0,5394	0,5284	0,5323	0,5079	0,5357	0,6010	0,6358	0,6039	0,5879	0,6100	-	0,3532	0,4435	0,4569
16	0,6336	0,6503	0,6003	0,6046	0,6208	0,6202	0,6149	0,5481	0,5120	0,5894	0,4910	0,4887	0,5149	0,4992	0,5220	-	0,3598	0,3454
17	0,6035	0,6336	0,5337	0,5559	0,5565	0,5551	0,5415	0,4956	0,5214	0,6349	0,8384	0,8173	0,5702	0,7808	0,6145	0,5292	-	0,8564
18	0,5959	0,6218	0,5066	0,5556	0,5360	0,5489	0,5351	0,4819	0,5203	0,6282	0,8381	0,8137	0,5643	0,7792	0,6272	0,5134	0,9227	-

Tabela 11. Valores dos coeficientes de similaridade entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir de dados do marcador AFLP (acima da diagonal: coeficiente de Anderberg; abaixo da diagonal: coeficiente de Ochiai).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,3481	0,2938	0,3152	0,3468	0,2999	0,3221	0,2834	0,2298	0,2800	0,2534	0,2455	0,2355	0,2472	0,2580	0,3018	0,2756	0,2693
2	0,6823	-	0,3117	0,3575	0,3696	0,3254	0,3185	0,2848	0,2423	0,2937	0,2753	0,2674	0,2436	0,2652	0,2441	0,3173	0,3018	0,2913
3	0,6258	0,6489	-	0,2912	0,2725	0,2717	0,3070	0,4250	0,1970	0,2306	0,2014	0,1981	0,2280	0,2108	0,2015	0,2730	0,2225	0,2043
4	0,6482	0,6906	0,6235	-	0,3025	0,2848	0,2690	0,2810	0,2226	0,2826	0,2200	0,2075	0,2469	0,2179	0,2328	0,2765	0,2383	0,2381
5	0,6805	0,7048	0,5998	0,6355	-	0,2913	0,2693	0,2381	0,2242	0,2806	0,2133	0,2061	0,2280	0,2273	0,2265	0,2904	0,2388	0,2241
6	0,6325	0,6631	0,5988	0,6160	0,6219	-	0,5524	0,2658	0,2000	0,2318	0,2241	0,2126	0,2204	0,2261	0,2188	0,2899	0,2377	0,2333
7	0,6564	0,6561	0,6393	0,5972	0,5959	0,8316	-	0,2849	0,2066	0,2404	0,2226	0,2151	0,2188	0,2246	0,2215	0,2853	0,2280	0,2234
8	0,6152	0,6212	0,7476	0,6133	0,5562	0,5919	0,6147	-	0,1825	0,1979	0,1840	0,1754	0,2351	0,1927	0,2051	0,2327	0,1972	0,1887
9	0,5462	0,5674	0,4955	0,5369	0,5367	0,5002	0,5104	0,4717	-	0,2918	0,2084	0,1990	0,2345	0,2101	0,2238	0,2078	0,2140	0,2133
10	0,6091	0,6274	0,5452	0,6127	0,6094	0,5470	0,5588	0,4973	0,6229	-	0,2672	0,2630	0,2500	0,2658	0,2735	0,2641	0,3030	0,2970
11	0,5760	0,6055	0,5025	0,5306	0,5204	0,5363	0,5342	0,4750	0,5138	0,5934	-	0,8430	0,2265	0,4919	0,3039	0,1943	0,5647	0,5641
12	0,5657	0,5959	0,4973	0,5122	0,5095	0,5194	0,5231	0,4605	0,4993	0,5881	0,9555	-	0,2167	0,4892	0,2760	0,1929	0,5279	0,5219
13	0,5534	0,5676	0,5416	0,5695	0,5418	0,5307	0,5284	0,5516	0,5507	0,5717	0,5399	0,5258	-	0,2265	0,2629	0,2097	0,2491	0,2446
14	0,5684	0,5940	0,5166	0,5281	0,5406	0,5390	0,5368	0,4889	0,5159	0,5916	0,7949	0,7931	0,5396	-	0,2811	0,1995	0,4710	0,4688
15	0,5832	0,5685	0,5023	0,5504	0,5396	0,5285	0,5323	0,5080	0,5357	0,6011	0,6364	0,6044	0,5879	0,6102	-	0,2145	0,2849	0,2961
16	0,6356	0,6567	0,6004	0,6075	0,6212	0,6203	0,6151	0,5481	0,5120	0,5897	0,4916	0,4893	0,5149	0,4995	0,5220	-	0,2194	0,2087
17	0,6035	0,6348	0,5346	0,5559	0,5570	0,5559	0,5424	0,4975	0,5233	0,6355	0,8388	0,8176	0,5715	0,7815	0,6160	0,5309	-	0,7489
18	0,5960	0,6238	0,5070	0,5560	0,5362	0,5493	0,5355	0,4831	0,5215	0,6284	0,8381	0,8137	0,5649	0,7795	0,6280	0,5144	0,9228	-

Tabela 12. Valores dos coeficientes de similaridade entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir de dados do marcador AFLP (acima da diagonal: coeficiente Simple Matching; abaixo da diagonal: coeficiente de Rogers e Tanimoto).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,6757	0,6615	0,6576	0,7067	0,6667	0,6886	0,6602	0,5995	0,6360	0,6059	0,5969	0,5995	0,6047	0,6266	0,6757	0,6214	0,6214
2	0,5102	-	0,6576	0,6796	0,7080	0,6705	0,6641	0,6382	0,5879	0,6276	0,6072	0,5982	0,5827	0,6008	0,5840	0,6693	0,6279	0,6227
3	0,4942	0,4899	-	0,6525	0,6550	0,6589	0,6938	0,7920	0,5840	0,5997	0,5646	0,5607	0,6150	0,5840	0,5827	0,6680	0,5801	0,5646
4	0,4899	0,5147	0,4842	-	0,6589	0,6447	0,6279	0,6512	0,5827	0,6318	0,5556	0,5388	0,6059	0,5594	0,5891	0,6434	0,5685	0,5762
5	0,5465	0,5480	0,4870	0,4913	-	0,6731	0,6512	0,6279	0,6111	0,6513	0,5736	0,5646	0,6085	0,5982	0,6072	0,6796	0,5943	0,5840
6	0,5000	0,5044	0,4913	0,4757	0,5073	-	0,8566	0,6628	0,5866	0,6025	0,5930	0,5788	0,6047	0,6021	0,6034	0,6835	0,5982	0,6008
7	0,5251	0,4971	0,5312	0,4576	0,4828	0,7492	-	0,6822	0,5956	0,6123	0,5917	0,5827	0,6034	0,6008	0,6072	0,6796	0,5866	0,5891
8	0,4928	0,4687	0,6556	0,4828	0,4576	0,4957	0,5176	-	0,5775	0,5704	0,5530	0,5413	0,6344	0,5724	0,5995	0,6357	0,5581	0,5556
9	0,4280	0,4163	0,4124	0,4111	0,4400	0,4150	0,4241	0,4060	-	0,6750	0,5853	0,5736	0,6331	0,5943	0,6214	0,6059	0,5801	0,5879
10	0,4663	0,4573	0,4283	0,4618	0,4829	0,4311	0,4412	0,3990	0,5095	-	0,6290	0,6248	0,6276	0,6360	0,6499	0,6444	0,6583	0,6583
11	0,4347	0,4360	0,3933	0,3846	0,4022	0,4215	0,4202	0,3821	0,4137	0,4588	-	0,9599	0,6008	0,8178	0,6848	0,5633	0,8501	0,8527
12	0,4254	0,4267	0,3896	0,3687	0,3933	0,4073	0,4111	0,3711	0,4022	0,4544	0,9230	-	0,5891	0,8165	0,6576	0,5620	0,8307	0,8307
13	0,4280	0,4111	0,4440	0,4347	0,4373	0,4333	0,4320	0,4645	0,4631	0,4573	0,4294	0,4176	-	0,6072	0,6576	0,6008	0,6163	0,6189
14	0,4333	0,4294	0,4124	0,3883	0,4267	0,4307	0,4294	0,4009	0,4228	0,4663	0,6918	0,6900	0,4360	-	0,6680	0,5775	0,7997	0,8023
15	0,4563	0,4124	0,4111	0,4176	0,4360	0,4320	0,4360	0,4280	0,4508	0,4814	0,5206	0,4899	0,4899	0,5015	-	0,6072	0,6563	0,6744
16	0,5102	0,5029	0,5015	0,4743	0,5147	0,5191	0,5147	0,4659	0,4347	0,4753	0,3921	0,3908	0,4294	0,4060	0,4360	-	0,5840	0,5788
17	0,4508	0,4576	0,4086	0,3971	0,4228	0,4267	0,4150	0,3871	0,4086	0,4906	0,7393	0,7105	0,4454	0,6663	0,4885	0,4124	-	0,9276
18	0,4508	0,4522	0,3933	0,4047	0,4124	0,4294	0,4176	0,3846	0,4163	0,4906	0,7432	0,7105	0,4481	0,6699	0,5088	0,4073	0,8651	-

Tabela 13. Valores dos coeficientes de similaridades entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir dos dados do marcador AFLP (acima da diagonal: coeficiente de Ochiai II; abaixo da diagonal: coeficiente de Russel e Rao).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,4581	0,4334	0,4322	0,4967	0,4405	0,4706	0,4300	0,3520	0,4021	0,3642	0,3532	0,3535	0,3616	0,3872	0,4517	0,3849	0,3839
2	0,3463	-	0,4375	0,4621	0,5062	0,4547	0,4461	0,4132	0,3499	0,3975	0,3715	0,3606	0,3434	0,3643	0,3449	0,4542	0,3956	0,3903
3	0,2817	0,3101	-	0,4241	0,4180	0,4212	0,4692	0,6156	0,3202	0,3504	0,3081	0,3035	0,3619	0,3280	0,3219	0,4300	0,3308	0,3096
4	0,3152	0,3566	0,2855	-	0,4330	0,4140	0,3923	0,4215	0,3353	0,3987	0,3072	0,2887	0,3644	0,3108	0,3440	0,4115	0,3227	0,3310
5	0,3114	0,3424	0,2584	0,2959	-	0,4429	0,4131	0,3784	0,3572	0,4175	0,3207	0,3101	0,3568	0,3476	0,3549	0,4490	0,3490	0,3341
6	0,2855	0,3178	0,2545	0,2829	0,2687	-	0,7277	0,4220	0,3240	0,3534	0,3420	0,3248	0,3495	0,3503	0,3476	0,4521	0,3527	0,3528
7	0,2959	0,3140	0,2713	0,2739	0,2571	0,3540	-	0,4487	0,3347	0,3656	0,3402	0,3292	0,3476	0,3485	0,3521	0,4464	0,3386	0,3385
8	0,2687	0,2881	0,3075	0,2726	0,2326	0,2442	0,2532	-	0,3057	0,3112	0,2907	0,2773	0,3814	0,3095	0,3365	0,3808	0,3027	0,2955
9	0,2390	0,2636	0,2041	0,2390	0,2248	0,2067	0,2106	0,1886	-	0,4454	0,3286	0,3145	0,3800	0,3361	0,3646	0,3428	0,3284	0,3335
10	0,2831	0,3096	0,2399	0,2901	0,2720	0,2399	0,2455	0,2120	0,2678	-	0,3911	0,3856	0,3836	0,3973	0,4137	0,4049	0,4317	0,4299
11	0,2674	0,2984	0,2196	0,2506	0,2313	0,2351	0,2339	0,2016	0,2183	0,2706	-	0,9207	0,3499	0,6647	0,4598	0,3039	0,7218	0,7249
12	0,2623	0,2933	0,2171	0,2416	0,2261	0,2274	0,2287	0,1951	0,2119	0,2678	0,4302	-	0,3355	0,6625	0,4224	0,3021	0,6890	0,6876
13	0,2468	0,2687	0,2274	0,2584	0,2313	0,2235	0,2222	0,2248	0,2248	0,2483	0,2339	0,2274	-	0,3549	0,4157	0,3403	0,3740	0,3739
14	0,2597	0,2881	0,2222	0,2455	0,2364	0,2326	0,2313	0,2041	0,2158	0,2636	0,3527	0,3514	0,2300	-	0,4339	0,3171	0,6379	0,6401
15	0,2597	0,2687	0,2106	0,2494	0,2300	0,2222	0,2235	0,2067	0,2183	0,2636	0,2752	0,2610	0,2442	0,2597	-	0,3480	0,4257	0,4469
16	0,2804	0,3075	0,2494	0,2726	0,2623	0,2584	0,2558	0,2209	0,2067	0,2552	0,2106	0,2093	0,2119	0,2106	0,2145	-	0,3337	0,3237
17	0,2881	0,3217	0,2403	0,2700	0,2545	0,2506	0,2442	0,2171	0,2287	0,2971	0,3889	0,3786	0,2545	0,3566	0,2739	0,2339	-	0,8602
18	0,2791	0,3101	0,2235	0,2649	0,2403	0,2429	0,2364	0,2067	0,2235	0,2887	0,3811	0,3695	0,2468	0,3488	0,2739	0,2222	0,4315	-

Tabela 14. Valores dos coeficientes de similaridades entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir dos dados do marcador RAPD (acima da diagonal: coeficiente de Jaccard; abaixo da diagonal: coeficiente de Sorensen-Dice).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,3019	0,2763	0,3137	0,2327	0,2357	0,2532	0,2649	0,2671	0,3778	0,2563	0,3657	0,2971	0,2683	0,2426	0,3226	0,2754	0,2651
2	0,4638	-	0,5033	0,5600	0,5369	0,5033	0,4868	0,4834	0,4630	0,3114	0,4534	0,3012	0,2398	0,4606	0,4464	0,5355	0,5000	0,4551
3	0,4330	0,6696	-	0,4832	0,4497	0,4459	0,4392	0,6107	0,4088	0,3121	0,3899	0,2930	0,2756	0,4161	0,4375	0,4331	0,4563	0,4198
4	0,4776	0,7179	0,6516	-	0,4966	0,4444	0,4194	0,5034	0,4085	0,3396	0,3902	0,3125	0,2716	0,3988	0,4192	0,4500	0,4545	0,4107
5	0,3776	0,6987	0,6204	0,6637	-	0,4595	0,4726	0,5214	0,4304	0,3000	0,4481	0,2813	0,2722	0,4557	0,4872	0,5033	0,4688	0,4410
6	0,3814	0,6696	0,6168	0,6154	0,6296	-	0,7317	0,4653	0,3576	0,2875	0,4539	0,2688	0,2675	0,4805	0,4839	0,4423	0,4472	0,4024
7	0,4041	0,6549	0,6103	0,5909	0,6419	0,8451	-	0,4685	0,3515	0,2813	0,4013	0,2547	0,2532	0,4459	0,4586	0,4452	0,4233	0,4136
8	0,4188	0,6518	0,7583	0,6697	0,6854	0,6351	0,6381	-	0,3899	0,2767	0,3797	0,2658	0,2895	0,4151	0,4551	0,4605	0,4375	0,4188
9	0,4216	0,6329	0,5804	0,5801	0,6018	0,5268	0,5202	0,5611	-	0,3416	0,4528	0,2909	0,2981	0,4601	0,3953	0,4596	0,4908	0,4545
10	0,5484	0,4749	0,4757	0,5070	0,4615	0,4466	0,4390	0,4335	0,5093	-	0,3148	0,4444	0,4580	0,3253	0,3214	0,2840	0,3235	0,3293
11	0,4080	0,6239	0,5611	0,5614	0,6188	0,6244	0,5727	0,5505	0,6234	0,4789	-	0,4483	0,2956	0,6549	0,5096	0,4233	0,6216	0,6122
12	0,5355	0,4630	0,4532	0,4762	0,4390	0,4236	0,4059	0,4200	0,4507	0,6154	0,6190	-	0,4462	0,3822	0,3354	0,2738	0,3789	0,4038
13	0,4581	0,3868	0,4322	0,4272	0,4279	0,4221	0,4040	0,4490	0,4593	0,6283	0,4563	0,6170	-	0,2754	0,2647	0,2353	0,3054	0,2798
14	0,4231	0,6307	0,5877	0,5702	0,6261	0,6491	0,6167	0,5867	0,6303	0,4909	0,7915	0,5530	0,4319	-	0,5443	0,4753	0,7153	0,6944
15	0,3905	0,6173	0,6087	0,5907	0,6552	0,6522	0,6288	0,6256	0,5667	0,4865	0,6751	0,5023	0,4186	0,7049	-	0,4260	0,5276	0,5472
16	0,4878	0,6975	0,6044	0,6207	0,6696	0,6133	0,6161	0,6306	0,6298	0,4424	0,5948	0,4299	0,3810	0,6444	0,5975	-	0,4788	0,4431
17	0,4319	0,6667	0,6266	0,6250	0,6383	0,6180	0,5948	0,6087	0,6584	0,4889	0,7667	0,5495	0,4679	0,8340	0,6908	0,6475	-	0,8864
18	0,4190	0,6255	0,5913	0,5823	0,6121	0,5739	0,5852	0,5903	0,6250	0,4955	0,7595	0,5753	0,4372	0,8197	0,7073	0,6141	0,9398	-

Tabela 15. Valores dos coeficientes de similaridades entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir dos dados do marcador RAPD (acima da diagonal: coeficiente de Anderberg; abaixo da diagonal: coeficiente de Ochiai).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,1778	0,1603	0,1860	0,1317	0,1336	0,1450	0,1527	0,1541	0,2329	0,1470	0,2237	0,1745	0,1549	0,1380	0,1923	0,1597	0,1528
2	0,4698	-	0,3363	0,3889	0,3670	0,3363	0,3217	0,3188	0,3012	0,1844	0,2932	0,1773	0,1362	0,2992	0,2874	0,3656	0,3333	0,2946
3	0,4353	0,6707	-	0,3186	0,2900	0,2870	0,2814	0,4396	0,2569	0,1849	0,2422	0,1716	0,1599	0,2627	0,2800	0,2764	0,2955	0,2656
4	0,4820	0,7182	0,6519	-	0,3304	0,2857	0,2653	0,3364	0,2567	0,2045	0,2424	0,1852	0,1571	0,2491	0,2652	0,2903	0,2941	0,2584
5	0,3800	0,6995	0,6204	0,6638	-	0,2982	0,3094	0,3527	0,2742	0,1765	0,2887	0,1636	0,1575	0,2951	0,3220	0,3363	0,3061	0,2829
6	0,3835	0,6707	0,6168	0,6157	0,6297	-	0,5769	0,3032	0,2177	0,1679	0,2936	0,1552	0,1544	0,3162	0,3191	0,2840	0,2880	0,2519
7	0,4061	0,6561	0,6103	0,5913	0,6419	0,8451	-	0,3059	0,2132	0,1636	0,2510	0,1459	0,1449	0,2869	0,2975	0,2863	0,2685	0,2607
8	0,4205	0,6535	0,7584	0,6704	0,6856	0,6351	0,6381	-	0,2422	0,1606	0,2344	0,1533	0,1692	0,2619	0,2946	0,2991	0,2800	0,2648
9	0,4262	0,6330	0,5809	0,5801	0,6021	0,5273	0,5208	0,5621	-	0,2060	0,2927	0,1702	0,1752	0,2988	0,2464	0,2984	0,3252	0,2941
10	0,5495	0,4771	0,4761	0,5083	0,4621	0,4469	0,4393	0,4336	0,5110	-	0,1868	0,2857	0,2970	0,1942	0,1915	0,1655	0,1930	0,1971
11	0,4117	0,6241	0,5614	0,5614	0,6190	0,6247	0,5731	0,5510	0,6234	0,4801	-	0,2889	0,1734	0,4869	0,3419	0,2685	0,4510	0,4412
12	0,5362	0,4658	0,4539	0,4779	0,4399	0,4243	0,4064	0,4203	0,4529	0,6155	0,6213	-	0,2871	0,2362	0,2015	0,1586	0,2337	0,2530
13	0,4583	0,3902	0,4334	0,4296	0,4294	0,4233	0,4051	0,4498	0,4627	0,6287	0,4589	0,6172	-	0,1597	0,1525	0,1333	0,1802	0,1626
14	0,4288	0,6307	0,5888	0,5705	0,6269	0,6504	0,6181	0,5883	0,6303	0,4934	0,7918	0,5567	0,4360	-	0,3739	0,3117	0,5568	0,5319
15	0,3963	0,6173	0,6102	0,5911	0,6564	0,6538	0,6306	0,6278	0,5668	0,4894	0,6756	0,5061	0,4230	0,7049	-	0,2707	0,3583	0,3766
16	0,4935	0,6975	0,6052	0,6208	0,6701	0,6141	0,6170	0,6319	0,6298	0,4441	0,5949	0,4322	0,3839	0,6444	0,5976	-	0,3147	0,2846
17	0,4394	0,6669	0,6287	0,6258	0,6400	0,6201	0,5971	0,6115	0,6589	0,4924	0,7676	0,5546	0,4737	0,8342	0,6908	0,6479	-	0,7959
18	0,4253	0,6256	0,5927	0,5827	0,6132	0,5753	0,5868	0,5924	0,6252	0,4984	0,7600	0,5798	0,4418	0,8197	0,7073	0,6142	0,9398	-

Tabela 16. Valores dos coeficientes de similaridades entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir dos dados do marcador RAPD (acima da diagonal: coeficiente Simple Matching; abaixo da diagonal: coeficiente de Rogers e Tanimoto).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,5763	0,5802	0,5992	0,5344	0,5420	0,5611	0,5763	0,5496	0,6794	0,5458	0,6756	0,6298	0,5420	0,5115	0,5992	0,5382	0,5344
2	0,4048	-	0,7137	0,7481	0,7366	0,7137	0,7023	0,7023	0,6679	0,5611	0,6641	0,5573	0,5038	0,6603	0,6450	0,7252	0,6870	0,6527
3	0,4086	0,5549	-	0,7061	0,6870	0,6870	0,6832	0,8053	0,6412	0,5878	0,6298	0,5763	0,5687	0,6412	0,6565	0,6603	0,6679	0,6412
4	0,4278	0,5976	0,5457	-	0,7137	0,6756	0,6565	0,7252	0,6298	0,5992	0,6183	0,5802	0,5496	0,6145	0,6298	0,6641	0,6565	0,6221
5	0,3646	0,5831	0,5233	0,5549	-	0,6947	0,7061	0,7443	0,6565	0,5725	0,6756	0,5611	0,5611	0,6718	0,6947	0,7137	0,6756	0,6565
6	0,3717	0,5549	0,5233	0,5101	0,5322	-	0,8740	0,7061	0,5954	0,5649	0,6832	0,5534	0,5611	0,6947	0,6947	0,6679	0,6603	0,6260
7	0,3899	0,5412	0,5188	0,4886	0,5457	0,7763	-	0,7099	0,5916	0,5611	0,6412	0,5420	0,5496	0,6679	0,6756	0,6718	0,6412	0,6374
8	0,4048	0,5412	0,6741	0,5689	0,5927	0,5457	0,5503	-	0,6298	0,5611	0,6260	0,5573	0,5878	0,6450	0,6756	0,6870	0,6565	0,6450
9	0,3789	0,5014	0,4719	0,4596	0,4886	0,4239	0,4201	0,4596	-	0,5954	0,6679	0,5534	0,5687	0,6641	0,6031	0,6679	0,6832	0,6565
10	0,5145	0,3899	0,4162	0,4278	0,4011	0,3936	0,3899	0,3899	0,4239	-	0,5763	0,7137	0,7290	0,5725	0,5649	0,5382	0,5611	0,5725
11	0,3753	0,4971	0,4596	0,4475	0,5101	0,5188	0,4719	0,4556	0,5014	0,4048	-	0,6947	0,5725	0,8130	0,7061	0,6412	0,7863	0,7824
12	0,5101	0,3862	0,4048	0,4086	0,3899	0,3826	0,3717	0,3862	0,3826	0,5549	0,5322	-	0,7252	0,6298	0,5840	0,5344	0,6183	0,6450
13	0,4596	0,3367	0,3973	0,3789	0,3899	0,3899	0,3789	0,4162	0,3973	0,5736	0,4011	0,5689	-	0,5382	0,5229	0,5038	0,5573	0,5382
14	0,3717	0,4929	0,4719	0,4435	0,5057	0,5322	0,5014	0,4761	0,4971	0,4011	0,6849	0,4596	0,3681	-	0,7252	0,6756	0,8435	0,8321
15	0,3436	0,4761	0,4886	0,4596	0,5322	0,5322	0,5101	0,5101	0,4317	0,3936	0,5457	0,4124	0,3540	0,5689	-	0,6298	0,7061	0,7252
16	0,4278	0,5689	0,4929	0,4971	0,5549	0,5014	0,5057	0,5233	0,5014	0,3681	0,4719	0,3646	0,3367	0,5101	0,4596	-	0,6718	0,6450
17	0,3681	0,5233	0,5014	0,4886	0,5101	0,4929	0,4719	0,4886	0,5188	0,3899	0,6478	0,4475	0,3862	0,7294	0,5457	0,5057	-	0,9427
18	0,3646	0,4844	0,4719	0,4515	0,4886	0,4556	0,4678	0,4761	0,4886	0,4011	0,6426	0,4761	0,3681	0,7124	0,5689	0,4761	0,8917	-

Tabela 17. Valores dos coeficientes de similaridades entre 18 linhagens de milho das populações BR-105 e BR-106, calculados a partir dos dados do marcador RAPD (acima da diagonal: coeficiente de Ochiai II; abaixo da diagonal: coeficiente de Russel e Rao).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**
1	-	0,3069	0,2907	0,3264	0,2392	0,2445	0,2655	0,2807	0,2702	0,4132	0,2609	0,4027	0,3295	0,2675	0,2363	0,3326	0,2706	0,2618
2	0,1832	-	0,5018	0,5549	0,5363	0,5018	0,4849	0,4842	0,4411	0,2979	0,4348	0,2913	0,2285	0,4324	0,4130	0,5219	0,4703	0,4230
3	0,1603	0,2901	-	0,4864	0,4552	0,4537	0,4475	0,6348	0,3991	0,3145	0,3818	0,2971	0,2830	0,4023	0,4240	0,4253	0,4417	0,4038
4	0,1832	0,3206	0,2748	-	0,4985	0,4431	0,4164	0,5130	0,3881	0,3371	0,3717	0,3110	0,2709	0,3712	0,3915	0,4337	0,4279	0,3819
5	0,1412	0,3053	0,2557	0,2824	-	0,4661	0,4820	0,5380	0,4204	0,2984	0,4443	0,2816	0,2769	0,4439	0,4771	0,5011	0,4525	0,4247
6	0,1412	0,2901	0,2519	0,2595	0,2595	-	0,7554	0,4789	0,3412	0,2868	0,4537	0,2698	0,2738	0,4751	0,4766	0,4357	0,4314	0,3841
7	0,1489	0,2824	0,2481	0,2481	0,2634	0,3435	-	0,4837	0,3359	0,2810	0,3960	0,2551	0,2587	0,4376	0,4496	0,4404	0,4058	0,3985
8	0,1527	0,2786	0,3053	0,2786	0,2786	0,2557	0,2557	-	0,3825	0,2783	0,3748	0,2699	0,3019	0,4060	0,4490	0,4608	0,4255	0,4077
9	0,1641	0,2863	0,2481	0,2557	0,2595	0,2252	0,2214	0,2366	-	0,3357	0,4383	0,2832	0,2976	0,4364	0,3593	0,4402	0,4645	0,4272
10	0,1947	0,1985	0,1870	0,2061	0,1832	0,1756	0,1718	0,1679	0,2099	-	0,3091	0,4752	0,4948	0,3124	0,3056	0,2696	0,3043	0,3146
11	0,1565	0,2786	0,2366	0,2443	0,2634	0,2634	0,2405	0,2290	0,2748	0,1947	-	0,4638	0,2980	0,6578	0,4946	0,4034	0,6168	0,6094
12	0,1870	0,1908	0,1756	0,1908	0,1718	0,1641	0,1565	0,1603	0,1832	0,2290	0,2481	-	0,4849	0,3821	0,3265	0,2628	0,3728	0,4046
13	0,1565	0,1565	0,1641	0,1679	0,1641	0,1603	0,1527	0,1679	0,1832	0,2290	0,1794	0,2214	-	0,2675	0,2532	0,2257	0,2960	0,2702
14	0,1679	0,2901	0,2557	0,2557	0,2748	0,2824	0,2672	0,2519	0,2863	0,2061	0,3550	0,2290	0,1756	-	0,5237	0,4522	0,7108	0,6909
15	0,1565	0,2863	0,2672	0,2672	0,2901	0,2863	0,2748	0,2710	0,2595	0,2061	0,3053	0,2099	0,1718	0,3282	-	0,3929	0,4974	0,5241
16	0,1908	0,3168	0,2595	0,2748	0,2901	0,2634	0,2634	0,2672	0,2824	0,1832	0,2634	0,1756	0,1527	0,2939	0,2748	-	0,4491	0,4125
17	0,1756	0,3130	0,2786	0,2863	0,2863	0,2748	0,2634	0,2672	0,3053	0,2099	0,3511	0,2328	0,1947	0,3931	0,3282	0,3015	-	0,8886
18	0,1679	0,2901	0,2595	0,2634	0,2710	0,2519	0,2557	0,2557	0,2863	0,2099	0,3435	0,2405	0,1794	0,3817	0,3321	0,2824	0,4466	-

REFERÊNCIAS BIBLIOGRÁFICAS

ANDERBERG, M.R. **Clustering analysis for applications.** London: Academic Press, 1973. 359p.

BARBOSA, A.M.M. Distâncias Genéticas entre linhagens e correlações com as performances de híbridos simples de milho, utilizando marcadores AFLP e SSR. Piracicaba, 2001. 93p. Tese (Doutorado) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

BENCHIMOL, L.L.; SOUZA Jr., C.L.; GARCIA, A.A.F.; KONO, P.M.S.; MANGOLIN, C.A.; BARBOSA, A.M.M.; COELHO, A.S.G.; SOUZA, A.P. Genetic diversity in tropical maize inbred lines: heterotic group assignment and hybrid performance determined by RFLP markers. **Plant Breeding**, v.119, p.491-496, 2000.

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

CARLINI-GARCIA, L.A. Estudo da estrutura genética populacional através de marcadores moleculares. Piracicaba, 1998. 118p. Monografia (Pós-graduação) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

CHATFIELD, C.; COLLINS, A.J. **Introduction to multivariate analysis.** New York: Chapman & Hall, 1980. 246p.

CLIFFORD, H.T.; STEPHENSON, W. **An introduction to numerical taxonomy.** London: Academic Press, 1975. 229p.

CRUZ, C.D.; VIANA, J.M.S. A Methodology of genetic divergence analysis based on sample unit projection on two-dimensional space. **Revista Brasileira de Genética,** v.17, p. 69-73, 1994.

CRUZ, C.D.; RAGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético.** Viçosa: Universidade Federal de Viçosa, 1997. 390p.

CRUZ, CD. Programa Genes: versão Windows; aplicativo computacional em genética e estatística. Viçosa: Universidade Federal de Viçosa, 2001. 648p.

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

DUARTE, M.C.; 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.427-432, 1999.

DUNN, G.; EVERRIT, B. S. **An introduction to mathematical taxonomy.** New York: Cambridge University Press, 1980. 152p.

EVERRIT, B. **Cluster analysis.** London: Heinemann Educational Books, 1974. 136p.

FERREIRA, M.E.; GRATTAPAGLIA, D. **Introdução ao uso de marcadores moleculares em análise genética.** Brasília: EMBRAPA, 1996. 220p.

GARCIA, A.A.F.; BENCHIMOL, L.L.; BARBOSA, A.M.M; GERALDI, I.O.; SOUZA Jr., C.L.; SOUZA, A.P. Comparing of RAPDs, RFLPs, AFLPs and SSRs for computing of genetic divergence in tropical maize inbred lines. **Theoretical and Genetics Journal,** 2002 /No prelo/

GOWER, J.C.; LEGENDRE, P. Metric and Euclidean properties of dissimilarity coefficients. **Journal of Classification,** v.3, p.5-48, 1986.

IEMMA, A.F. Estatística Descritiva. Piracicaba: φσρ publicações, 1992, 182p.

JACCARD, P. Étude comparative de la distribuition florale dans une portion des Alpes et des Jura. **Bulletin de la Societé Voudoise des Sciencies Natureller**, n.37, p.547-579, 1901.

JACKSON, A.A.; SOMERS, K.M.; HARVEY, H.H. Similarity coefficients: measures for co-occurrence and association or simply measures of occurrence? **American Naturalist.,** v.133, p.436-453, 1989.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** New Jersey: Prentice-Hall, 1988. 607p.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis.** 3. ed. New Jersey: Prantice Hall, 1992. 642p.

KAUFMANN, L.; ROSSEEUW, P.J. **Finding groups in data: an introduction to cluster analysis.** New York: Jonh Wiley, 1990. 342p.

KRUSKAl, J.B. Multidimensional scaling by optimizing goodness of fit to a nom-metric hypothesis. **Psychometrika,** v.29, p.1-27, 1964.

LANZA, L.L.B.; SOUZA Jr., C.L.; OTTOBONI, L.M.M.; VIEIRA, M.L.C.; SOUZA, A.P. Genetic distance of inbred lines and prediction of maize single-cross performace using RAPD markers. **Theoretical and Genetics journal,** v.94, p.1023-1030, 1994.

LIU, B.H. **Statistical genomics:** linkage, mapping and QTL analysis. Boca Raton: CRC Press, Boca Raton, 1998.

MANLY, B.F.J. **Multivariate statistical metholds.** 2.ed. London: A Primer, 1994. 215p.

MARDIA, K.V.; KENT, J.T.; BIBBY, J.M. **Multivariate analysis.** London: Academic Press, 1979. 520p.

MORRISON, D.F. **Multivariate statistical metholds.** New York: McGraw-Hill Book Company, 1976. 415p.

NEI,M. Genetic distance beteewm populations. **American Naturalist,** v.106, p.283-292, 1972.

OCHIAI, A. Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions. **Bulletin of the Japanese Society for Fish Science,** v.22, p. 526-530, 1957.

PINTO, R.M.C. Comparações de marcadores moleculares e cruzamentos dialélicos na alocação de linhagens de milho em grupos heteróticos. Piracicaba, 2000. 147p. Tese (Doutorado)- Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo.

PINTO, R.M.C; GARCIA, A.A.F. Alocação de linhagens de milho derivadas das populações BR – 105 e BR – 106 em grupos heteróticos. **Scientia Agricola,** n.3**,** v.58, p.541-548, 2001.

RAO, R.C. **Advanced statistical methods in biometric reserch.** New York: J. Wiley, 1952. 390p.

ROGERS, J.S; TANIMOTO, T.T. A computer program for classying plants. **Science,** n132, p.1115-1118, 1960.

ROHLF, F.J. Consesurs indices for comparing classifications. **Mathematical Bioscience,** n.59, p.131-144, 1982.

ROHLF, F.J. Numerical taxonomy and multivariate analysis system: version 1.70, (software). New York, 1992. 470p.

ROMESBURG, H.C. **Cluster analysis for researchers.** California: Lifetime Learning Publications, 1984. 334p.

RUSSEL, P.F.; RAO, T.R. On habitat and association of species of anopheline larvae in south-eastern Madras. **Journal of Malaria India Institute,** n3, p.153-178, 1940.

SAS INSTITUTE. Statistical analysis system: release 6.08, (software). Cary, 1992. 620p.

SKROCH, P.; TIVANG, J.; NIENHUIS, J. Analysis of genetic relationships using RAPD marker data in: applications of RAPD technology to plant breeding symposia series, Madison. **CCSA, ASHS, and AGMA,** p.26-30, 1992.

SNEATH, P.H.A.; SOKAL, R.R. **Numeric taxonomy:** the principles and practice of numerical classification**.** San Francisco: W.H. Freeman, 1973. 573p.

SOKAL, R.R.; MICHENER, C.D. A statistical method for evaluating systematic relationships. **Bulletin of the Society University of Kansas,** n.38, p.1409-1438, 1958.

SOKAL, R.R.; SNEATH, P.H. A. **Principles of numeric taxonomy.** San Francisco: W.H. Freeman, 1963. 359p.

STAT SOFT STATISTICA. ('99 Edition)-Quick Reference. Tulsa, 1999. 231p.

TOTTI, R. Utilização de métodos de agrupamentos hierárquicos em acessos de *paspalum* (Gramineae (Poaceae)). Piracicaba, 1998. 75p. Dissertação (Mestrado) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo.

VICTÓRIA, D.C.; GARCIA, A.A.F.; SOUZA JR., C.L; SOUZA, A.P. Desenvolvimento de um programa SAS para cálculo de coeficiente de similaridade de dados de marcadores moleculares utilizando bootstrap. (compact disc) In: CONGRESSO NACIONAL DE GENÉTICA, 47, Águas de Lindóia: Sociedade Brasileira de Genética, 2001.

VOS, P.; ROGERS, R.; BLEEKER, M.; REIJANS, M.; VAN DE LEE, T.; HORNES, M.; FRIJTERS, A.; POT, J.; PELEMAN, J.; KUIPE, M.; ZABEAU, M. AFLP: A new tecnique for DNA fingerprinting. **Nucleic Acids Research,** v. 23, p. 4407-4414, 1995.

WEIR, B.S. **Genetic data analysis II.** 2. ed. Sunderland: Sinauer Associates, 1996. 445p.

WILLIANS, J.G.K; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A.; TINGEY, S.V. DNA polimorfisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research,** v. 18, p.7213-7218, 1990.

APÊNDICES

APÊNDICE 1

Exemplo: de cálculo de matrizes de similaridade

1..Dados de marcadores

Foi considerado um exemplo hipotético, com 5 linhagens e um marcador molecular do tipo dominante. Os dados de presença (1) ou ausência (0) de bandas, estão apresentados na Tabela 1.

Tabela A1. Dados fictícios considerando-se 5 linhagens e 30 marcadores (bandas). A presença da banda é codificada como 1 e a ausência como 0.

Bandas	L1	L2	L3	L4	L5
1	0	0	0	0	0
2	0	0	0	0	0
3	1	1	1	1	1
4	1	0	0	1	0
5	0	1	1	0	1
6	1	0	0	0	0
7	1	0	0	0	0
8	0	0	0	0	0
9	0	1	1	1	1
10	0	1	1	1	1
11	1	0	0	0	0
12	0	0	0	0	0
13	1	1	1	1	1
14	1	1	1	1	1
15	1	0	0	0	0
16	0	0	0	0	0
17	1	0	0	1	0
18	0	0	0	0	0
19	0	0	0	0	0
20	1	0	0	0	0
21	0	0	0	0	0
22	1	1	1	1	1
23	0	0	0	0	0
24	0	0	0	0	0
25	0	0	1	0	1
26	1	1	1	1	0
27	1	0	0	0	0
28	1	1	1	1	1
29	0	0	0	0	0
30	0	0	0	1	0

A partir dos dados da Tabela 1, foram obtidas estimativas de similaridade e dissimilaridade entre cada par de linhagens *i* e *j*, para a construção da matriz de similaridades. Por exemplo, para as linhagens 1 e 2, tem-se (Tabela 2):

Tabela A2. Similaridades e dissimilaridades entre as linhagens 1 e 2.

		Linhagem 1		
		1	0	
Linhagem 2	1	$a = 6$	$b = 8$	$a + b = 14$
	0	$c = 3$	$d = 13$	$c + d = 16$
		$a + c = 9$	$b + d = 21$	$p = a + b + c + d = 30$

2 Obtenção das matrizes de similaridade

Para calcular a similaridade entre as linhagens, foram utilizados os coeficientes de Jaccard (Jaccard, 1901), Sorensen-Dice (Dice, 1945), Anderberg (Anderberg, 1973), Ochiai (Ochiai, 1957), Simple Matching (Sokal & Michener, 1958), Rogers e Tanimoto (Rogers & Tanimoto, 1960), Ochiai II (Ochiai, 1957) e Russel e Rao (Russel & Rao, 1940).

Usando novamente as linhagens 1 e 2 como exemplo, os cálculos para a construção das matrizes de similaridades são obtidos do seguinte modo ($S_{(1,2)}$ = similaridade entre as linhagens 1 e 2):

a) Coeficiente de Jaccard

$$S_{(1,2)} = \frac{a}{a+b+c} = \frac{6}{6+8+3} = 0{,}3529$$

b) Sorensen-Dice

$$S_{(1,2)} = \frac{2a}{2a+b+c} = \frac{12}{12+8+3} = 0{,}5217$$

c) Simple Matching

$$S_{(1,2)} = \frac{a+d}{a+b+c+d} = \frac{6+13}{6+8+3+13} = 0{,}6333$$

d) Rogers e Tanimoto

$$S_{(1,2)} = \frac{a+d}{a+2b+2c+d} = \frac{6+13}{6+16+6+13} = 0{,}4634$$

e) Anderberg

$$S_{(1,2)} = \frac{a}{a+2(b+c)} = \frac{6}{6+2(8+3)} = 0{,}2143$$

f) Russel e Rao

$$S_{(1,2)} = \frac{a}{a+b+c+d} = \frac{6}{6+8+3+13} = 0{,}2000$$

g) Ochiai

$$S_{(1,2)} = \frac{a}{\sqrt{(a+b)(a+c)}} = \frac{6}{\sqrt{(6+8)+(6+3)}} = 0{,}5345$$

h) Ochiai II

$$S_{(1,2)} = \frac{ad}{\sqrt{(a+b)(a+c)(d+b)(d+c)}} = \frac{78}{\sqrt{(6+8)(6+30(13+8)(13+3)}} = 0{,}3791$$

Procedendo do mesmo modo com as demais linhagens, obtém-se as seguintes matrizes de similaridade (Tabela 3):

Tabela A3. Matrizes de similaridade obtidas com diversos coeficientes.

Jaccard				
	L2	L3	L4	L5
L1	0,3529	0,3333	0,4706	0,2778
L2		0,9000	0,6667	0,8000
L3			0,6154	0,9000
L4				0,5385
Sorensen-Dice				
	L2	L3	L4	L5
L1	0,5217	0,5000	0,6400	0,4348
L2		0,9474	0,8000	0,8889
L3			0,7619	0,9474
L4				0,7000

Tabela A3. Matrizes de similaridade obtidas com diversos coeficientes (continuação).

		Concordância Simples		
	L2	L3	L4	L5
L1	0,6333	0,6000	0,7000	0,5667
L2		0,9667	0,8667	0,9333
L3			0,8333	0,9667
L4				0,8000
		Rogers		
	L2	L3	L4	L5
L1	0,4634	0,4286	0,5385	0,3953
L2		0,9355	0,7647	0,8750
L3			0,7143	0,9355
L4				0,6667
		Anderberg		
	L2	L3	L4	L5
L1	0,2143	0,2000	0,3077	0,1613
L2		0,8182	0,5000	0,6667
L3			0,4444	0,8182
L4				0,3684
		Russel		
	L2	L3	L4	L5
L1	0,2000	0,2000	0,2667	0,1667
L2		0,3000	0,2667	0,2667
L3			0,2667	0,3000
L4				0,2333
		Ochiai		
	L2	L3	L4	L5
L1	0,5345	0,5071	0,6447	0,4454
L2		0,9487	0,8040	0,8889
L3			0,7628	0,9487
L4				0,7035
		Ochiai II		
	L2	L3	L4	L5
L1	0,3791	0,3402	0,4807	0,2916
L2		0,9258	0,7245	0,8466
L3			0,6652	0,9258
L4				0.5987

APÊNDICE 2

Exemplo: de cálculo da correlação de Spearman

1. Coeficiente de correlação de Spearman.

O coeficiente de correlação ordinal ou de Spearman (r_s) é obtido com base na ordem (ou posto) dos dados:

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)},$$

em que *n* é o número de pares e d_i é a diferença entre os postos para as variáveis.

Usando como exemplo as matrizes dos coeficientes de Jaccard e Sorensen-Dice, provenientes do Apêndice 1, são feitos os seguintes cálculos (Tabela A4):

Tabela A4: Cálculo da correlação de Spearman entre os coeficientes de Jaccard e Dice. Dados hipotéticos.

Par de linhagens	Jaccard	Ordem	Dice	Ordem	d_i	d_i^2
L1, L2	0,3529	3	0,5217	3	0	0
L1, L3	0,3333	2	0,5000	2	0	0
L1, L4	0,4706	4	0,6400	4	0	0
L1, L5	0,2778	1	0,4348	1	0	0
L2, L3	0,9000	9,5	0,9474	9,5	0	0
L2, L4	0,6667	7	0,8000	7	0	0
L2, L5	0,8000	8	0,8889	8	0	0
L3, L4	0,6154	6	0,7619	6	0	0
L3, L5	0,9000	9,5	0,9474	9,5	0	0
L4, L5	0,5385	5	0,7000	5	0	0

Usou-se valor 9,5 na ordem porque houve empate, ou seja, dois pares de linhagem apresentaram mesma classificação. Tomou-se então a média das ordens 8 e 10. Os valores de d_i representam a diferença entre as ordens.

A correlação entre elas é:

$$r_s = 1 - \frac{0}{10(99)} = 1$$

Pois $\sum d_i^2$ é igual a zero.

APÊNDICE 3

Exemplo: da construção do dendrograma

1. Construção do dendrograma

Neste exemplo ilustrativo, usou-se o método UPGMA, que é um processo hierárquico, e em cada passo diminui uma dimensão da matriz de semelhança, pela reunião de pares semelhantes até reunir todos os pontos em um único grupo. Para tanto, foi considerada a matriz de semelhança obtida para os dados hipotéticos da Tabela A1 (Tabela A5)

Tabela A5. Matrizes de similaridade usando o coeficiente de Jaccard.

	Jaccard			
	L2	L3	L4	L5
L1	0,3529	0,3333	0,4706	0,2778
L2		0,9000	0,6667	0,8000
L3			0,6154	0,9000
L4				0,5385

Na aplicação desse algoritmo, o primeiro passo é observar na matriz de similaridade quais linhagens são mais próximas. Nesse exemplo, os pares mais similares serão aqueles que apresentarem maiores valores, pois, usou-se um coeficiente de similaridade. Percebe-se que há dois pares de linhagens com o mesmo valor de proximidade L2-L3 e L3-L5. Nesses casos, o processo recomenda selecionar aleatoriamente um destes pares para criar um novo grupo. De acordo com Bussab et al. (1990) e Sneath & Sokal (1973), os pacote computacionais, por facilidade de programação, escolhem o primeiro par que aparece para agrupar. Assim, seram agrupados L2 e L3, formando um único grupo, e obtendo desse modo 4 grupos:

L1, (L2L3), L4 e L5

É necessário reconstruir uma nova matriz de similaridade considerando o novo grupo formado. Como as linhagens L1, L4 e L5 não sofreram alterações, as distâncias entre elas continuam as mesmas. Contudo, deve-se calcular as distâncias entre

o novo grupo (L2 L3) e as demais linhagens. Vale ressaltar que é nessa etapa que a maioria dos métodos hierárquicos se diferenciam.

O método UPGMA define a distância entre dois grupos como a média não ponderada dos valores individuais de similaridade dos objetos com os valores de similaridade de cada um dos outros indivíduos do outro grupo:

$$S_{L1,(L2L3)} = \frac{1}{2}(S_{L1,L2} + S_{L1,L3)} = \frac{1}{2}(0{,}3529 + 0{,}3333) = 0{,}3431$$

$$S_{L4,(L2L3)} = \frac{1}{2}(S_{L4,L2} + S_{L4,L3}) = \frac{1}{2}(0{,}6667 + 0{,}6154) = 0{,}6411$$

$$S_{L5,(L2L3)} = \frac{1}{2}(S_{L5,L2} + S_{L5,L3}) = \frac{1}{2}(0{,}8000 + 0{,}9000) = 0{,}8500$$

E assim, obteve-se a nova matriz de similaridade:

	L4	L5	L2L3
L1	0,4706	0,2778	0,3431
L4		0,5385	0,6411
L5			0,8500

Analisando a nova matriz, nota-se que os pares mais similares são L5 e L2L3. Como no processo anterior, as distâncias entre L1 e L4 não se alteram. As distâncias de L2L3L5 com as demais serão:

$$S_{L1,(L2L3L5)} = \frac{1}{3}(S_{L1,L2} + S_{L1,L3} + S_{L1,L5}) = \frac{1}{3}(0{,}3529 + 0.3333 + 0.2778) = 0{,}3213$$

$$S_{L4,(L2L3L5)} = \frac{1}{3}(S_{L4,L2} + S_{L4,L3} + S_{L4,L5}) = \frac{1}{3}(0{,}6667 + 0{,}6154 + 0{,}0{,}5385) = 0{,}6069$$

Assim, tem-se:

	L4	L2L3L5
L1	0,4706	0,3213
L4		0,6069

O próximo passo consistiu em agrupar L4 com L2L3L5, repetindo-se novamente todo o processo:

$$S_{L1,(L2\,L3L4L5)} = \frac{1}{4}(S_{L1,L2} + S_{L1,L3} + S_{L1,L4} + S_{L1,L5}) = \frac{1}{4}(0{,}3529 + 0{,}3333 + 0{,}4706 + 0{,}2778) = 0{,}3587$$

Obtém-se assim:

	L2L3L4L5
L1	0,3587

O processo encerra reunindo num único grupo os conjuntos L2 L3 L4 L5 com L1, que foram iguais a um nível 0,3587 de similaridade.

Os resultados desses agrupamentos, são representados graficamente, através do dendrograma, apresentado a seguir (Figura A1). Verifica-se que essa representação gráfica ilustra todos os passos do algoritmo, mostrando as distâncias entre os grupos formados.

Caso o agrupamento tivesse iniciado com L3 e L5, a única mudança seria nos dois primeiros agrupamentos; para os demais, a seqüência permaneceria a mesma.

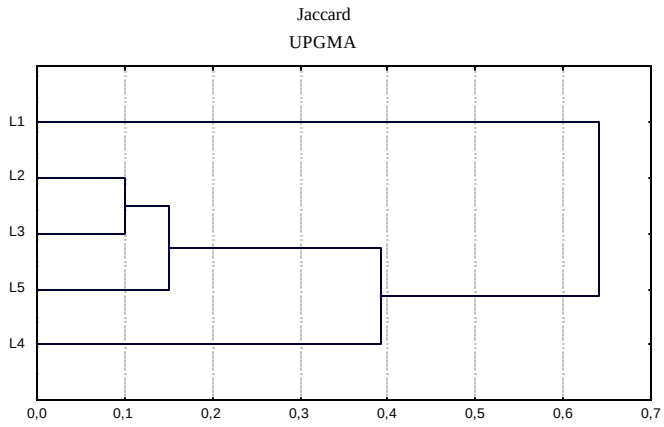

Figura A1 - Dendrograma obtido com o coeficiente de Jaccard e método UPGMA, para os dados da Tabela A5.

APÊNDICE 4

Exemplo: de cálculo da correlação cofenética, grau de distorção e estresse

1 Correlação cofenética

Tal correlação é calculada por meio de (Bussab et al., 1990):

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\overline{c})(s_{ij}-\overline{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\overline{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-\overline{s})^2}}, \text{ em que}$$

c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade;

$$\overline{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}c_{ij};$$

$$\overline{s} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}s_{ij}.$$

Usando novamente os dados hipotéticos apresentados na Tabela A1 e o dendrograma apresentado na Figura A1, obtém-se a seguinte matriz cofenética, que representa o dendrograma

Tabela A6. Matriz cofenética entre as linhagens do exemplo hipotético

	L2	**L3**	**L4**	**L5**
L1	0,6413	0,6413	0,6413	0,6413
L2		0,1000	0,3931	0,1500
L3			0,3931	0,1500
L4				0,3931

Considerando a matriz de dissimilaridade correspondente, usando Jaccard:

Tabela A7. Matriz de dissimilaridade entre as linhagens do exemplo hipotético

	L2	**L3**	**L4**	**L5**
L1	0,6471	0,6667	0,5294	0,7222
L2		0,1000	0,3333	0,2000
L3			0,3846	0,1000
L4				0,4615

Obtém-se assim o valor da correlação cofenética:

$$r_{cof} = 0{,}9649$$

2 Grau de distorção

O grau de distorção (1-α) para essas matrizes é calculado usando (Kruskal, 1964):

$$\alpha = \frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n} c_{ij}}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} s_{ij}}, \text{ em que}$$

c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade.

O valor nesse exemplo é: $\alpha = 0{,}9999$. Portanto o grau de distorção foi igual a 1- α = 0,0001, ou 0,01% de distorção.

3 Estresse

O estresse é calculado por (Kruskal, 1964):

$$S = \sqrt{\frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(s_{ij} - c_{ij})^2}{\sum_{i=1}^{n-1}\sum_{j=2}^{n} s_{ij}}}$$, em que

c_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

s_{ij}: valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade;

Para o exemplo em questão, o valor do estresse foi igual a 12,27%.

4 Índice de Consenso

Para aplicar o Índice de Consenso, foram tomados dois dendrogramas construídos a partir das matrizes de similaridade apresentadas na Tabela A3. Os Dendrogramas foram construídos usando o método UPGMA, para os coeficientes de Jaccard e de Sorensen-Dice (Figuras A2 e A3)

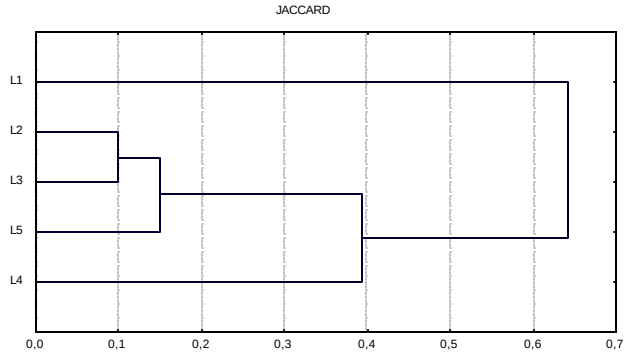

Figura A2 – Dendrograma usando método UPGMA para o coeficiente de Jaccard.

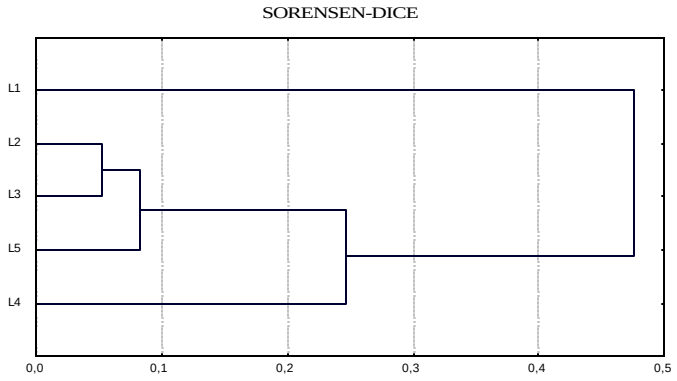

Figura A3 - Dendrograma usando método UPGMA para o coeficiente de Jaccard.

Nesse caso, temos n - 2 = 5 - 2 possíveis passos. No primeiro deles, em ambos os casos foram agrupados L2 e L3. No segundo, L2L3 e L5 e no terceiro L4 com L2L3L5. Assim, tem-se:

$$CI_c = \frac{3}{5-2} = 1,$$

uma vez que todos coincidiram.

APÊNDICE 5

Exemplo: do método de agrupamento de Tocher

1 Método de Otimização de Tocher.

No presente caso, o método de otimização de Tocher, citado por Rao (1952), foi aplicado aos dados da Tabela A1 (Apêndice A1) e A7 (Apêndice A4).

O critério usado para considerar a inclusão de uma linhagem num grupo baseou-se no valor máximo da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo, conforme será mostrado.

O processo teve os seguintes passos (S_{ij} = similaridade entre as linhagens i e j):

i) Estimação da maior distância, dentre o conjunto de menores distâncias, entre cada linhagens:

Linhagens	L1	L2	L3	L4	L5
$d = \text{mín}(1 - S_{ij})$	0,5294	0,1000	0,1000	0,3333	0,1000

Assim, é estabelecido α = 0,5294 (maior valor de d) como o limite de acréscimo na média das distâncias intragrupo, para a inclusão de um elemento ou formação de um novo grupo.

ii) Formação dos grupos.

Assim como para os dendrogramas, a formação dos grupos se inicia agrupando as linhagens mais similares. Portanto, como neste caso trabalhou-se com matriz de dissimilaridades, os pares mais similares foram aqueles que apresentarem os menores valores.

O par de linhagens mais similares foi contituído por L2 e L3, portanto o grupo I contém tais linhagens. Uma vez formado o grupo, calculam-se as medidas de dissimilaridade entre este grupo e as demais linhagens. Como o critério de inclusão de

novos objetos em um grupo baseia-se no acréscimo da distância média do grupo, é apropriado o cálculo da distância entre linhagem-grupo, por meio de

$$d_{(ij)k} = d_{ik} + d_{jk}$$

em que d é qualquer medida de dissimilaridade. Assim, para o exemplo em questão, tem-se:

$$d_{(2,3)1} = d_{(1,2)} + d_{(3,1)} = 0{,}6471 + 0{,}6667 = 1{,}3138$$

$$d_{(2,3)4} = d_{(2,5)} + d_{(3,4)} = 0{,}3333 + 0{,}3846 = 0{,}7179$$

$$d_{(2,3)5} = d_{(2,5)} + d_{(3,5)} = 0{,}2000 + 0{,}1000 = 0{,}3000$$

Com base nas distâncias entre linhagem-grupo, notou-se que a linhagem L5 é a mais similar ao grupo formado por L2 e L3. O acréscimo médio no valor de d a esse grupo, pela inclusão desta linhagem L5, é dado por:

$$\frac{d_{(2,3,5)} - d_{(2,3)}}{C_{3,2} - C_{2,2}} = \frac{d_{(2,3)5}}{2} = \frac{0{,}3000}{2} = 0{,}1500\ ,$$

em que $C_{3,2}$ e $C_{2,2}$ são, respectivamente, combinação de três, dois a dois, e combinação de dois, dois a dois. Como o valor 0,1500 é inferior à α (passo i), a inclusão da linhagem L5 no grupo I foi permitida.

Uma maneira mais simples para se avaliar a possibilidade de inclusão de uma linhagem i em um grupo, é obtida usando o critério a seguir (n = número de linhagens que constitui o grupo original):

Se $\dfrac{d_{(grupo)i}}{n} \leq \alpha \Rightarrow$ inclui-se a linhagem i ao grupo;

Se $\dfrac{d_{(grupo)i}}{n} > \alpha \Rightarrow$ a linhagem i não deve ser incluída ao grupo.

As novas distâncias das linhagens não incluídas, ao grupo formado (2, 3, 5), foram:

$$d_{(2,3,5)1} = d_{(2,3)1} + d_{(5)1} = 1{,}3138 + 0{,}7222 = 2{,}0360$$

$$d_{(2,3,5)4} = d_{(2,3)4} + d_{(5)4} = 0{,}7179 + 0{,}4615 = 1{,}1794$$

A linhagem L4 é a mais similar ao grupo que está sendo formado. A inclusão da linhagem é avaliada por meio de:

$$\frac{d_{(2,3,5)4}}{3} = \frac{1{,}1794}{3} = 0{,}2949$$

Como 0,2949 é menor que α, a inclusão da linhagem L4 foi permitida.

As novas distâncias das linhagens não incluídas, em relação ao grupo em formação (2, 3, 4, 5), foram:

$$d_{(2,3,4,5)1} = d_{(2,3,5),1} + d_{(4)1} = 2{,}0360 + 0{,}5294 = 2{,}5654$$

Avaliando a possibilidade de inclusão, obteve-se:

$$\frac{d_{(2,3,4,5)1}}{4} = \frac{2{,}5654}{4} = 0{,}6414$$

Como 0,6414 é maior que α, a L1 não foi incluída no grupo I.

Portanto, o agrupamento com base no Tocher permitiu o estabelecimento dos grupos apresentados na tabela A8.

Tabela A8. Grupos obtidos com o método de otimização de Tocher para os dados da Tabela A1 do Apêndice 1.

Grupo	Linhagens
I	L2, L3, L4, L5
II	L1

APÊNDICE 6

Exemplo: de obtenção de projeção no plano bidimensional

1 Projeção da matriz de dissimilaridade no plano bidimensional

A abordagem de Cruz e Viana (1994) consiste em realizar a projeção da matriz de dissimilaridade num espaço bidimensional, baseada nos seguintes passos:

i) Estimativa da ordem de estabelecimento de cálculo de distância no espaça bidimensional

Para o exemplo em questão (Tabela A7), o par de indivíduos L1 e L5 foi o que apresentou o maior valor de dissimilaridade, sendo eles considerados os dois primeiros indivíduos.

Os próximos indivíduos mais divergentes foram aqueles que apresentaram maiores valores de $d_{(51)k}$, dados por:

$$d_{(51)k} = d_{5k} + d_{1k}$$

$$d_{(51)2} = d_{52} + d_{12} = 0,6471 + 0,2000 = 0,8471$$

$$d_{(51)3} = d_{53} + d_{13} = 0,6667 + 0,1000 = 0,7667$$

$$d_{(51)4} = d_{54} + d_{14} = 0,5294 + 0,4615 = 0,9909$$

A linhagem 4 foi a que apresentou o maior valor; portanto, essa linhagem foi considerada o terceiro indivíduo mais divergente. Em seguida, obtém-se:

$$d_{(514)2} = d_{52} + d_{12} + d_{42} = 0,6471 + 0,2000 + 0,3333 = 1,1804$$

$$d_{(514)3} = d_{53} + d_{13} + d_{43} = 0,6667 + 0,1000 + 0,3846 = 1,1513$$

Nota-se que, o quarto e o quinto indivíduo mais divergente são as linhagem 2 e 3, respectivamente.

Assim, é possível obter uma ordem decrescente de dissimilaridade: L5, L1, L4, L2 e L3.

ii) Estimativas das coordenadas a serem utilizadas na dispersão gráfica

As coordenadas para os dois primeiros indivíduos (5 e 1) são estabelecidas arbitrariamente, como $(0,0)$ e $(d_{51},0)$ respectivamente. As coordenadas para o terceiro indivíduo foram obtidas por:

$$x_4 = \frac{d_{14}{}^2 - d^2{}_{54} - d_{51}{}^2}{-2d_{51}} = 0{,}3145$$

$$y_4 = (d_{54}{}^2 - X_4{}^2)^{\frac{1}{2}} = 0{,}3377$$

As coordenadas para o quarto e sucessivos indivíduos foram estimadas por:

$$C = P^{-1}Q;$$

sendo

$$C = [k\ x_l\ y_l], \text{ em que}$$

k : constante;

x_l : abscissa do indivíduo l;

y_l : ordenada do indivíduo l;

$$P = \begin{bmatrix} m & -2\sum_m x_m & -2\sum_m y_m \\ -2\sum_m x_m & 4\sum_m x_m{}^2 & 4\sum_m x_m y_m \\ -2\sum_m y_m & 4\sum_m x_m y_m & 4\sum_m y_m{}^2 \end{bmatrix} \text{ e } Q = \begin{bmatrix} \sum d^2{}_{lm} - \sum (x^2{}_m + y^2{}_m) \\ -2\sum x_m d^2{}_{lm} + 2\sum x_m (x^2{}_m + y^2{}_m) \\ -2\sum y_m d^2{}_{lm} + 2\sum y_m (x^2{}_m + y^2{}_m) \end{bmatrix}$$

O cálculo das coordenadas para o quarto indivíduo (3) foram estimadas com base nos seguintes cálculos:

Tabela A9. Cálculos intermediários para obtenção das coordenadas no plano bidimensional.

Ordem	linhagem	X_m	Y_m	X_m^2	Y_m^2	X_mY_m
1	5	0,0000	0,0000	0,0000	0,0000	0,0000
2	1	0,7222	0,0000	0,5216	0,0000	0,0000
3	4	0,3145	0,3377	0,0989	0,1140	0,1062

Tablela A10. Cálculos intermediários para obtenção das coordenadas no plano bidimensional.

Ordem	Linhagem	$X_m^2+Y_m^2$	$X_m(X_m^2+Y_m^2)$	$Y_m(X_m^2+Y_m^2)$	d_{2m}^2	$X_md_{2m}^2$	$Y_md_{2m}^2$
1	5	0,0000	0,0000	0,0000	0,0400	0,0000	0,0000
2	1	0,5216	0,3767	0,0000	0,4187	0,3024	0,0000
3	4	0,2129	0,0670	0,0719	0,1111	0,0349	0,0375

$$P = \begin{bmatrix} 3 & -2,0734 & -0,6754 \\ -2,0734 & 2,4820 & 0,4248 \\ -0,6754 & 0,4248 & 0,4560 \end{bmatrix},$$

$$Q = \begin{bmatrix} -0,1647 \\ 0,2128 \\ 0,0688 \end{bmatrix},$$

$$C' = \begin{bmatrix} 0,0401 \\ 0,0991 \\ 0,1181 \end{bmatrix}$$

Portanto as coordenadas para o quarto indivíduo (2) são 0,0991 e 0,1181.

As coordenadas para o quinto indivíduo(3) foram obtidas de forma análoga. Finalmente, as coordenadas são apresentadas na Tabela A11.

Tabela A11. Coordenadas estimadas para as cinco linhagens no espaço bidimensional.

Linhagem	X_m	Y_m
1	0,7222	0,0000
2	0,0991	0,1181
3	0,0607	0,0622
4	0,3145	0,3377
5	0,0000	0,0000

Tais coordenadas são então projetadas num gráfico bidimensional (Figura A4).

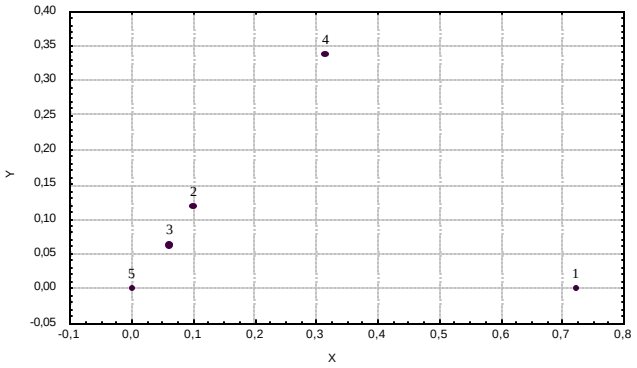

Figura A4– Projeção de 5 linhagens de milho no plano bidimensional com base nas coordenadas da Tabela A11.

iii) Cálculo da eficiência da projeção gráfica

Inicialmente, é necessário calcular uma nova matriz de dissimilaridade, usando a distância Euclidiana

Tabela A12. Distância euclidiana entre as linhagens com base na projeção bidimensional.

	L2	L3	L4	L5
L1	0,6342	0,6644	0,5294	0,7222
L2		0,0678	0,3076	0,1542
L3			0,3746	0,0869
L4				0,4615

Essa matriz é comparada com a matriz original de dissimilaridades (Tabela A7)

a) Correlação entre as distâncias originais e as distâncias obtidas pela representação gráfica da dispersão bidimensional:

É obtida a correlação entre as matrizes das Tabelas A7 e A12, resultando no valor $(r) = 0.9582$

b) Grau da distorção (1 - α), dado por:

$$\alpha = \frac{\sum_{i<j} dg_{ij}}{\sum_{i<j} do_{ij}}, \text{ sendo } \alpha = \frac{4{,}0028}{4{,}1448} = 0{,}9657$$

Portanto o grau da distorção foi igual a 1 - α = 0,0343, ou 3,43 (%) de distorção.

c) Valor do estresse *(S)*, dado por:

$$S = \sqrt{\frac{\sum_{i<j} (do_{ij} - dg_{ij})^2}{\sum_{i<j} do^2{}_{ij}}}.$$

Para o exemplo em questão, o valor do estresse foi igual a 4,41%.