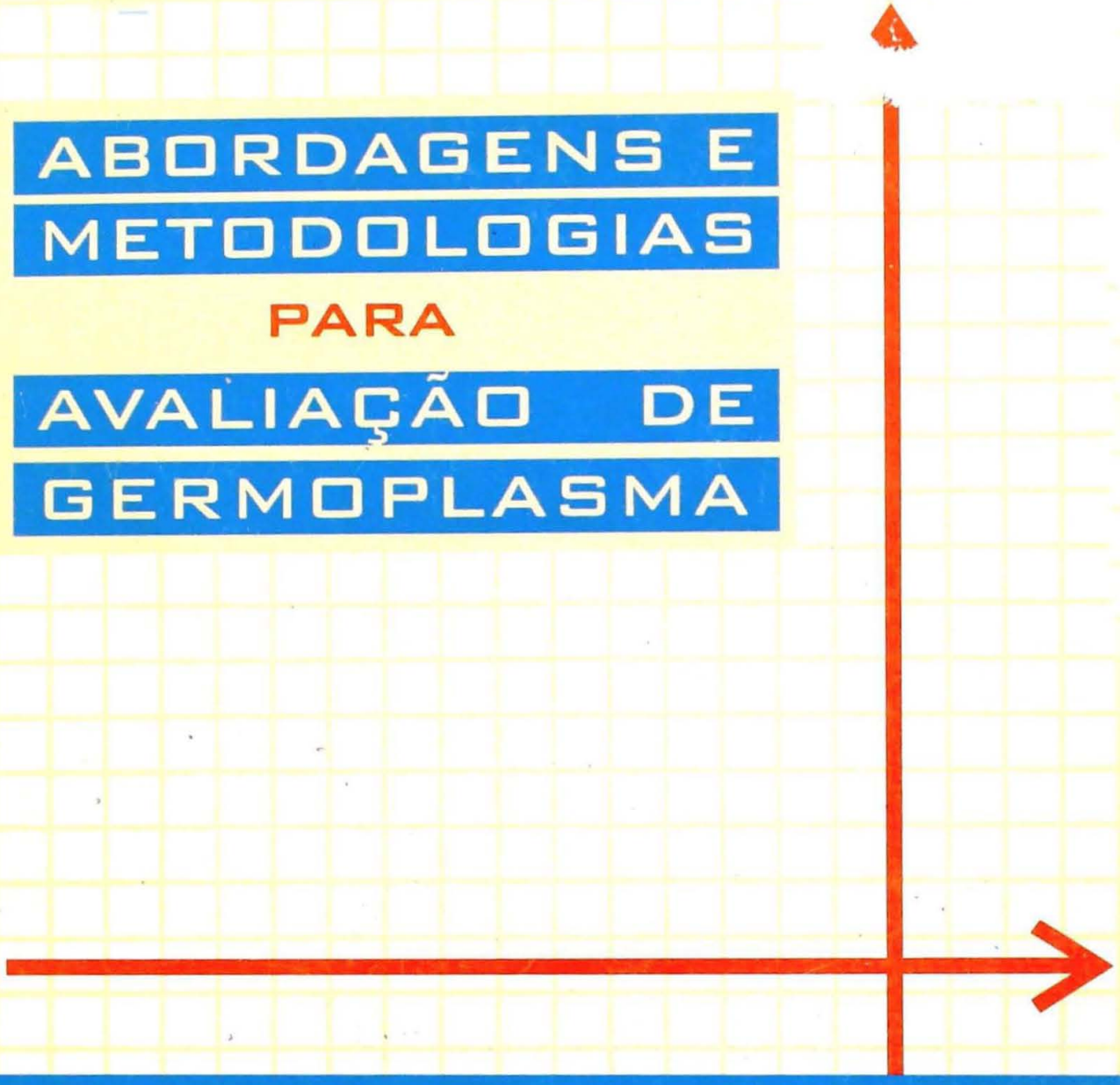
ABORDAGENS E METODOLOGIAS PARA AVALIAÇÃO DE GERMOPLASMA



JOSÉ DE ALENCAR NUNES MOREIRA
JOSÉ WELLINGTON DOS SANTOS
ROBSON DE MEDEIROS OLIVEIRA

Ministério da Agricultura, do Abastecimento e da Reforma Agrária-MAARA
Empresa Brasileira de Pesquisa Agropecuária - EMBRAPA
Centro Nacional de Pesquisa de Algodão - CNPA
Campina Grande, Paraíba, Brasil

ABORDAGENS E METODOLOGIAS PARA AVALIAÇÃO DE GERMOPLASMA

JOSÉ DE ALENCAR NUNES MOREIRA
JOSÉ WELLINGTON DOS SANTOS
STANLEY ROBSON DE MEDEIROS OLIVEIRA

EMBRAPA-CNPA/SPI
1994

Exemplares desta publicação podem ser solicitadas à
EMBRAPA/CNPA

Rua Osvaldo Cruz nº 1143 - Centenário
Caixa Postal 174
Telefone (083) 341-3608
Fax (083) 322-7751
Telex (083) 3231
58.107-720 - Campina Grande, PB

Tiragem: 800 exemplares

Capa: Die Presse Editorial Ltda
Designer: Ewandro Magalhães Junior

Moreira, José de Alencar Nunes.
Abordagens e metodologias para avaliação de germoplasma / José de Alencar Nunes Moreira, José Wellington dos Santos, Stanley Robson de Medeiros Oliveira ; Empresa Brasileira de Pesquisa Agropecuária, Centro Nacional de Pesquisa de Algodão. — Campina Grande : EMBRAPA-CNPA ; Brasília : EMBRAPA-SPI, 1994.
115p.

ISBN: 85-85760-01-X

1. Germoplasma. 2. Análise Multivariada. I. Santos, José Wellington dos. II. Oliveira, Stanley Robson de Medeiros. III. Empresa Brasileira de Pesquisa Agropecuária. Centro Nacional de Pesquisa de Algodão (Campina Grande, PB). IV. Título.

CDD 575.1

APRESENTAÇÃO

A organização de uma extensa coleção de germoplasma é o passo inicial na longa trajetória de um programa de melhoramento genético fadado ao sucesso. A coleção, na verdade, funciona como um rico manancial que permite ao melhorista renovar a variabilidade quando esta vai sendo exaurida no decurso dos processos de seleção.

É por esta razão que vem assumindo tanta importância o estudo destas coleções, em especial o realizado no Brasil, sob os auspícios da rede dos Bancos Ativos de Germoplasma com as diversas culturas, coordenado pelo Centro Nacional de Recursos Genéticos e Biotecnologia.

O Centro Nacional de Pesquisa de Algodão (CNPA) no ensejo do lançamento da presente publicação não teve outro intuito senão o de oferecer sua contribuição, mesmo que modesta, a assunto de tanta relevância como é este direcionado à avaliação dos germoplasma.

ROBSON DE MACÊDO VIEIRA
Chefe do CNPA

PREFÁCIO

O armazenamento da variabilidade em grandes coleções de germoplasma é tarefa básica para o bom andamento de qualquer programa de melhoramento. À custa deste valioso recurso, os melhoristas contam com uma espécie de reserva gênica que pode ser acionada à medida que a variabilidade vai sendo exaurida com o progresso do melhoramento.

Acontece que não basta apenas organizar a coleção, sendo necessário, em complemento, analisá-la segundo diferentes abordagens para que os dados obtidos possam servir aos propósitos, quer do melhoramento ou para estudos outros, como os relacionados com a evolução da espécie em consideração. Resulta, daí, a grande dificuldade relacionada com o manuseio das metodologias adequadas para analisar os dados obtidos em tais estudos. Em primeiro lugar, pela complexidade matemática envolvida no uso desses métodos, os quais englobam, via de regra, procedimentos da análise estatística multivariada. Em segundo, porque eles se acham descritos em diferentes fontes, o que, nem sempre, os torna acessíveis ao público interessado nas suas aplicações.

O presente trabalho procura sanar estas dificuldades e enfoca as metodologias para três das abordagens mais comumente utilizadas nos estudos com as coleções de germoplasma. É seu objetivo propiciar uma espécie de roteiro prático para os interessados em aplicar estas metodologias na análise dos dados de tais coleções.

SUMÁRIO

1. Abordagens e Metodologias para avaliação de germoplasma 9
   1.1. Introdução 9
   1.2. Principais Abordagens 9
   1.3. Metodologias Adotadas 11
2. Análise dos componentes principais 13
   2.1. Introdução 13
   2.2. Desenvolvimento da metodologia 13
   2.3. Exemplo 17
3. Análise das coordenadas principais 26
   3.1. Introdução 26
   3.2. Desenvolvimento da metodologia 26
   3.3. Exemplo 30
4. Análise de agrupamento 32
   4.1. Introdução 32
   4.2. Medidas de similaridade 32
      4.2.1. Distância de Mahalanobis 32
      4.2.2. Coeficiente simples de emparelhamento 39
      4.2.3. Coeficiente de Jaccard 41
      4.2.4. Distância de Rogers 42
   4.3. Métodos de agrupamento 45
      4.3.1. Métodos aglomerativos 45
         4.3.1.1. Método do vizinho mais próximo 46
         4.3.1.2. Método do vizinho mais distante 50
      4.3.2. Métodos de otimização 51
         4.3.2.1. Método de Tocher 51
5. Metodologias para o estudo da diversidade genética 57
   5.1. Introdução 57
   5.2. Índice de Shannon e Weaver 57
      5.2.1. Exemplo 58
   5.3. Análise da diversidade genética de Nei 64
      5.3.1. Exemplo 64
   5.4. Distância genética de Nei 72

5.4.1. Exemplo 74
6. Metodologia para o estudo da divergência genética 78
6.1. Introdução 78
6.2. Passos da metodologia 78
6.3. Exemplo 78
6.4. Relação entre divergência genética e heterose 84
7. Programas de computador para o emprego das técnicas de análise multivariada nos estudos dos germoplasmas 88
7.1. Introdução 88
7.2. Descrição dos programas 88
7.2.1. PROGRAMA: Estandar.cm 88
7.2.2. PROGRAMA: Varcovar.cm 90
7.2.3. PROGRAMA: Corre.cm 91
7.2.4. Programa: Conver.cm 92
7.2.5. Programa: Cpp.cm 93
7.2.6. Programa: Matriz.cm 94
7.2.7. Programa: Dme2.cm 95
7.2.8. Programa: dme4.cm 96
7.2.9. Programa: Binário.cm 97
7.3. Descrição das rotinas 99
7.3.1. Exemplo de rotina para cálculo dos componentes principais 103
7.3.2. Exemplo de rotina para cálculo de distâncias 104
7.3.3. Exemplo de rotina para cálculo das coordenadas principais 105
7.3.4. Rotina para geração da matriz de similaridade, a partir de dados binários 106
8. Referências Bibliográficas 108

1. ABORDAGENS E METODOLOGIAS PARA AVALIAÇÃO DE GERMOPLASMA

1.1. INTRODUÇÃO

A garantia para o uso eficiente dos germoplasma passa, necessariamente, pelos trabalhos relacionados com a sua caracterização e avaliação. É por intermédio desses estudos que se aquilata o potencial dos germoplasma para emprego imediato ou futuro na renovação da variabilidade que está sendo trabalhada pela seleção. Estes primeiros trabalhos geram, também, a estrutura de dados para ulteriores estudos, os quais, ao lado dos anteriores, formam o quadro de referência acerca das potencialidades dos germoplasma do ponto de vista prático ou científico. Neste capítulo pretende-se dar uma idéia acerca das principais abordagens e metodologias usadas nos estudos dos germoplasmas, complementares ao trabalho da simples caracterização e avaliação.

1.2. PRINCIPAIS ABORDAGENS

Os dados obtidos nos estudos dos germoplasma podem ser visualizados segundo diferentes enfoques; todavia, três abordagens principais vêm sendo utilizadas pelos especialistas nestes estudos, que são: 1) Classificação dos acessos da coleção; 2) Análise da diversidade genética e 3) Estudo da divergência genética.

A classificação dos acessos da coleção pode ser feita utilizando-se os procedimentos da análise estatística multivariada ou da análise numérica, cujas metodologias principais são abordadas nos itens a seguir. Nestes estudos são adotados os caracteres morfo-fisiológicos e de produção ou os dados derivados das análises dos sistemas isoenzimáticos.

A importância da classificação é que, por seu intermédio, pode-se proceder ao agrupamento dos acessos com base em suas similaridades e com isto tornar mais fácil o manuseio da coleção para os trabalhos de avaliação ou do melhoramento.

No caso de cultivares, o objetivo com o emprego da classificação é verificar se o agrupamento reflete a similaridade baseada nas relações de pedigree do material em estudo e ou se ela proporciona informações que possam ser usadas nos programas de melhoramento.

A diversidade genética diz respeito ao grau em que o material genético difere em uma população. O material genético, no caso de uma planta, é aqui entendido como correspondente a todo o seu DNA, quer genómico ou citoplasmático.

A importância do estudo da diversidade genética repousa no fato de que ela é a ferramenta do melhoramento genético. Deste modo, o esclarecimento de sua estrutura é vital para uso eficiente da coleção de germoplasma pelos melhoristas.

De outro lado, as informações derivadas desses estudos podem prestar-se a outras finalidades, entre as quais destacamos: 1) Estudos básicos de evolução da espécie com que se está trabalhando; 2) Planejamento eficiente das expedições de coleta de germoplasma; 3) Seleção dos acèssos de coleção para os programas de melhoramento e 4) Fornecer as indicações para o estabelecimento das coleções-núcleo (core colection) em uma espécie de planta e seus parentes selvagens.

A divergência genética relaciona-se com o grau em que as populações se distanciam uma da outra quanto ao conjunto de caracteres que lhe são peculiares. A divergência genética pode, assim, ser avaliada em termos da distância entre as populações que estão sendo comparadas.

A importância dos estudos de divergência genética deve-se ao fato de que já é lugar comum, entre os melhoristas, a crença de que a superioridade dos híbridos é proporcional à distância genética entre os seus respectivos progenitores. Deste modo, nas espécies onde esta relação é verificada, os melhoristas podem contar com um critério rápido e fácil na escolha dos progenitores para os programas de hibridação.

Segundo Wilches (1973) a avaliação da divergência genética também é de grande importância no contexto da evolução das espécies, pois o seu conhecimento não só provê informações sobre os recursos genéticos disponíveis como, ainda, pode auxiliar na localização e intercâmbio de tais recursos.

1.3. METODOLOGIAS ADOTADAS

As metodologias adotadas para os estudos das coleções de germoplasma, segundo as abordagens citadas, baseiam-se em métodos estatísticos uni ou multivariados e trabalham com dados derivados, quer da análise dos caracteres morfo-fisiológicos e de produção ou com as freqüências dos alelos, quando são utilizados os resultados das análises dos sistemas isoenzimáticos.

A classificação dos acessos da coleção é realizada adotando-se os métodos de ordenação e da análise de agrupamento, ambos pertencentes ao domínio da análise estatística multivariada.

A ordenação é um procedimento muito útil para o propósito de sumariar a informação acerca das relações implicadas para a série completa de caracteres medidos. Por exemplo, a medida de 20 caracteres nos acessos de uma coleção de germoplasma exigiria um elevado número de eixos cartesianos para que se pudesse apreciar em um gráfico o valor numérico de um germoplasma particular desta coleção. Ora, a representação dos dados em um espaço multidimensional não é possível pelos meios convencionais. É por esta razão que são empregados os métodos de ordenação, com os quais se pode projetar as entidades, no caso os germoplasma, em um espaço de menos dimensões do que o original e que seja capaz de englobar todas as variáveis estudadas. Neste contexto, a ordenação constitui-se, em essência, num poderoso instrumento para projetar os dados em dois ou três eixos com vistas a facilitar a sua inspeção e representação.

Os principais métodos multivariados usados com as finalidades de ordenação, são: 1) Análise dos componentes principais e 2) Análise das coordenadas principais.

A análise de agrupamento é a técnica que permite condensar as relações multivariadas entre as unidades de observação taxonomica (OTU's) utilizando a representação gráfica dos dados em duas dimensões. Esta análise, na verdade, engloba um conjunto de métodos, entre os quais se destacam os aglomerativos e os divisivos. O objetivo dos primeiros é proceder a uma série de fusões dos **n** objetos em grupos, até que todos eles formem um só aglomerado. Estes são os de emprego mais generalizado nos estudos que estão sendo considerados. Nos divisivos, pelo contrário, a meta é a quebra do conjunto em subconjuntos, até que, com a partição sucessiva, cada subconjunto se apresente como uma simples OTU.

A análise da diversidade genética pode ser realizada a partir dos caracteres morfológicos de natureza qualitativa ou quantitativa ou dos dados derivados dos sistemas isoenzimáticos. No primeiro caso, o procedimento mais amplamente adotado é o Índice de Shannon e Weaver (1961) e no segundo a maioria dos autores adota a Análise da Diversidade de Nei (1949).

Para os estudos da divergência genética, adota-se a Análise dos Componentes Principais, Variáveis Canônicas e os Métodos de Agrupamento.

2. ANÁLISE DOS COMPONENTES PRINCIPAIS

2.1. INTRODUÇÃO

A análise dos componentes principais foi descrita, originalmente, por Karl Pearson, no início do século na área da geometria espacial. Mais tarde Hotelling (1933, 1936) ampliou o emprego da técnica a outros campos, o que tornou possível a sua utilização inclusive no melhoramento genético.

Basicamente, a análise dos componentes principais opera condensando a variância de um conjunto de dados em uns poucos eixos, de modo que se torna possível visualizar a maior parte da variabilidade dos dados originais em duas ou três novas dimensões (componentes). Uma das vantagens desta técnica é que, ao contrário das variáveis originais que estão correlacionadas, os componentes obtidos, por não gozarem desta condição, podem ser interpretados independentemente. Nas considerações a seguir serão desenvolvidos os aspectos principais desta técnica, principalmente no tocante aos conceitos e emprego da metodologia nos estudos dos dados das coleções de germoplasma.

2.2. DESENVOLVIMENTO DA METODOLOGIA

Na apresentação dos conceitos da análise dos componentes principais consideram-se as variáveis $X_1, X_2, X_3, ..., X_p$, com

$$\text{Var}(X_i) = \sigma_i^2,\ i = 1, 2, \ldots, p$$

e $$\text{COV}(X_i, X_j) = \rho_{ij}\,\sigma_i^2\,\sigma_j^2,\ i = 1, 2, \ldots, p$$

onde ρ_{ij} é o coeficiente de correlação entre X_i e X_j. Nestas condições, a matriz de variâncias e covariâncias referente às **p** variáveis será:

$$S = \begin{bmatrix} V(X_1) & COV(X_1,X_2) & \cdots & COV(X_1,X_p) \\ COV(X_1,X_2) & V(X_2) & \cdots & COV(X_2,X_p) \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ COV(X_1,X_p) & COV(X_2,X_p) & \cdots & V(X_p) \end{bmatrix}$$

que é simétrica e real.

A análise dos componentes principais parte de uma matriz deste tipo ou, então, da que envolve as correlações entre as variáveis com a estrutura a seguir:

$$\begin{bmatrix} 1 & r12 & r13 & \cdots & r1p \\ r21 & 1 & r23 & \cdots & r2p \\ r31 & r32 & 1 & \cdots & r3p \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ rp1 & rp2 & rp3 & \cdots & 1 \end{bmatrix}$$

que é também simétrica, isto é, são iguais os elementos acima e abaixo da diagonal ($r_{ij} = r_{ji}$). A diagonal principal é ocupada pela unidade porque a correlação de uma variável com ela mesma é igual a 1.

É comum quando as variáveis são medidas em unidades diferentes proceder a sua padronização, antes de se obter a matriz das variâncias e covariâncias. Isto é conseguido dividindo-se o valor de cada variável pelo respectivo desvio padrão ou subtraindo-se a média de cada observação e dividindo-se o resultado pelo desvio-padrão. Quando se trabalha com as variáveis padronizadas, a análise, que é feita usando-se a matriz de correlação, é semelhante à que é realizada com as variâncias e covariâncias.

A análise dos componentes principais consiste em transformar um conjunto P de variáveis X_1, X_2 ... X_p, em um novo Y_1, Y_2 ... Y_p, onde cada Y_i ou componente principal é uma combinação dos X'S, isto é:

$$Y_1 = C_{11} X_1 + C_{12} X_2 + \ldots + C_{1p} X_p$$
$$Y_2 = C_{21} X_1 + C_{22} X_2 + \ldots + C_{2p} X_p$$

$$\vdots$$

$$Y_p = C_{p1} X_1 + C_{p2} X_2 + \ldots + C_{pp} X_p$$

ou utilizando a notação matricial $Y = CX$

Os coeficientes das combinações lineares devem ser escolhidos para atender às seguintes condições:

1. variância de $Y_1 \geq$ variância $Y_2 \geq \ldots \geq$ variância Y_p;
2. os valores de quaisquer um dos componentes principais não são correlacionados;
3. para qualquer componente principal a soma dos quadrados dos coeficientes é igual à unidade (valores normalizados).

Em palavras, isto significa dizer que Y_1 é a combinação linear com mais alta variância, Y_2 a de maior grandeza depois de Y_1, Y_3 idem depois de Y_1 e Y_2, e assim por diante, com a condição, porém, de que nenhuma delas seja correlacionada entre si.

Na última expressão os elementos C da matriz são formados pelos autovetores associados aos autovalores da matriz S ou R. Normalmente, trabalha-se com os autovetores normalizados e, assim, o vetor Y passará a ser $Y = Kx$; então, os componentes principais serão:

$$Y_1 = K_{11} X_1 + K_{12} X_2 + \ldots + K_{1p} X_p$$
$$Y_2 = K_{21} X_1 + K_{22} X_2 + \ldots + K_{2p} X_p$$

$$\vdots$$

$$Y_p = K_{p1} X_1 + K_{p2} X_2 + \ldots + K_{pp} X_p$$

com as variâncias e covariâncias equivalentes a:

$$V(Y_1) = \lambda_1,\ V(Y_2) = \lambda_2 \ldots V(Y_p) = \lambda_p$$
$$COV(Y_1, Y_2) = COV(Y_1, Y_3) = \ldots COV(Y_{p-1}, Y_p) = 0$$

e

$$K_{p1}^2 + K_{p2}^2 + \ldots + K_{pp}^2 = 1$$

A obtenção das estimativas dos componentes principais é conseguida pela solução do sistema:

$$[R_{(n)} - \lambda_{I(n)}]_n X_1 = {}_n\varnothing_1 \quad (1)$$

Então, se $R_{(n)}$ é uma matriz quadrada de ordem m, real e simétrica, *a polinomial de* grau m em λ de det $|R_{(n)} - \lambda I(n)| = \varnothing$ (2) é definida como equação característica ou polinômio característico de $R_{(n)}$. Os escalares λ_i = 1, 2, ..., m, raízes de (2) são as raízes características ou autovalores de $R_{(n)}$ e as correspondentes soluções não nulas para X em (1) são os autovetores associados aos autovalores de $R_{(n)}$, onde R é a matriz de correlação entre as variáveis, I é a matriz identidade de dimensões p x p, λ_i as raízes características ou autovalores da matriz de correlação ou com as variâncias e covariâncias e a_i é o vetor característico ou autovetor. Esta quantidade representa o conjunto de transformações ortogonais através das quais as variáveis originais padronizadas devem ser multiplicadas para obtenção das variáveis transformadas.

A solução do sistema é tal que $a \neq \varnothing$ e, neste caso, o determinante da expressão $|R - \lambda I| = \varnothing$. Esta condição torna o sistema indeterminado e, assim, a solução passa a ser escolhida entre aquelas que satisfaçam a relação $a_i' . a_i = 1$.

Estimados os componentes principais, pode-se medir a sua importância relativa, o que é feito através da expressão:

$$\frac{V(Y_{ij})}{\text{Traço de R}} = \frac{\lambda_i}{\text{Traço de R}}$$

onde o traço de R é a soma dos elementos da diagonal principal da matriz de correlação ou com as variâncias e covariâncias, que é igual a:

$$\text{Traço de R} = V(Y_{i1}) + V(Y_{i2}) + \ldots V(Y_{ip})$$

Um exemplo não só ajudará a esclarecer essas noções como também, poderá servir para mostrar como são estimados os componentes principais a partir de um conjunto P de variáveis.

2.3. EXEMPLO

Considere os dados a seguir correspondentes a sete acessos de uma coleção de germoplasma de sisal (*Agave sisalana*, Perr.) avaliados no Campo Experimental de Monteiro, PB, com respeito aos caracteres número de folhas por planta, peso de 1 folha, peso da fibra seca por planta e rendimento de fibra.

TABELA 1. Médias e desvios padrões para os caracteres medidos em sete acessos de uma coleção de germoplasma de sisal avaliados no Campo Experimental de Monteiro, PB, dados de 1989-1991

	Médias para os caracteres			
Acessos	Nº de folhas por planta (g)	Peso de 1 folha (kg)	Peso da fibra seca p/planta (%)	Rendimento da fibra (%)
A_1	109,25	588,97	2,29	4,02
A_2	99,50	551,65	1,73	4,24
A_3	69,32	645,90	1,64	4,03
A_4	48,65	505,75	0,71	2,90
A_5	72,37	716,50	1,81	3,28
A_6	78,00	519,37	1,23	3,25
A_7	81,00	557,37	1,62	4,03
Média	79,72	583,64	1,58	3,67
D. Padrão	19,98	74,68	0,49	0,52

Os dados da Tabela correspondem às variáveis originais (X_i) e, como para obter os componentes principais eles precisam ser padronizados, tem-se então as variáveis Z_i depois de subtrair de cada X_i a média e dividir o resultado pelo respectivo desvio-padrão.

Acessos	Variáveis			
	Z_1	Z_2	Z_3	Z_4
A_1	1,477	0,071	1,446	0,656
A_2	0,989	-0,488	0,312	1,078
A_3	-0,521	0,834	0,130	0,675
A_4	-1,555	-1,043	-1,753	-1,495
A_5	-0,368	-1,779	0,474	-0,765
A_6	0,086	-0,861	-0,700	-0,823
A_7	0,064	-0,352	0,090	0,675

É a partir destas variáveis Z_i que é obtida a matriz de correlação e, como as variáveis são padronizadas, ela corresponde, também, à matriz de variâncias e covariâncias.

A matriz em apreço equivale a:

$$R = \begin{bmatrix} 1,00 & 0,04403 & 0,83256 & 0,73416 \\ & 1,00 & 0,553367 & 0,136002 \\ & & 1,00 & 0,710989 \\ & & & 1,00 \end{bmatrix}$$

Na prática, a análise dos componentes principais envolve o cálculo dos autovalores ou raízes características e autovetores da matriz de correlação ou com as variâncias e covariâncias. Para a matriz R em questão os autovalores são calculados pela expressão:

$|R - \lambda_i I| = \varnothing$ ou, de forma mais explicita, do seguinte modo:

$$\left[\begin{bmatrix} 1,0 & 0,04403 & 0,0832568 & 0,73416 \\ & 1,0 & 0,553367 & 0,136002 \\ & & 1,0 & 0,710989 \\ & & & 1,0 \end{bmatrix} - \lambda_i \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \right] = 0$$

que equivale a:

$$\begin{bmatrix} 1,0-\lambda & 0,04403 & 0,0832568 & 0,73416 \\ & 1,0-\lambda & 0,553367 & 0,136002 \\ & & 1,0-\lambda & 0,710989 \\ & & & 1,0-\lambda \end{bmatrix} = 0$$

Vê-se que a expressão obtida origina uma equação do quarto grau que deve ter, no caso geral, quatro raízes não negativas λ_1, λ_2, λ_3 e λ_4. O cálculo destas raízes é bastante trabalhoso quando realizado manualmente, de modo especial nas matrizes de grandes dimensões, porém, com o emprego de computadores, torna-se muito fácil e rápido a sua obtenção.

As raízes, no caso, fornecidas pelo computador, são:

$\lambda_1 = 2,634559$
$\lambda_2 = 1,048053$
$\lambda_3 = 0,298356$
$\lambda_4 = 0,019031$

Deve ser lembrado que os autovalores são as variâncias amostrais dos valores de Y_1, Y_2 ... Y_p das novas variáveis ou componentes principais. Os dados a seguir correspondem à variância de cada componente e sua importância relativa em relação à variância total, isto é, a soma dos valores de λ_i.

Componente	Variância	Variância (%)	Variância Acumulada
1	2,634459	65,75	65,75
2	1,048053	26,75	92,00
3	0,298356	0,45	99,45
4	0,019031	0,47	100,00
Total	3,999899		

Observa-se, desses dados, que os dois primeiros componentes principais totalizam 92% da variância total existente. Desta forma, esses dois componentes podem, perfeitamente, ser utilizados para representar o conjunto das variáveis medidas nos acessos da coleção, uma vez que eles incorporam mais de 90% daquela variância. Do ponto de vista geométrico, isto quer dizer que se pode trabalhar num espaço bidimensional representado por dois eixos cartesianos para a representação das novas variáveis ou

componentes principais.

Às vezes, pode ser necessário acrescentar um terceiro eixo ou componente principal quando, por exemplo, os dois primeiros autovalores não encerram a maior parte da variância da amostra. Neste caso, quando são usados três eixos (k = 3) como responsáveis pela maior parte da variação, então, um espaço tridimensional deve ser adotado para representar os dados. A norma usual, contudo, é considerar somente os autovalores que apresentem a maior parte da variação, isto é, um mínimo de 80% para identificação do número de eixos a empregar.

Obtidos os autovalores e definidos o número de eixos, o próximo passo consta do cálculo dos coeficientes (h_{ij}), pois nas equações para a obtenção dos componentes principais são conhecidos os valores das variáveis originais (X_i). Os coeficientes, do ponto de vista teórico, são os autovetores da matriz de correlação ou com as variâncias e covariâncias. Destas matrizes são obtidos tanto autovetores quantos sejam os autovalores e são os primeiros que, substituídos nas equações de Y_i, irão permitir a obtenção das novas variáveis ou componentes principais.

O cálculo dos autovetores é feito através da seguinte expressão matricial:

$$[R - \lambda_i I]\ [h_{ij}] = \varnothing$$

No caso do primeiro autovetor correspondente ao primeiro autovalor ou raiz característica $\lambda_1 = 2{,}634559$, a expressão anterior equivale a:

$$\begin{bmatrix} 1{,}0 - 2{,}634559 & 0{,}04403 & 0{,}832658 & 0{,}73416 \\ 0{,}04403 & 1{,}0 - 2{,}634559 & 0{,}553367 & 0{,}136002 \\ 0{,}832568 & 0{,}55367 & 1{,}0 - 2{,}6344559 & 0{,}710989 \\ 0{,}73416 & 0{,}136002 & 0{,}710989 & 1{,}0 - 2634559 \end{bmatrix} \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{14} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Procedendo as operações no sistema matricial obtem-se as seguintes equações homogêneas.

$$\begin{aligned} -1{,}634559\, h_{11} + 0{,}04403\ h_{12} + 0{,}832568\, h_{13} + 0{,}73416\ h_{14} &= 0 \\ 0{,}04403\ h_{11} - 1{,}634554\, h_{12} + 0{,}553367\, h_{13} + 0{,}136002\, h_{14} &= 0 \\ 0{,}832568\, h_{11} + 0{,}553363\, h_{12} - 1{,}634559\, h_{13} + 0{,}710989\, h_{14} &= 0 \\ 0{,}73416\ h_{11} + 0{,}13600\ h_{12} + 0\,710989\, h_{13} - 1{,}634559\, h_{14} &= 0 \end{aligned}$$

O sistema obtido é indeterminado e, assim, uma solução seria conseguida atribuindo-se, por exemplo, o valor 1 à variável h_{14} e considerando-se só as três primeiras equações, isto é:

$$\begin{aligned} 1{,}634559\, h_{11} + 0{,}04403\ h_{12} + 0{,}832568\, h_{13} &= 0{,}73416 \\ 0{,}04403\ h_{11} - 1{,}634559\, h_{12} + 0{,}553367\, h_{13} &= 0{,}136002 \\ 0{,}832568\, h_{11} + 0{,}553367\, h_{12} - 1{,}634559\, h_{13} &= 0{,}710989 \end{aligned}$$

O uso dos computadores permite obter o vetor solução do sistema já normalizado que, no exemplo considerado, equivale a:

$$h = \begin{bmatrix} h_{11} \\ h_{12} \\ h_{13} \\ h_{14} \end{bmatrix} = \begin{bmatrix} -0{,}54724 \\ -0{,}260472 \\ -0{,}596028 \\ -0{,}52672 \end{bmatrix}$$

A normalização é conseguida dividindo-se o vetor solução pela norma de h, isto é por:

$$||h|| = \sqrt{h' \times h}$$

ou para o exemplo em questão:

$$||h|| = \sqrt{h_{11}^2 + h_{12}^2 + h_{13}^2 + h_{14}^2}$$

A partir do vetor-solução obtém-se a expressão para o primeiro componente principal, que corresponde à seguinte equação:

$$Y_1 = -0{,}54724\, Z_1 - 0{,}260472\, Z_2 - 0{,}596028\, Z_3 - 0{,}52672\, Z_4$$

O processo para se chegar aos demais componentes segue o mesmo raciocínio, isto é, considerando-se a segunda, terceira e quarta raízes e resolvendo-se os sistemas correspondentes. Assim, repetindo o processo obtém-se os demais componentes principais cujas expressões são as seguintes:

$$Y_2 = -0,3463822\ Z_1 + 0,882204\ Z_2 + 0,170644\ Z_3 - 0,269543\ Z_4$$
$$Y_3 = 0,515642\ Z_1 - 0,097317\ Z_2 + 0,279624\ Z_3 - 0,804023\ Z_4$$
$$Y_4 = 0,56098\ Z_1 + 0,379999\ Z_2 - 0,733104\ Z_3 + 0,058818\ Z_4$$

Na Tabela 2 são encontrados os autovetores e autovalores dos dois primeiros componentes principais correspondentes a cada variável para o exemplo que está sendo apresentado.

TABELA 2. Autovalores e autovetores dos dois primeiros componentes principais correspondentes a cada variável estudada na coleção de germoplasma de Sisal avaliada no Campo Experimental de Monteiro-PB

Variáveis	Componente Principal	
	1°	2°
Número de folhas p/planta	-0,54724	-0,346388
Peso de 1 folha	0,260472	0,882204
Peso da Fibra Seca p/planta	-0,596028	0,170644
Rendimento	-0,52672	-0,269543
Autovalores	2,634559	1,048053
% Acumulada	65,75	92,00

Conforme já foi referido, os dois primeiros componentes principais concentram 92% da variância total e, assim, na interpretação dos resultados, somente eles devem ser considerados. (Tabela 2). Observa-se nos coeficientes do primeiro componente principal que os de maiores cargas (peso) são o número de folhas por planta, peso da fibra seca por planta e rendimento de fibra. Portanto, o primeiro componente principal está relacionado com estas variáveis, em especial com o peso da fibra seca por planta cujo coeficiente foi o de maior carga. (-0,596028). O exame dos coeficientes no segundo componente revela que o de maior valor foi o peso de uma folha; portanto, ele corresponde ao eixo maior da variação na variável considerada.

No emprego da técnica dos componentes principais, estas informações podem ser complementadas, ainda, com a estimação dos valores das correlações entre as variáveis estudadas e os componentes principais escolhidos.

A correlação entre a i-ésima componente principal e a j-ésima variável é calculada empregando-se a expressão:

$$r_{ij} = C_{ij}\sqrt{\lambda_i}$$

onde λ_i corresponde à variância do componente principal ou autovalor e C_{ij} o coeficiente correspondente à variável constante na Tabela 2.

No exemplo que estamos considerando, a correlação entre o número de folhas e o primeiro componente principal é calculada tomando-se a raiz de λ_1 (2,634559) que equivale ao primeiro autovalor ou raiz característica, multiplicada por (-0,54724) referente ao coeficiente da variável considerada que, consta na tabela já citada.

Portanto,

$$r_{11} = -0,54724 \times \sqrt{2,634559}$$
$$r_{11} = -0,88$$

O sinal negativo indica que, a medida em que a variável aumenta, o valor do componente principal diminui e, assim, a interpretação tem o mesmo significado que nos estudos convencionais de correlação.

Para a estimação das demais correlações, o processo é o mesmo porém, quando se tratar do segundo componente principal, o valor a ser trabalhado é $\lambda_2 = 1.048053$ correspondente ao segundo autovalor tomando-se o C_{ij} da variável cuja correlação se quer estimar. Por intermédio do conhecimento destas correlações pode-se reforçar a importância das variáveis na determinação dos componentes principais e, com isto, desprezar aquelas com baixos valores de **r**.

A etapa final no emprego da técnica dos componentes principais consta da dispersão dos dados num sistema de eixos coordenados. Neste caso, os valores correspondentes aos genótipos estudados são referidos em um gráfico cujas coordenadas são os componentes principais.

No exemplo considerado são escolhidos somente dois eixos, representando os primeiro e segundo componentes principais, pois as duas encerram, como já se viu, mais de 90% da variância total dos dados.

Os valores a seguir correspondem aos escores dos genótipos para os primeiros e segundo componentes a partir dos quais pode ser constituído o gráfico com a dispersão dos 7 genótipos estudados.

Genótipos	Escores	
	Y_1	Y_2
1	-0,38	-2,03
2	-0,96	-1,18
3	0,76	-0,36
4	-0,28	2,95
5	1,98	-0,14
6	-0,63	1,12
7	-0,50	-0,35

Na Fig. 1 encontra-se o gráfico com a dispersão dos genótipos tomando-se os dois eixos correspondentes aos primeiro e segundo componentes principais. Nota-se, a partir do exame do gráfico, que podem, visualmente, ser detectados cinco conglomerados, dos quais os I, III e IV acham-se constituídos por apenas um membro e os demais por dois cada um, conforme assinalado na Fig. 1.

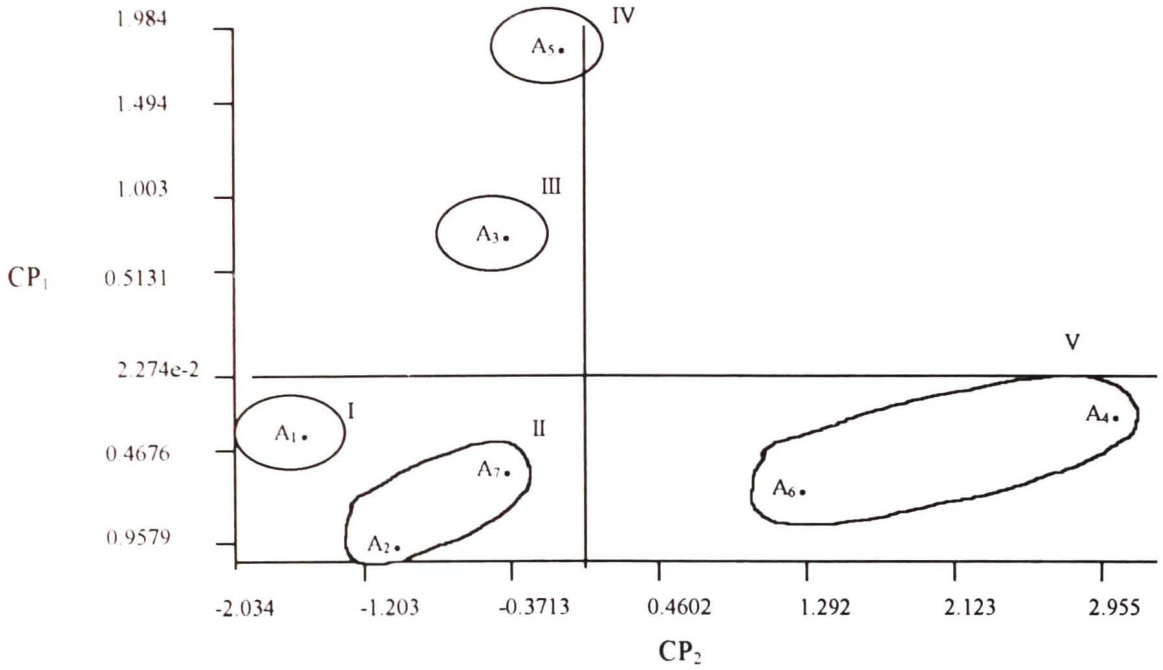


FIG. 1 - Distribuição dos pontos referentes aos acessos de sisal nos dois eixos correspondentes ao primeiro e segundo componentes principais

A_1 - Híbrido 11648; A_2 - Híbrido da Paraíba; A_3 - Mutante da Bahia; A_4 - Sisalana; A_5 - IAC 034; A_6 - IAC 069 e A_7 - Mutante da Paraíba

3. ANÁLISE DAS COORDENADAS PRINCIPAIS

3.1. INTRODUÇÃO

A análise das coordenadas principais foi desenvolvida por Grower (1936) e constitui-se num grande avanço nas técnicas de ordenação. Graças a esta metodologia, pode-se computar os componentes principais de qualquer matriz com as distâncias euclidianas, sem que para isto se necessite dos dados originais (Sneath e Sokal, 1973).

No caso de se trabalhar com distâncias baseadas nos caracteres padronizados, os resultados desta técnica são equivalentes aos encontrados com a análise dos componentes principais, empregada a partir da matriz de correlação. Nos itens a seguir serão abordados os fundamentos da metodologia e do seu emprego aos dados obtidos nos estudos com os germoplasma.

3.2. DESENVOLVIMENTO DA METODOLOGIA

A metodologia da análise das coordenadas principais pressupõe certos conhecimentos indispensáveis ao seu pleno entendimento, em que o principal deles é o relacionado com a medida de similaridade adotada para agrupar os genótipos.

A similaridade pode ser avaliada através de várias medidas e estas dependem do caráter que está sendo estudado. Para detalhes sobre as mesmas, consultar Sneath e Sokal (1973) e Dun & Everitt (1982) os quais tratam o assunto de forma bastante detalhada.

A medida comumente empregada para avaliar a similaridade é a distância Euclidiana (d_{ij}). Para definir esta distância, considere que são estudados **n** genótipos e medidos, em cada um deles, **p** características, de modo que se obtém um conjunto de dados X_{ij} disposto segundo a matriz:

$$X = \begin{bmatrix} X_{11} & X_{12} & X_{13} & \cdots & X_{1n} \\ X_{21} & X_{22} & X_{23} & \cdots & X_{2n} \\ & & & & \\ & & & & \\ X_{p1} & X_{p2} & X_{p3} & \cdots & X_{pn} \end{bmatrix}$$

cujo primeiro índice se refere ao genótipo e o segundo à característica medida, tal que i = 1,2...n e j = 1,2...p.

A distância Euclidiana para p caracteres medidos em duas Unidades de Observação Taxonômica (OTU's) i e i' é definida pela expressão:

$$dii' = \left[\sum_{j=1}^{n} (x_{ij} - X_{i'j})^2 \right]^{\frac{1}{2}}$$

Para tornar claro como se calcula esta distância, considere as 3 OTU's 1, 2 e 3, nas quais foram medidos o peso em gramas e a altura em centímetros, conforme os resultados a seguir:

OUT's	Peso (g)	Altura (cm)
1	25	1,5
2	20	1,0
3	15	0,5

A distância Euclidiana entre os OTU's 1 e 2 utilizando a expressão anterior é obtida calculando-se:

$$d_{12} = \left[(25-20)^2 + (1,5-1,0)^2 \right]^{\frac{-1}{2}} = 25,25^{\frac{-1}{2}} \text{ ou } d_{12} = 5,02$$

Um dos inconvenientes no emprego da distância Euclidiana é a alteração em seu valor com a mudança de escala de medição das variáveis estudadas. É por esta razão que no cálculo desta distância deve-se proceder à padronização dos dados, a exemplo do que foi feito com o estudo da análise dos componentes principais.

A padronização, no caso, é obtida calculando-se uma nova variável Z_{ij}, isto é:

$$Z_{ij} = \frac{X_{ij} - \overline{X}_{.j}}{S(X_j)}$$

onde X_{ij} é a variável original e $S(X_j)$ é o desvio padrão da característica j.

Outro ponto a considerar com respeito à distância Euclidiana é que ela aumenta com o número de caracteres e, assim, ao invés de medir d_{ij} pode ser mais conveniente computar a distância Euclidiana média que é obtida pela expressão:

$$d_{ij} = \sqrt{d_{ij} / p}$$

A distância Euclidiana, apesar de sua simplicidade, tem uma restrição, que é a de não se ajustar a dados nos quais os p caracteres não podem ser representados em p eixos ortogonais. Devido à correlação entre os caracteres, esta condição nem sempre é satisfeita e, nesta situação, outras medidas devem ser adotadas em substituição à Euclidiana; uma delas é a distância generalizada D^2 de Mahalanobis, a qual será tratada no item referente à análise de agrupamento.

Quando as distâncias são avaliadas para os p caracteres entre diversas unidades, obtém-se, com os valores calculados, uma matriz p x p da forma:

$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & \dots & d_{1n} \\ d_{21} & d_{22} & d_{23} & \dots & d_{2n} \\ & & & & \\ & & & & \\ d_{p1} & d_{p2} & d_{p3} & \dots & 0 \end{bmatrix}$$

chamada matriz de distância ou de dissimilaridade. É chamada dissimilaridade porque a diagonal é ocupada por zeros, pois a dissimilaridade de uma entidade

com ela mesma é nula. A citada matriz é simétrica, o que equivale a dizer que são iguais os elementos acima e abaixo da diagonal.

A análise das coordenadas principais parte, exatamente, de uma matriz de distância deste tipo. Para isto, a matriz tem de ser transformada através dos seguintes passos:

1. Gera-se uma nova matriz,

$$A_{pp} = -\frac{1}{2} d_{ij}, \text{ com } i, j = 1,2\cdots3\cdots p, \text{ isto é:isto é:}$$

$$App = \begin{bmatrix} 0 & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ . & & & & \\ . & & & & \\ . & & & & \\ a_{p1} & a_{p2} & a_{p3} & \cdots & 0 \end{bmatrix}$$

2. A partir da matriz A_{pp} constroe-se outra B_{pp}:

$$Bpp = \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & a_{22} & \cdots & b_{2n} \\ . & & & \\ . & & & \\ . & & & \\ b_{p1} & b_{p2} & \cdots & b_{pp} \end{bmatrix}$$

também simétrica, cujos elementos são formados a partir de A_{pp} da seguinte maneira:

$$b_{11} = a_{11} . \bar{a}_1 . \bar{a}_1 . \bar{a}..$$

.

.

.

$$b_{pp} = a_{pp} . \bar{a}_p . \bar{a}_p . \bar{a}..$$

onde $a_{p.}$ e $a_{.p}$ são, respectivamente, as médias das linhas e colunas, e a., a média geral.

Feito isto, são extraídos os autovalores e autovetores desta última matriz. A partir daí, a obtenção e interpretação das coordenadas principais são em tudo semelhantes à técnica dos componentes principais.

3.3. EXEMPLO

Para exemplificar como se trabalha com as coordenadas principais, considere os mesmos dados com os sete germoplasma discutidos no item anterior.

As distâncias Euclidianas médias determinadas para as combinações duas a duas entre os diversos genótipos permitem obter a matriz de dissimilaridade a seguir de dimensões, 7 x 7.

$$\begin{bmatrix} 0 & 0,698 & 1,255 & 2,514 & 1,523 & 1,589 & 1,002 \\ & 0 & 1,008 & 2,105 & 1,592 & 1,222 & 0,518 \\ & & 0 & 1,791 & 0,881 & 1,223 & 0,661 \\ & & & 0 & 1,927 & 0,968 & 1,673 \\ & & & & 0 & 1,451 & 1,317 \\ & & & & & 0 & 0,887 \\ & & & & & & 0 \end{bmatrix}$$

Aplicando-se nesta matriz as operações definidas anteriormente, chega-se à matriz A_{pp} e, finalmente a B_{pp} com os dados a seguir:

$$\begin{bmatrix} 1,177 & 0,647 & -0,019 & -1,471 & -0,058 & -0,435 & 0,160 \\ 0,647 & 0,605 & -0,026 & -0,812 & -0,450 & -0,205 & 0,242 \\ -0,019 & -0,026 & 0,359 & -0,324 & 0,304 & -0,329 & 0,035 \\ -1,471 & -0,812 & -0,324 & 2,203 & -0,242 & 0,872 & -0,224 \\ -0,058 & -0,450 & 0,304 & -0,242 & 1,027 & -0,300 & -0,280 \\ -0,435 & -0,205 & -0,329 & 0,872 & -0,300 & 0,478 & -0,080 \\ 0,160 & 0,242 & 0,035 & -0,224 & -0,280 & -0,080 & 0,147 \end{bmatrix}$$

Os dois primeiros autovalores desta matriz, obtidos por processo semelhante ao adotado no estudo dos componentes principais, totalizam mais

de 90% da variância total. Logo as duas primeiras coordenadas principais podem, perfeitamente, descrever o conjunto dos genótipos estudados, uma vez que elas envolvem a maior parte da variação total existente.

O primeiro autovalor $\lambda_1 = 3,9518$ permite estimar os autovetores ou coeficientes de ponderação da primeira coordenada principal cuja expressão é dada por:

$$Y_1 = 0,5116\ Z_1 + 0,2977\ Z_2 + 0,0918\ Z_3 - 0,7431\ Z_4 + 0,0355\ Z_5 - 0,2822\ Z_6 + 0,0885\ Z_7$$

O segundo $\lambda_2 = 1,5720$ possibilita obter a segunda coordenada principal ou:

$$Y_2 = 0,1510\ Z_1 + 0,3819\ Z_2 - 0,3015\ Z_3 + 0,1106\ Z_4 - 0,7912\ Z_5 + 0,2500\ Z_6 + 0,1989\ Z_7$$

Destas expressões obtém-se os escores das duas coordenadas principais e, a partir deles, pode-se estudar as dispersões dos genótipos nos eixos correspondentes às primeira e segunda coordenadas. A interpretação das posições dos genótipos nos dois eixos é feita de modo semelhante ao que já foi discutido no item anterior, com o estudo dos componentes principais.

4. ANÁLISE DE AGRUPAMENTO

4.1. INTRODUÇÃO

A análise de agrupamento constitui-se em uma técnica com a finalidade de evidenciar as relações multidimensionais entre um conjunto de observações em um gráfico denominado dendograma ou fenograma, por intermédio do estabelecimento de grupos entre elas.

A técnica, na verdade, envolve uma série de procedimentos distintos, tais como o estabelecimento da relação entre as observações (expressa em matrizes de distância, associação ou correlação) o agrupamento propriamento dito dessas observações e a avaliação do fenograma (Sokal, 1986). Nos itens a seguir serão descritos os aspectos essenciais da metodologia e mostrado como ela é empregada nos dados obtidos nos estudos dos germoplasma.

4.2. MEDIDAS DE SIMILARIDADE

Dois aspectos de particular importância têm de ser considerados na análise de agrupamento, que são a medida de similaridade adotada para formar os grupos e a escolha do método de agrupamento.

4.2.1. DISTÂNCIA DE MAHALANOBIS

A medida de similaridade mais amplamente adotada nestes estudos é a distância Euclidiana (d_{ij}) já definida no item referente ao estudo das coordenadas principais.

Uma outra distância, também muito difundida para os propósitos de agrupamento dos acessos de uma coleção de germoplasma, é a distância generalizada D^2 proposta por MAHALANOBIS (1936). Esta distância tem a vantagem de levar em consideração a correlação entre os caracteres analisados e, como tal, em certas situações deve substituir a Euclidiana, já mencionada.

A distância D^2 é calculada através da expressão:

$$D2 = (\overline{X} - \overline{Y})' \cdot W^{-1} (\overline{X} - \overline{Y})$$

onde X e Y são os vetores das médias entre dois grupos, isto é $X' = (\overline{X}_1 \overline{X}_2 \cdots \overline{X}_p)$ $Y' = (\overline{Y}_1 \overline{Y}_2 \cdots \overline{Y}_p)$ e W é a matriz polida de dimensões p x p das dispersões dentro dos dois grupos. Neste caso, $W = (W_1 + W_2)/ 2$ e W_1 e W_2 são matrizes, também, p x p da soma de quadrados e produtos para os dois grupos.

A expressão para D^2 pode tomar também a forma a seguir:

$$D_{ii} = b_1 d_1 + b_2 d_2 + \ldots + b_j d_j$$

na qual os d_i's correspondem a diferenças entre as médias de dois genótipos i e i' para uma característica j e os b's são coeficientes estimados para tornar a razão correspondente à variância entre genótipos pela variância dentro, maximizada.

Em termos de variância e covariância, a distância D_{ii}, é definida também pela expressão:

$$D_{ii'} = [d_1 \; d_2 \ldots d_j] S^{-1} \begin{bmatrix} d_1 \\ d_2 \\ \cdot \\ \cdot \\ \cdot \\ d_j \end{bmatrix}$$

na qual S é a matriz de dispersão amostral comum a todas as unidades.

O problema da estimação dos valores D^2_{ii}, está na obtenção desta matriz S ou da W, já referida. Quando os genótipos são testados em um delineamento experimental apropriado envolvendo a casualização e a repetição, é comum, no cálculo de D_{ii}', tomar-se como S a matriz de dispersão residual comum a todos os genótipos. Esta matriz é mais apropriada neste caso, pois os quadrados médios residuais têm distribuição normal multivariada (Goodman, 1968, citado em Cruz, 1987).

Para ilustrar o cálculo de D_{ii} com o uso desta matriz vão ser considerados os dados de 8 cultivares de sisal (*Agave sisalana* Perr) testados em um experimento em blocos ao acaso com 6 repetições. Os genótipos foram

avaliados para o caracteres peso da folha, comprimento da fibra, peso de fibra úmida e peso de fibra seca.

O resumo das análises de variância envolvendo cada característica isoladamente é encontrado na tabela a seguir:

TABELA 3. Resumo das análises de variância para os caracteres peso da folha, comprimento da fibra, peso da fibra úmida e peso da folha seca em cultivares de sisal testados no Campo Experimental de Monteiro, PB

F.V.	G.L.	Q.M			
		Peso da folha (PF)	Comprimento da fibra (CF)	Peso da fibra úmida (P.F.U.)	Peso da folha seca (P.F.S.)
Genótipos	7	19.950,45	608,48	161,50	48,53
Resíduo	35	1.066,90	36,99	22,92	7,03

As análises da covariância para cada par de características constam da Tabela 4 e, para os cálculos, adotou-se o mesmo modelo empregado nas análises da variância.

Para obtenção dos produtos médios (PM) com respeito a um par de combinações de variáveis, considera-se a variável Z resultante da soma de duas características quaisquer, X e Y, isto é:

$$Z = X + Y$$

Desta expressão obtém-se a variância de Z, que equivale a:

$$V(Z) = V(X) + V(Y) + 2\,Cov\,(X, Y)$$

e, daí, a exemplo do caso dos quadrados médios, estima-se o produto médio de X, Y, isto é:

$$PM(X,Y) = \frac{1}{2}\left[Q.M.(X+Y) - QM(X) - QM(Y)\right]$$

Por intermédio desse artifício foram, então, obtidos os produtos médios a seguir:

TABELA 4. Resumo das análises da covariância para os pares de características avaliados em cultivares de sisal testados no Campo Experimental de Monteiro, PB

F.V.	G.L.	P.M					
		P.F x C.F.	P.F x P.F.U.	P.F x P.F.S.	C.F x P.F.U.	C.F x P.F.S.	P.F.S x P.F.U.
Genótipos	7	3.070,24	1.609,15	898,91	276,69	162,66	87,93
Resíduo	35	169,17	128,41	73,83	25,38	13,37	12,34

P.F. = Peso da Folha; P.F.U. = Peso da Fibra úmida; P.F.S. = Peso da Fibra Seca.; C.F = Comprimento da Fibra

Das análises procedidas pode-se, então, obter a matriz S ou de dispersão residual com os elementos:

$$S = \begin{bmatrix} 1.066,90 & 169,07 & 128,41 & 73,83 \\ & 36,99 & 26,38 & 13,77 \\ & & 22,92 & 12,34 \\ & & & 7,03 \end{bmatrix}$$

cuja inversa é:

$$S^{-1} = \begin{bmatrix} 0,00462 & -0,013361 & 0,017172 & -0,052612 \\ & 0,151483 & -0,158031 & 0,120891 \\ & & 0,962089 & -1,559588 \\ & & & 3,194542 \end{bmatrix}$$

Na Tabela 5 são encontradas as médias das características dos 8 germoplasmas de sisal, a partir dos quais são calculadas as diferenças entre os genótipos (d_i) para a obtenção dos valores de D_{ii}.

TABELA 5. Médias das características avaliadas nos germoplasmas de sisal testados no Campo Experimental de Monteiro, PB, no ano de 1991

Genótipos	Média das características avaliadas			
	Peso da folha	Comprimento da fibra	Peso da fibra úmida	Peso da fibra seca
1	102,37	60,00	15,96	7,87
2	224,75	75,58	21,33	11,46
3	130,29	51,58	14,42	7,25
4	258,29	79,37	26,96	14,58
5	160,79	59,12	14,50	7,46
6	97,40	56,58	10,04	5,71
7	171,79	62,58	15,17	8,04
8	123,62	71,83	14,58	7,84

Para estimar a distância D_{ii}', entre os genótipos 1 e 2 procede-se, primeiramente, ao cálculo das diferenças entre as linhas 1 e 2, como segue:

$$\begin{aligned} 102,37 - 222,47 &= -122,38 \\ 60,00 - 75,58 &= -15,58 \\ 15,96 - 21,33 &= -5,37 \\ 7,87 - 11,46 &= -3,59 \end{aligned}$$

posto isto, substituem-se os valores de d_i e S na expressão de D_{ii}, isto é:

$$D^2_{12} = \begin{bmatrix} -122,38 & -15,38 & -5,37 & -3,59 \end{bmatrix} \begin{bmatrix} 0,00462 & 0,013351 & 0,017172 & -0,052512 \\ & 0,151483 & -0,158031 & 0,120891 \\ & & 0,962089 & -1,559588 \\ & & & 3,194542 \end{bmatrix} \begin{bmatrix} -122,38 \\ -15,58 \\ -5,37 \\ -3,59 \end{bmatrix}$$

Realizando-se as operações matriciais, obtém-se o valor de D^2_{12}. Raciocínio semelhante permite calcular os valores das demais distâncias

O cálculo da distância generalizada D^2 de Mahalanobis pode ser muito simplificado quando as características em estudo são independentes ou não correlacionadas. Neste caso e quando são tomadas as variáveis padronizadas, esta distância é equivalente à Euclidiana (d_{ij}).

$$D_{ii'} = \sum_{j=1}^{j} (Y_{ij} - Y_{i'j})^2$$

onde as variáveis Y são médias das variáveis não correlacionadas e padronizadas, isto é;

$$y_j = \frac{Y_j}{s(Y_j)}$$

Deste modo, são importantes para o cálculo de D^2 os métodos que permitem transformar um conjunto de variáveis x_j em outras não correlacionadas Y_j.

Dois métodos principais podem ser usados com este objetivo, que são a condensação pivotal e o emprego dos escores dos primeiros componentes principais ou das variáveis canônicas.

O método da condensação pivotal é muito trabalhoso pois consiste num processo de triangulação da matriz de dispersão S, através das operações com as linhas e colunas. No final, o processo fornece as funções de transformações por intermédio das quais são obtidas as novas variáveis Y_j não correlacionadas. O processo fornece, também, as variâncias com as quais é possível padronizar as novas variáveis obtidas. Para detalhes sobre esta metodologia consultar Cruz (1987) o qual traz um exemplo de como opera a condensação pivotal na matriz S para o cálculo das Distâncias D^2 .

Desta forma, o processo apresentado será o uso dos escores dos componentes principais para transformar um conjunto de variáveis em outro não correlacionado. Conforme já foi salientado, os componentes principais são estimados de modo tal que as correlações entre eles sejam nulas. Assim sendo, adotando-se para os componentes os escores correspondentes aos dois primeiros autovalores com maior percentual acumulado da variância total, consegue-se não só a independência das novas variáveis como, ainda, a garantia de que elas incorporem a maior parte da variação total dos dados.

Para exemplificar de como se opera com este método serão tomadas as variâncias padronizadas (Z) do exemplo considerado no ítem 2.3, relativo ao estudo dos componentes principais. Conforme já foi relatado, a matriz de correlação derivada destas variáveis apresentou os seguintes autovalores:

$\lambda_1 = 2{,}634559$
$\lambda_2 = 1{,}048053$
$\lambda_3 = 0{,}298356$
$\lambda_4 = 0{,}019031$

e autovetores constando na matriz a seguir:

$$\begin{bmatrix} 0,56098 & 0,515642 & -0,346388 & -0,54724 \\ 0,37999 & -0,097317 & 0,882204 & -0,260472 \\ -0,733104 & 0,279624 & 0,170644 & -0,496028 \\ 0,058818 & -0,8040023 & -0,269543 & -0,52672 \end{bmatrix}$$

As duas últimas colunas desta matriz são os autovetores correspondentes aos dois primeiros autovalores (λ_1 e λ_2) com maior participação acumulada na variância total (90%). Logo, são os escores que deverão ser utilizados para a obtenção do conjunto de variáveis não correlacionadas. Para isto, multiplica-se a matriz padronizada (Z), pela matriz de autovetores, definida anteriormente. A partir dessa operação gera-se a matriz dos escores:

$$\begin{bmatrix} -0,1655 & 0,6322 & -0,3787 & -2,0343 \\ 0,2267 & -0,2274 & -0,9579 & -1,1839 \\ -0,0311 & -0,8558 & 0,7560 & -1,1839 \\ -0,0719 & 0,0113 & -0,2775 & 2,9548 \\ 0,0766 & 0,3849 & 1,9840 & -0,1414 \\ 0,0891 & 0,5050 & -0,6269 & 1,1220 \\ -0,1239 & -0,4503 & -0,4990 & -0,3520 \end{bmatrix}$$

É dessa matriz que são calculadas as distâncias generalizadas D^2 de Mahalanobis e como as variáveis não são correlacionadas elas são estimadas do mesmo modo como no caso da distância Euclidiana.

Assim, para a distância entre os genótipos 1 e 2 obtém-se:

$$D^2_{12} = (-0,1655 - 0,2267)^2 + (0,6322 + 0,2274)^2 + (-0,3785 + 0,9579)^2 + (-2,0343 + 1,1839)^2 = 9,3512$$

Procedendo-se de modo semelhante, são estimadas as demais distâncias envolvendo os outros pares de genótipos.

Às vezes, os dados disponíveis são provenientes dos estudos com as isoenzimas e a partir deles deseja-se proceder ao agrupamento dos genótipos da coleção. Neste caso, pode-se trabalhar com as bandas tomando-as como descritor fenotípico que vai servir para caracterizar os acessos da coleção, ou com as freqüências das bandas isoenzimáticas nas populações ou acessos cuja similaridade se deseja estimar.

Na primeira situação, cada banda pode ser considerada uma variável qualitativa, atribuindo-se-lhe o valor 1 quando ela está presente, e zero quando ausente. Assim sendo, quando dois genótipos diferem, no mínimo, em uma banda (eletromorfo) eles serão, então, considerados como exibindo diferentes zimotipos. Existem dois coeficientes que podem ser de muita utilidade para a análise de dados binários, como os citados anteriormente, e que servem para medir a similaridade entre dois OTU's. Estas medidas são o coeficiente simples de emparelhamento (simple matching coeficient) e o coeficiente de Jaccard:

4.2.2. COEFICIENTE SIMPLES DE EMPARELHAMENTO

Para definir este coeficiente, considere o quadro a seguir:

		OTUi		
		1	0	
OTUj	1	a	b	a + b
	0	c	d	c + d
				P = a + b + c + d

no qual dois caracteres estão representados por presença (1) e ausência (0) nos OTU's i e j, onde:

P = número de caracteres binários estudados
a = número de caracteres com ambos os OTU's codificados por 1
b = número de caracteres com o OTU j codificado por 1 e o i por zero
c = número de caracteres com o OTU j codificado por 0 e o i por 1
d = número de caracteres com ambos os OTU's codificados como 0
O coeficiente simples de emparelhamento s_{ij} é definido pela expressão:

$$s_{ij} = \frac{a + d}{P}$$

Portanto, é medido pela razão entre o número total de parelhas 1-1 e 0-0 (a + d) e o número de caracteres (P).

Este coeficiente varia entre zero, quando os dois OTU's deixam de formar parelhas sobre qualquer uma das P variáveis e 1 quando, pelo contrário, formam parelhas sobre cada caráter.

Para exemplificar como é feito o cálculo deste coeficiente, considere que no estudo de 10 germoplasma para uma isoenzima foram obtidos 9 padrões de bandas distribuídos entre os acessos da maneira a seguir:

Germoplasma	PADRÕES								
	1	2	3	4	5	6	7	8	9
1	1	0	0	0	0	0	0	0	0
2	1	1	0	0	0	1	0	0	0
3	1	1	0	0	0	0	0	0	0
4	1	0	0	1	0	0	0	0	0
5	1	0	0	1	1	0	0	0	0
6	1	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	0	0
8	1	1	1	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0	0
10	1	0	0	0	0	0	0	0	0

Vê-se, desta Tabela, que os OTU's 1 e 2 diferem quanto aos padrões 2 e 6, pois estes estão ausentes em 1, porém presentes no 2.

O coeficiente simples de emparelhamento para os germoplasma em apreço é calculado observando-se que para o par 1-1 (a) existe, apenas, uma combinação, 1-0 (b) não ocorre nenhuma, 0-1 (c) existem 2 e, finalmente, para o par 0-0 são encontradas 6 combinações.

Logo, substituindo-se estes valores na fórmula para s_{ij} tem-se:

$$s_{12} = \frac{a + d}{P} = \frac{1 + 6}{9}$$

$$s_{12} = 0,77 \text{ ou } 77\%$$

pois $p = a + b + c + d = 9$

4.2.3. COEFICIENTE DE JACCARD

Este coeficiente é definido pela expressão:

$$s_{ij} = \frac{a}{a + b + c}$$

isto é, corresponde à razão entre o número de parelhas 1-1 (positivas) para o número total de caracteres sem considerar o número de pares 0-0, referente ao valor de d.

No caso dos germoplasma 1 e 2 do exemplo considerado, o coeficiente de Jaccard é calculado organizando-se a Tabela a seguir:

		Germoplasma 2	
		1	0
Germoplasma 1	1	a = 1	b = 0
	0	c = 2	d = 6

Como o valor de d não é considerado, o coeficiente em questão para os dados apresentados será:

$$s_{12} = \frac{a}{a + b + c} \text{ ou}$$

$$s_{12} = \frac{1}{1 + 0 + 2}$$

$$s_{12} = 0{,}33 \text{ ou } 33\%$$

Como se pode notar, este valor difere de 0,77 calculado para o primeiro coeificiente envolvendo os mesmos germoplasma. Dunn & Everitt (1982) discutem as razões desta discrepância e apontam que ela se deve ao fato dos valores de similaridade não formarem uma série monotônica, isto é, que aumenta ou diminui ao longo de toda a sua amplitude. Para detalhes consultar

esta fonte, onde podem ser encontradas maiores referências sobre os coeficientes apresentados.

Em outras situações, a genética das isoenzimas é conhecida de modo que é possível identificar os loci e os respectivos alelos envolvidos em sua manifestação. Quando isto acontece pode-se determinar as freqüências dos alelos envolvidos com cada sistema enzimático sendo estas, então, usadas para o cálculo das distâncias entre as populações da coleção de germoplasma.

Quando se trabalha com as freqüências dos alelos, pode--se usar a distância Euclidiana usual, conforme definida no item 3.2, ou a distância genética de Rogers (1972).

4.2.4. DISTÂNCIA DE ROGERS

A distância de Rogers (1972) é definida através da expressão:

$$D_{ii'j} = \frac{1}{L}\sum_{j=1}^{L}\left(\frac{1}{2}\sum_{k=1}^{n}\left(P_{kji} - P_{kji'}\right)^2\right)^{\frac{1}{2}}$$

onde

P_{kji} e $P_{kji'}$ são as frequências do alelo k no loco j nas populações i i', respectivamente

L = número de loci examinados
n = número de alelos no loco j

Para exemplificar como são calculadas as distâncias Euclidianas e de Rogers (1972) considere as freqüências constantes das Tabelas 6 e 7, a seguir, correspondentes a 6 loci com três alelos cada em duas populações 1 e 2.

As distâncias Euclidianas entre as populaçoes 1 e 2 em relação a cada loco, a partir das freqüências alélicas dadas nas tabelas citadas, são:

TABELA 6. Freqüências alélicas correspondentes a cada loco na população 1

Loco	Freqüência Alélica		
	P(a)	P(b)	P(c)
1	0,600	0,200	0,100
2	0,500	0,500	0,100
3	0,700	0,200	0,100
4	0,300	0,350	0,350
5	1,000	0,000	0,000
6	0,600	0,300	0,100

TABELA 7. Freqüências alélicas correspondentes a cada loco na população 2

Loco	Freqüência Alélica		
	P(a)	P(b)	P(c)
1	0,100	0,800	0,100
2	0,600	0,250	0,150
3	0,200	0,600	0,200
4	1,000	0,000	0,000
5	0,900	0,100	0,000
6	0,200	0,300	0,500

$$^{\text{Loco }1}D_{12,1} = \sqrt{(0,600-0,100)^2 + (0,200-0,800)^2 + (0,100-0,100)^2}$$

$$= 0,707$$

$$^{\text{Loco }2}D_{12,2} = \sqrt{(0,500-0,600)^2 + (0,500-0,250)^2 + (0,100-0,150)^2}$$

$$= 0,308$$

$$^{\text{Loco }3}D_{12,3} = \sqrt{(0,700-0,200)^2 + (0,200-0,600)^2 + (0,100-0,200)^2}$$

$$= 0,648$$

$$^{\text{Loco }4}D_{12,4} = \sqrt{(0,300-1,000)^2 + (0,350-0,000)^2 + (0,350-0,000)^2}$$

$$= 0,857$$

$$^{\text{Loco }5}D_{12,5} = \sqrt{(1,000-0,900)^2 + (0,000-0,100)^2 + (0,000-0,000)^2}$$

$$= 0,141$$

$$^{\text{Loco }6}D_{12,6} = \sqrt{(0,600-0,200)^2 + (0,300-0,300)^2 + (0,100-0,500)^2}$$

$$= 0,566$$

A distância de Rogers (1972) depois destes cálculos equivale a:

$$D_{12} = \frac{1}{6}\left(\sqrt{(0,707 + 0,308 + 0,648 + 0,857 + 0,141 + 0,566) \ /2}\right)$$

$$D_{12} = \frac{1}{6}\left(\sqrt{1,6135}\right)$$

$$D_{12} = \frac{1}{6}(1,2702)$$

$$D_{12} = 0,212$$

Os valores destas medidas obtidas, quer pelos coeficientes apresentados ou através das distâncias Euclidiana ou de Rogers (1972), permitem formar a matriz de similaridade a partir da qual se pode proceder ao agrupamento dos germoplasma, conforme será visto a seguir.

4.3. MÉTODOS DE AGRUPAMENTO

Os mais usados com os propósitos de classificar acessos nas coleções de germoplasma são os métodos hierárquicos e os de otimização.

Nos primeiros, a classificação dos acessos é realizada por um processo que se repete, até que se forme uma estrutura ramificada ou "árvore". Estes métodos classificam-se em aglomerativos e divisivos e destes serão tratados, apenas, os aglomerativos.

Nos segundos, os grupos são formados pela otimização de algum critério de agrupamento.

4.3.1. MÉTODOS AGLOMERATIVOS

São métodos cujo ponto de partida é a matriz de similaridade baseada na distância Euclidiana e sobre a qual é adotado algum critério de

agrupamento; no final, obtém-se o dendograma ou diagrama de "árvore" que permite avaliar as relações entre os OTU'S.

Entre os métodos aglomerativos destacam-se como os de mais amplo uso o método do Vizinho mais Próximo, Vizinho mais Distante, o Centróide e o método de Ward.

4.3.1.1. MÉTODO DO VIZINHO MAIS PRÓXIMO

Para exemplificar de como opera um método aglomerativo, será considerado o do vizinho mais próximo (Single linkage) que é de uso bastante freqüente no estudo que está sendo considerado.

Considere a matriz de dissimilaridade a seguir, segundo Freire e Moreira (1991) referente ao estudo de 8 acessos em uma coleção com diferentes raças e espécies de algodoeiros.

$$D_1 = \begin{array}{c} \text{OTU} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \end{array} \\ \left[\begin{array}{cccccccc} 0,0 & & & & & & & \\ 6,8 & 0,0 & & & & & & \\ 5,6 & 6,3 & 0,0 & & & & & \\ 5,0 & 6,2 & 2,8 & 0,0 & & & & \\ 4,4 & 6,9 & 3,2 & 2,7 & 0,0 & & & \\ 3,8 & 6,7 & 3,7 & 3,0 & 1,5 & 0,0 & & \\ 4,2 & 6,7 & 3,9 & 2,3 & 2,3 & 2,9 & 0,0 & \\ 44,7 & 6,5 & 4,4 & 3,8 & 2,3 & 2,6 & 3,4 & 0,0 \end{array}\right] \end{array}$$

Neste método a distância entre um grupo (ij) em um indivíduo (K) é obtida pela expressão:

$$d_{(ij)k} = \text{mim } \{d_{ik} ; d_{jk}\}$$

isto é, equivale ao menor elemento do conjunto das distâncias dos pares que se formam combinando os elementos i e j do grupo com k correspondente ao indivíduo.

A distância entre grupos obedece ao mesmo princípio e é dada por:

$$d_{(ij)(ke)} = \text{mim } \{d_{ik} ; d_{ie} ; d_{jk} ; d_{je}\}$$

Neste caso, a distância corresponde ao menor elemento do conjunto formado pelas distâncias entre os pares de indivíduos

(i e k), (i e **e**), (j e k) e (j e **e**).

Para operar com o método começa-se identificando o menor valor da matriz D_1, que no caso em questão equivale a 1,5 correspondente à distância entre as Unidades de Observação Taxonômicas (OTU"s) 5 e 6. Logo, os dois são fundidos para formar um aglomerado com dois membros, isto é, 5 e 6. É calculada, em seguida, a dissimilaridade entre este grupo e os outros 6 OTU"s restantes da maneira a seguir:

$d_{(56)1}$ = mim $\{d_{15}$, $d_{16}\}$ = mim {4,4 ; 3,8} = 3,8 menor distância)

$d_{(56)2}$ = mim $\{d_{25}$, $d_{26}\}$ = mim {6,9 , 6,7} = 6,7

e assim por diante, até $d_{(56)8}$ que é calculada como:

$d_{(56)8}$ = mim $\{d_{58}$, $d_{68}\}$ = mim {2,3 , 2,6} = 2,3

Com estas quantidades obtém-se uma nova matriz cujos elementos são as dissimilaridades entre os OTU's e do aglomerado formado com os demais, isto é:

$D_2 =$

OTU	1	2	3	4	5	(56)	7	8
1	0,0							
2	6,8	0,0						
3	5,6	6,3	0,0					
4	5,0	6,2	2,8	0,0				
5	4,4	6,9	3,2	2,7	0,0			
(56)	3,8	6,7	3,7	3,0	1,5	0,0		
7	4,2	6,7	3,9	2,3	2,3	2,9	0,0	
8	4,7	6,5	4,4	3,8	2,3	2,6	3,4	0,0

A matriz (D_2) conserva as dissimilaridades entre os OTU's da anterior (D_1) e no caso do aglomerado (56) encerra as distâncias dele com os OTU's restantes 1, 2, 3, 4 e 8. Como na etapa anterior, identifica-se, novamente, o menor valor na matriz D_2, que no caso considerado equivalente a 2,3

correspondente à distância entre os OTU's 4 e 7. Fundem-se os dois OTU's e calculam-se as dissimilaridades entre este novo grupo com os outros OTU's restantes e o primeiro aglomerado (56) da maneira seguinte:

$d_{(47)1} = \text{mim}\ \{d_{14}, d_{17}\} = \text{mim}\ \{5{,}0, 4{,}2\} = 4{,}2$ (menor valor)
$d_{(47)(2)} = \text{mim}\ \{d_{24}, d_{27}\} = \text{mim}\ \{6{,}2, 6{,}7\} = 6{,}2$

•
•
•

$d_{(47)(56)} = \text{mim}\ \{d_{45}, d_{46}, d_{57}, d_{67}\} = \text{mim}\ \{2{,}7, 3{,}0, 2{,}3, 2{,}9\} = 2{,}3$
$d_{(47)8} = \text{mim}\ \{d_{48}, d_{78}\} = \text{mim}\ \{3{,}8, 3{,}4\} = 3{,}4$

Forma-se, com estas quantidades, uma nova matriz D_3 com os elementos a seguir:

$$D_3 = \begin{array}{c} \text{OTU} \\ 1 \\ 2 \\ 3 \\ (47) \\ (56) \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & (47) & (56) & 8 \end{array} \\ \begin{bmatrix} 0{,}0 & & & & & \\ 6{,}8 & 0{,}0 & & & & \\ 5{,}6 & 6{,}3 & 0{,}0 & & & \\ 3{,}8 & 6{,}2 & 2{,}8 & 0{,}0 & & \\ 4{,}2 & 6{,}7 & 3{,}2 & 3{,}4 & 0{,}0 & \\ 4{,}7 & 6{,}5 & 4{,}4 & 2{,}3 & 2{,}3 & 0{,}0 \end{bmatrix} \end{array}$$

Esta matriz, a exemplo da anterior, conserva as dissimilaridades entre os OTU's existentes em D_2 e para os aglomerados (5 e 6) e (4 e 7) encerra não só as distâncias entre eles, como deles, em relação aos demais 1, 2, 3 e 8.

O processo se repete seguindo os mesmos passos já apresentados, até que no final todos os grupos aparecem fundidos e formando um só aglomerado.

A continuidade do processo permitirá chegar à matriz (D_5) com os elementos:

$$D_4 = \begin{array}{c} \text{OTU} \\ 1 \\ 2 \\ 3 \\ (45678) \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & (45678) \end{array} \\ \begin{bmatrix} 0{,}0 & & & \\ 6{,}8 & 0{,}0 & & \\ 5{,}6 & 6{,}3 & 0{,}0 & \\ 3{,}8 & 6{,}2 & 3{,}2 & 0{,}0 \end{bmatrix} \end{array}$$

Nesta matriz o menor valor corresponde à dissimilaridade entre o OTU 3 e o grupo (45678). Logo, funde 3 com este OTU e calculam-se as distâncias correspondentes:

$$d_{(345678)1} = \text{mim } \{d_{13}, d_{14}, d_{15}, d_{16}, d_{17}, d_{18}\} = 3,8$$

$$d_{(345678)2} = \text{mim } \{d_{23}, d_{24}, d_{25}, d_{26}, d_{27}, d_{28}\} = 6,2$$

A matriz D_6 final será:

$$D_5 = \begin{array}{c} \text{OTU} \\ 1 \\ 2 \\ (345678) \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 2 & (345678) \end{array} \\ \begin{bmatrix} 0,0 & & \\ 6,8 & 0,0 & \\ 3,8 & 6,2 & 0,0 \end{bmatrix} \end{array}$$

Observa-se, dessa matriz, que no final estarão fundidos os OTU's 1 e 2 com o aglomerado (345678) dado que eles são os restantes e os que não foram fundidos nas etapas anteriores.

O processo pode ser resumido conforme o quadro a seguir:

Passo	Junção	Nível
1	5 e 6	1,5
2	4 e 7	2,3
3	(56) e 8	2,3
4	(568) e (47)	2,3
5	(56847) e 3	3,0
6	(568473) e 1	3,8
7	(568472) e 2	6,2

Daí, obtém-se o dendograma ou diagrama de árvore, isto é, a representação dos resultados em um gráfico, que é muito útil e bastante usada na análise de agrupamento. Na construção deste gráfico utiliza-se de uma escala vertical à esquerda para indicar o nível de similaridade. No eixo horizontal são marcados os acessos, numa ordem conveniente, de modo que as linhas verticais que partem dos acessos têm altura correspondente ao nível em que eles são considerados semelhantes.

A Figura 2 contém o dendograma mostrando as relações entre as amostras de germoplasmas estudados para o exemplo em discussão.

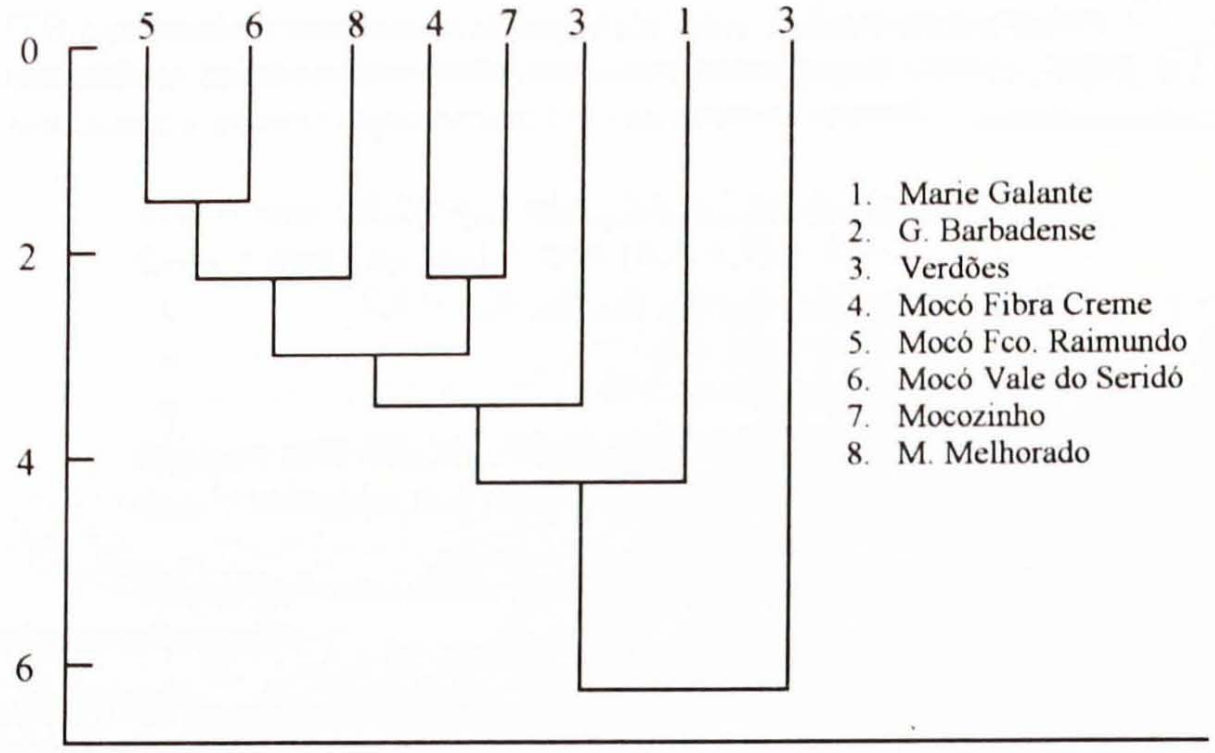


FIG. 2. Dendograma mostrando as relações entre as amostras de germoplasma estudados

4.3.1.2. MÉTODO DO VIZINHO MAIS DISTANTE

O método do vizinho mais distante (Complete Linkage Method) apresenta sistemática semelhante ao método anterior, porém difere deste com respeito às distâncias que são medidas para formar os grupos.

No caso da distância entre um grupo e um indivíduo é usado a expressão:

$$D_{(ij)k} = \max \{d_{ik} ; d_{jk}\}$$

na qual a distância, ao invés de menor, é a maior.

A distância entre dois grupos é dada por:

$$d_{(ij)(ke)} = \max \{d_{ik} ; d_{ie} ; d_{jk} ; d_{je}\}$$

isto é, corresponde ao maior elemento do conjunto formado pelas distâncias entre os pares de indivíduos (i e k) , (i e e), e (j e e).

Para detalhes de como se trabalhar com os outros métodos aglomerativos, consultar, entre outros, Sneath e Sokal (1973), Dun & Everitt (1982) e Gama (1980).

4.3.2. MÉTODOS DE OTIMIZAÇÃO

Nestes métodos a formação dos grupos se dá pela otimização de algum critério de agrupamento. Deste modo, a partição do grupo de indivíduos é realizada por meio da maximização ou minimização de alguma medida pré-fixada.

Diferem dos métodos hierárquicos porque os grupos formados são mutuamente exclusivos. A norma neste método de agrupamento é que a média dos valores D^2 intra-conglomerados deve ser menor do que a dos valores de D^2 inter-conglomerados.

4.3.2.1. MÉTODO DE TOCHER

Para ilustrar como opera este método considere a matriz com as distâncias D^2 de Mahalanobis (1936) envolvendo 8 cultivares de algodoeiro avaliados para os caracteres: rendimento (kg/ha), percentagem de fibra, peso médio de 1 capulho (g), peso de 100 sementes (g), comprimento da fibra (mm), uniformidade, finura e resistência.

$D_1 =$

OTU	1	2	3	4	5	6	7	8
1	-							
2	12,84	-						
3	17,40	8,59	-					
4	25,22	15,94	20,83	-				
5	20,99	11,16	5,02	13,09	-			
6	15,03	10,67	8,65	17,45	7,24	-		
7	20,28	12,67	12,25	23,90	8,09	16,60	-	
8	14,41	10,50	26,47	8,55	24,01	23,68	26,70	-

A primeira etapa no desenvolvimento do método é a determinação do acréscimo máximo permitido para inclusão de um genótipo no grupo. Isto é feito inspecionando-se a matriz D_1 para identificação dos genótipos com menores valores de D^2. No exemplo em discussão, os valores mínimos de D^2 para cada cultivar em relação a outra qualquer são os seguintes:

CULTIVAR	D^2	CULTIVAR MAIS PRÓXIMA
1	12,84	2
2	8,59	3
3	5,02	8
4	8,55	8
5	7,24	6
6	16,60	7
7	12,25	3
8	10,50	2

O maior valor da tabela é 16,60; logo, será este o limite estabelecido para a inclusão de um elemento num aglomerado. Portanto, uma cultivar só será incluída no aglomerado se o acréscimo médio no valor de D^2 não ultrapassar este valor limite (16,60).

O segundo passo é a formação do primeiro aglomerado. Para isto inspeciona-se, novamente, a matriz D_1, para a identificação da menor distância e esta, no exemplo em questão, equivale a $D^2 = 5,02$, envolvendo as cultivares 3 e 5. Como esta distância se situa abaixo do limite pré-fixado (16,600), então estas cultivares são agrupadas para a formação do primeiro aglomerado.

Em seguida, são calculadas as distâncias do aglomerado formado com as demais cultivares, adotando-se as seguintes expressões:

$$D_{(35)1} = D_{13} + D_{15} = 17,40 + 20,99 = 38,39$$
$$D_{(35)2} = D_{23} + D_{25} = 8,59 + 11,16 = 19,75$$
$$D_{(35)6} = D_{36} + D_{56} = 8,65 + 7,24 = 15,89$$
$$D_{(35)7} = D_{37} + D_{37} = 12,85 + 18,09 = 30,34$$
$$D_{(35)8} = D_{38} + D_{58} = 26,47 + 24,01 = 50,48$$

Com estas distâncias obtém-se uma nova matriz D_2 com os elementos a seguir:

$$D_2 = \begin{array}{c} \text{OTU} \\ 1 \\ 2 \\ 3 \\ 4 \\ 6 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccccccc} 1 & 2 & 3 & 4 & 6 & 7 & 8 \end{array} \\ \begin{bmatrix} - & & & & & & \\ 38,39 & - & & & & & \\ 19,75 & 12,84 & - & & & & \\ 25,85 & 25,22 & 15,94 & - & & & \\ 15,89 & 15,03 & 10,67 & 17,45 & - & & \\ 30,34 & 20,28 & 12,42 & 23,98 & 16,60 & - & \\ 50,48 & 14,41 & 10,50 & 8,55 & 23,68 & 26,70 & - \end{bmatrix} \end{array}$$

A cultivar (6) nesta matriz é a mais próxima do primeiro aglomerado (35) pois a distância 15,89 é a menor entre as demais envolvendo este grupo. Resta saber, no entanto, se a citada cultivar pode ser incluída neste grupo e, para isto, calcula-se o acréscimo médio no valor de D^2 caso a mesma fosse colocada no aglomerado em questão.

Este valor é calculado dividindo-se 15,89 por **n**, isto é, pelo número de cultivares no grupo original no caso **n**= 2. Como este coeficiente equivale a 7,94 e é inferior ao limite permitido (16,60) então a cultivar 6 pode ser incluída no mesmo grupo formado pela 3 e 5.

Calculam-se em seguida as distâncias do novo grupo com os demais, a exemplo do que foi feito na etapa anterior:

$$D_{(356)1} = D_{(35)1} + D_{16} = 38,39 + 15,03 = 53,42$$
$$D_{(356)2} = D_{(35)2} + D_{26} = 19,75 + 10,67 = 30,42$$
$$D_{(356)7} = D_{(35)7} + D_{(67)} = 15,89 + 16,60 = 32,49$$
$$D_{(356)8} = D_{(35)8} + D_{(68)} = 50,48 + 23,68 = 74,16$$

Obtem-se uma nova matriz D_3 com estas distâncias equivalentes a:

$$D_3 = \begin{array}{c} \text{OTU} \\ (356) \\ 1 \\ 2 \\ 4 \\ 7 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccccccc} (356) & 1 & 2 & 4 & 7 & 8 \end{array} \\ \begin{bmatrix} - & & & & & \\ 53,42 & - & & & & \\ 30,42 & 12,84 & - & & & \\ 43,30 & 25,22 & 15,94 & - & & \\ 32,49 & 20,28 & 12,42 & 23,98 & - & \\ 74,16 & 14,41 & 10,50 & 8,55 & 26,70 & - \end{bmatrix} \end{array}$$

Verifica-se, da nova matriz, que a cultivar mais próxima ao grupo (35 e 6) é a 2, pois a distância 30,42 é a menor entre as demais envolvendo o grupo citado.

O acréscimo médio da inclusão de 2 é calculado dividindo-se 30,42 por 3, isto é, o número de cultivares no grupo anterior. O resultado obtido (10,14) está abaixo do limite permitido (16,60) e, assim, a cultivar 2 pode ser incluída no grupo já formado por (3,5 e 6).

Calculam-se, novamente, as distâncias e repete-se o processo segundo as mesmas normas estabelecidas nas etapas anteriores. O primeiro aglomerado só é dado por encerrado quando o acréscimo médio da inclusão de determinada cultivar ultrapassar o valor de 16,60 pré-fixado.

No exemplo considerado, isto ocorreu quando se obteve a matriz D_5 com os elementos seguintes:

$$D_4 = \begin{array}{c} \text{OTU} \\ (3567) \\ 1 \\ 4 \\ 8 \end{array} \begin{array}{c} \begin{array}{cccc} (3567) & 1 & 4 & 8 \end{array} \\ \begin{bmatrix} - & & & \\ 86{,}54 & - & & \\ 83{,}22 & 83{,}22 & - & \\ 111{,}36 & 25{,}22 & 8{,}55 & - \end{bmatrix} \end{array}$$

Nesta matriz a cultivar mais próxima do grupo (356,7) é a 4; todavia, o acréscimo médio proporcionado pela inclusão de 4 é 83,22/5 = 16,64. Este número é ligeiramente maior do que o limite estabelecido de 16,60. Portanto, a partir daí começa-se a formar o segundo aglomerado.

Nesta etapa forma-se a matriz D_6, cujos elementos são os restantes após a exclusão do primeiro aglomerado, isto é:

$$D_5 = \begin{array}{c} \text{OTU} \\ 1 \\ 4 \\ 8 \end{array} \begin{array}{c} \begin{array}{ccc} 1 & 4 & 8 \end{array} \\ \begin{bmatrix} - & & \\ 25{,}22 & - & \\ 14{,}81 & 8{,}55 & - \end{bmatrix} \end{array}$$

Os genótipos mais próximos são o 4 e 8 com distância de 8,55 e, assim, o primeiro grupo consta destas cultivares. Calcula-se a distância de (48) com 1 que é a cultivar restante e que não entrou em nenhum grupo. Esta distância é calculada pela expressão:

$$D_{(48)1} = D_{14} + D_{18} = 25{,}22 + 14{,}41 = 39{,}63$$

Daí, obtém-se a matriz D_7 a seguir:

$$D_6 = \begin{array}{c} \text{OTU} \\ (48) \\ 1 \end{array} \begin{array}{c} \begin{array}{cc} (48) & 1 \end{array} \\ \begin{bmatrix} - & \\ 39{,}63 & - \end{bmatrix} \end{array}$$

O acréscimo médio de 1 no grupo formado por (4 e 8) é 39,63 dividido por 2. Este valor equivale a 19,81, e, assim, é bem maior do que o limite estabelecido e portanto, não permitindo a sua inclusão no grupo citado.

Deste modo, no final do processo são formados os grupos I constituido pelas cultivares (2, 3, 5, 6 e 7), II pela (4 e 8) e o III representado, apenas, pela cultivar 1.

A classificação dos acessos das coleções de germoplasma vem sendo realizada por vários autores nas mais diversas espécies agrícolas de inportância econômica. Na Tabela 8 acham-se relacionados os autores, técnicas utilizadas, fontes e tipos de plantas, onde, estes estudos já foram realizados.

TABELA 8. Relação dos autores, técnicas utilizadas, fontes para consulta de trabalhos dirigidos para a classificação dos acessos das Coleções de Germoplasma de diferentes plantas

Autor(es)	Técnica(s) Utilizada(s)	Fonte(s)	Planta(s)
Edye *et al.* (1970)	Análise de conglomerados	Aust. J. Agric. Res. v.21, p.57-69	Cevada
Martin & Rhodes (1977)	Método de Agrupamento - K. médias	Trop. Agri., v.17, p.257-263	Inhame
Hussaini *et al* (1977)	Componentes principais e análise da variável canônica	Crop Sci, v.17, p.257-263	Milheto
Martin & Rhodes (1978)	Componentes principais e análise de conglomerados	Trop. Agri., v.55, p.193-206	Inhame
Martin & Rhodes (1979)	Componentes principais e análise de conglomerados	Euphytica, v,28, p.367-383	Beringela
Bartual *et al* (1985)	Análise fatorial, componentes principais e análise de conglomerados	Euphytica, v.34, p.113-123	Soja
Veronesi e Falcinelli (1988)	Componentes principais e análise de conglomerados	Euphytica, v.38, p.211-220	Festuca
Zeven & Shache (1989)	Componentes principais e análise de conglomerados	Euphytica, v.41, p.235-246	Trigo
Caradus *et al* (1989)	Componentes principais e análise de conglomerados	Euphytica, v.42, p.183-196	Trevo branco
Hamon & Touré (1990)	Análise fatorial e de conglomerados	Euphytica, v.47, p.179-187	Inhame
Rezai e Frey (1990)	Análise de conglomerados, componentes principais e análise discriminante canônica	Euphytica, v.49, p.111-119	Aveia
Smith *et al* (1991)	Análise de conglomerados e componentes principais	Crop Sci. v.31, p.1.159-163	Alfafa
Zeven & Van Hintum (1992)	Análise de conglomerados e componentes principais	Euphytica, v.59, p.33-47	Trigo

5. METODOLOGIAS PARA O ESTUDO DA DIVERSIDADE GENÉTICA

5.1. INTRODUÇÃO

Os estudos sobre a diversidade genética nas coleções de germoplasma podem ser realizados a partir dos caracteres morfológicos de natureza qualitativa ou quantitativa ou com os dados derivados da análise dos sistemas isoenzimáticos.

No primeiro caso, o procedimento mais em voga é o índice de Shannon e Weaver (1962). Este índice, designado por H' foi, primeiramente, empregado por Lewontin (1973) na análise da variabilidade das populações humanas e só depois é que passou a ser empregado nos estudos com plantas.

Quando os dados utilizados derivam dos sistemas isoenzimáticos o procedimento mais utilizado é o da análise da diversidade genética de Nei (1973). Neste caso, trabalha-se com as freqüências dos alelos das isoenzimas e a metodologia é desenvolvida de tal forma que permite o cálculo da diversidade genética total (Ht) e a correspondente entre (D_{st}) e dentro (Hs) da população de germoplasma estudada. Nas considerações que se seguem serão apresentados os aspectos principais dessas duas metodologias e exemplos de como elas funcionam com os dados das coleções de germoplasma.

5.2. ÍNDICE DE SHANNON E WEAVER

No emprego do índice os diversos caracteres medidos nos acessos que fazem parte da coleção são, dependendo de sua variabilidade, subdivididos em diferentes classes fenotípicas.

Posto isto, são determinadas as freqüências fenotípicas dos diversos caracteres dentro de cada classe dos acessos estudados. É a partir dessas freqüências que o índice H' é calculado usando-se, para isto, a expressão:

$$H' = -\left[\sum_{i=1}^{n} P_i \log_2 P_i\right]$$

onde n = número de classes em que se divide o caráter e P_i é a proporção fenotípica de i-ésima classe para o caráter avaliado.

Deve ser observado que o índice baseia-se no logarítimo tomado na base 2, porém alguns autores preferem trabalhar mesmo com os logarítimos neperianos. Segundo Bogyo et al (1980) é irrelevante a base em que o logaritimo é calculado e a desvantagem de usar a base 2 é que as calculadoras eletrônicas não operam com esta base. Portanto, para facilitar a operacionalização, converte-se para a base 10.

Neste caso, a expressão do correspondente A H' será:

$$H' = -3{,}3219\left[\sum_{i=1}^{n} P_i \log_{10} . P_i\right]$$

5.2.1. EXEMPLO

Negassa (1985) fez uso deste índice no estudo do padrão da diversidade fenotípica em uma coleção de germoplasma contendo 485 raças locais de cevada da Etiópia, coletadas em várias províncias e, a partir de critérios geográficos, foram agrupadas em diversas regiões. O estudo baseou-se em 07 caracteres medidos em 10 espigas tomadas ao acaso, de cada parcela.

Na Tabela 9 são encontradas as percentagens fenotípicas das diversas classes para cada caráter medido nos acessos de cada província e nas regiões respectivas. Para regiões, os valores da Tabela citada correspondem às percentagens médias ponderadas pelo número de entradas estudadas em cada província. Os valores relativos à Etiópia, como um todo, correspondem à população total (485) também ponderados pelo número de acessos em cada região. O exame desta Tabela permite tirar conclusões acerca da diversidade fenotípica da população no tocante aos caracteres estudados. Por exemplo, vê-se que o caráter espiga lisa ocorre em todas as províncias e que a sua freqüência pelas regiões decresce das províncias do norte para as situadas na zona este. O caráter em questão é importante para a cevada nas regiões, onde a colheita não é mecanizada. Deste modo, conhecendo-se como varia este caráter pode-se, então, orientar a coleta de germoplasma que encerre esta característica para uso nos trabalhos de melhoramento.

TABELA 9. Percentagem das classes fenotípicas nos acessos de cada província e percentagem ponderada média de cada região para os sete caracteres estudados em cevada

Região	Nº Obs	Nº de Linhas		Comprimento da espiga		Tipo da Arista		Densidade da espiga				Cariopse		Cor do Grão			
		2	6	Curto	Longo	Rugosa	Lisa	Densa	Laxo	Curta	Longa	Coberta	Nua	Branca	Preta	Púrp.	Azul
Terras Altas do Nordeste																	
Begemidir	17	12	88	59	41	88	12	6	94	88	12	100	0	41	24	6	29
Tigre	66	58	42	45	55	92	8	6	94	61	39	100	0	50	11	11	28
Região	83	49	51	48	52	91	9	6	94	67	33	100	0	48	14	10	28
Platô Central																	
Gojjan	21	57	43	14	86	52	48	0	100	33	67	100	0	29	0	52	19
Shoa	171	23	77	49	51	91	9	8	92	76	24	99	1	41	21	19	19
Região	192	27	73	45	55	87	13	7	93	71	29	99	1	40	19	22	19
Terras Altas de Arussi-Rale																	
Arussi	108	20	80	51	39	94	6	13	87	70	30	100	0	44	17	17	22
Bale	29	14	86	55	45	90	10	7	93	66	34	100	0	45	17	4	34
Região	137	19	81	60	40	93	7	12	88	69	31	100	0	44	17	14	25
Platô Sudeste																	
Gemugofa	35	26	74	43	57	86	14	6	94	60	40	86	14	60	29	11	0
Sidamo	23	30	70	35	65	96	4	0	100	52	48	100	0	57	9	30	4
Wollega	15	47	52	27	73	33	67	0	100	40	60	100	0	27	13	13	47
Região	73	32	68	37	63	78	22	3	97	53	47	93	7	52	19	17	12
Etiópia	485	29	71	49	51	88	12	8	92	67	33	99	1	44	18	17	21

FONTE: MEGASSA, MULUGETA. 1985

A Tabela 10 contém os valores de H' calculados a partir das percentagens constantes na tabela anterior.

Para exemplificar como são calculados estes valores, considere o caráter número de linhas na espiga para a província Begemidir, o qual se acha subdividido em duas classes fenotípicas, isto é, 2 e 6, respectivamente, nas freqüências 0,12 e 0,88. (Tabela 9).

O cálculo de H', aplicando-se o logaritimo neperiano, é feito do seguinte modo:

$$H' = -\sum_{i=1}^{n} P_i \ln P_i \quad \text{ou} \quad H' = (-0{,}12 \times \ln 0{,}12) + (-0{,}88 \times \ln 0{,}88)$$

Como ln 0,12 = -2,120 e ln 0,88 = -0,128, tem-se substituindo na expressão anterior:

$$H' = [-0{,}12 \times (-2{,}120)] + [-0{,}88 \times (-0{,}128)] \text{ ou}$$
$$H' = 0{,}2544 + 0{,}113 = 0{,}367$$

Este valor tem de ser normalizado, isto é, deve ser dividido pelo logaritmo neperiano do número de classes (2) que é 0,693. Logo, o valor de H' normalizado será:

$$H' = \frac{0{,}367}{0{,}693} = 0{,}52$$

como se encontra na Tabela 10 citada.

TABELA 10. Estimativa dos índices de diversidade fenotípica (H') para as várias províncias, regiões e caracteres, e média da diversidade (H') e seu erro padrão para todos os caracteres em cevada

	Nº de Linha	Comprimento da espiga	Tipo da Arista	Densidade da espiga	Peso na Ráquila	Cariopse	Cor do Grão	$H' \pm SE$
Terras Altas do Nordeste								
Begemidir	0,52	0,98	0,52	0,33	0,52	0,00	0,89	$0,54 \pm 0,11$
Tigre	0,98	1,00	0,40	0,33	0,97	0,00	0,86	$0,65 \pm 0,04$
Região	1,00	1,00	0,45	0,33	0,92	0,00	0,88	$0,65 \pm 0,04$
Plantô Central								
Gojjan	0,98	0,59	1,00	0,00	0,92	0,00	0,74	$0,60 \pm 0,07$
Shoa	0,79	1,00	0,45	0,40	0,79	0,09	0,96	$0,64 \pm 0,03$
Região	0,84	1,00	0,56	0,38	0,86	0,09	0,97	$0,67 \pm 0,03$
Terras Altas de Arussi-Bale								
Arussi	0,72	0,97	0,33	0,56	0,88	0,00	0,93	$0,63 \pm 0,08$
Bale	0,59	1,00	0,46	0,38	0,92	0,00	0,82	$0,60 \pm 0,04$
Região	0,71	0,98	0,38	0,52	0,89	0,00	0,93	$0,63 \pm 0,04$
Plantô Sudeste								
Gemugofa	0,82	0,98	0,59	0,33	0,98	0,59	0,66	$0,71 \pm 0,08$
Sidamo	0,88	0,94	0,25	0,00	1,00	0,00	0,74	$0,54 \pm 0,06$
Wollega	1,00	0,84	0,92	0,00	0,98	0,00	0,89	$0,66 \pm 0,05$
Região	0,89	0,95	0,75	0,20	1,00	0,38	0,87	$0,72 \pm 0,05$
Ethiopia	0,87	1,00	0,52	0,40	0,92	0,09	0,94	$0,68 \pm 0,02$

FONTE: NEGASSA, MULUGETA. 1985

A normalização pode ser feita também diretamente, usando-se a expressão $H' = - 3,3219 [0,12 \log 0,12 + 0,88 \log 0,88]$, isto é, com o logarítmo transformado na base 10. O resultado é 0,52 é, assim, igual ao obtido com a expressão anterior.

O mesmo procedimento pode ser adotado para obter os demais valores de H' da Tabela em questão.

O índice H', no caso de se trabalhar com amostras grandes, tem distribuição que se aproxima da normal com variância equivalente a:

$$V_{H'} = \frac{\sum_{i=1}^{n} P_i \log_2^2 P_i - \left(\sum_{i=1}^{n} P_i \log_2 P_i\right)^2}{N} + \frac{n-1}{2N^2}$$

onde n e P_i têm os mesmos significados referidos anteriormente e N é o tamanho da amostra. Tomando o logarítmo na base 10, $V_{H'}$ é expresso por

$$V_{H'} = \frac{11,035\left[\sum_{j=1}^{n} P_j \log_{10}^2 P_j - \left(\sum_{j=1}^{n} P_j \log_{10} P_j\right)^2\right]}{N} + \frac{n-1}{2N^2}$$

Portanto, pode-se comparar os valores de H' de dois grupos usando-se para isto o teste "t", isto é:

$$t = \frac{H_1' - H_2'}{(V_{H_1'} + V_{H_2'})^{1/2}}$$

onde H_1' e H_2' são os valores do índice para os dois grupos ou caracteres e $V_{H_1'}$ e $V_{H_2'}$ respectivamente, as variâncias para os primeiro e segundo grupos ou caráter. O número de graus de liberdade no caso é obtido pela expressão de Satterthwaite (1946):

$$M = \frac{\left(V_{H1'} + V_{H2'}\right)^2}{\frac{VH_1^{2'}}{N1} + \frac{VH_2^{'2}}{N2}}$$

onde N_1 e N_2 são os graus de liberdade associados a cada uma das amostras.

Para exemplificar de como este teste é feito, considere os valores de H' para a província Begemidir, que é de 0,54 e o da Gemugota 0,72, respectivamente, o menor e o maior dos índices (H') constantes na Tabela 10.

Como as variâncias são 0,0121 e 0,0064, respectivamente para os valores de H' nas primeira e segunda províncias, então o valor t será:

$$t = \frac{0,71 - 0,54}{(0,0121 + 0,0064)^{1/2}}$$

$$t = \frac{0,17}{0,14} = 1,2$$

e, portanto, não significativo ($P > 0,20$).

Deste modo, da comparação realizada pode-se dizer, então, que em termos da diversidade fenotípica média estimada por H' as províncias não diferem significativamente. Este resultado pode significar que, em termos da amostragem da variabilidade nas raças locais de cevada estudadas, estas duas províncias podem ser consideradas um único sítio/ecogeográfico.

Os valores de H' têm a propriedade de aditividade e em razão desta pode-se combinar caracteres nas coleções de germoplasma contendo acessos de diferentes países. Assim sendo, a análise da variância, segundo o modelo hierárquico, pode ser adotada para testar a significância dos vários componentes da variação determinados pelos valores do índice H' como, por exemplo, regiões, países dentro de regiões e caracteres dentro de países. (Pielou, 1969).

O tipo mais simples de análise da variância é aquela realizada para cada caráter individualmente e que envolve as fontes de variação entre e dentro regiões. Esta análise é procedida com os valores de H' normalizados e no estudo de Negassa (1985) foi observado que mais de 70% da variância das estimativas de diversidade foi devida às diferenças entre os caracteres estudados.

No entanto, a análise pode envolver outras fontes de variação, como regiões, províncias dentro de regiões e caracteres dentro de províncias, conforme consta da análise da variância na Tabela 11.

TABELA 11. Análise da Variância segundo o modelo hierárquico para os valores dos índices (H')

C.V.	G.L.	Q.M.	F
Regiões	3	0,0012	< 1
Províncias Dentro de Regiões	5	0,0157	< 1
Caracteres Dentro de Províncias	54	0,1469	
Total			

FONTE: MEGASSA, MULUGETA 1985

Verifica-se, da Tabela 11, que muito da variação foi devida às diferenças no nível de diversidade dos diferentes caracteres, pois o quadrado médio correspondente a esta fonte foi 0,1469

5.3. ANÁLISE DA DIVERSIDADE GENÉTICA DE NEI

A análise da diversidade genética de Nei (1973) é adotada a partir dos dados obtidos no estudo dos sistemas isoenzimáticos. Portanto, o material de trabalho consta das freqüências dos genes envolvidos na determinação das isoenzimas. Diversos sistemas isoenzimáticos estão sendo utilizados com o propósito de caracterizar as coleções de germoplasma nas diferentes espécies de plantas. Para cada espécie já existe um padrão isoenzimático particular, freqüentemente de herança conhecida a partir do qual é realizado o trabalho de caracterização.

5.3.1. EXEMPLO

Para mostrar como esta análise é realizada suponha que existam **n** alelos para um locus e que a freqüência do i-ésimo alelo é x_i na população. Segundo Nei (1973) a probabilidade de identidade e não identidade de dois genes escolhidos ao acaso nesta população é $J = \sum x^2_i$ e $H = 1 - J$, respectivamente.

A probabilidade de não identidade (H) é uma medida da variação genética na população e, usualmente, é chamada de heterogozidade. Nei (1973) considera o termo inapropriado para uma população que não é cruzada ao acaso e, assim, prefere chamar esta quantidade de diversidade gênica.

Para exemplificar como são calculados estes valores, considere um locus qualquer com os alelos 2^S, 2^F e 2^a dados a seguir:

ALELOS	x_i	x_i^2
2^S	0,00	0,000
2^F	0,09	0,008
2^a	0,91	0,828
3		0,836

$\sum_{i=1} x^2_i$

A identidade e não identidade desses alelos são definidas como:

$$J = \sum_{i=1}^{n} x_i^2 = 0{,}836$$

$$H = 1 - \sum_{i=1}^{n} x_i^2 = 0{,}164$$

Em alguns autores a não identidade gênica (H) é simbolizada por **h** ao invés do **H** maiúsculo, como representada por Nei (1973). O termo **H** é, neste caso, reservado para a medida da diversidade gênica média correspondente a todos os loci em uma população.

Considerando **m** loci obtém-se **H** ou a diversidade gênica média através da expressão:

$$H = \sum_{l=1}^{m} h_l / m$$

onde **m** = número de loci e l refere-se ao l-ésimo locus.

A contribuição de Nei (1973) residiu em provar que no caso de subpopulações o valor da diversidade gênica total (H_t) pode ser fracionado nas componentes H_S e D_{st} correspondentes, respectivamente, à diversidade gênica dentro e entre populações, isto é:

$$H_t = H_S + D_{st}$$

Com estas quantidades pode-se calcular G_{st}, que é a diversidade gênica entre populações relativa a H_t, a qual é definida como:

$$G_{st} = \frac{(H_t - H_s)}{H_t}$$

Está fora dos objetivos deste trabalho apresentar a dedução dessas equações e, assim, o leitor interessado deve consultar o trabalho de Nei (1973) citado na bibliografia.

Para exemplificar como opera a análise da diversidade genética de Nei (1973) vão ser considerados os dados de Nevo *et al* (1986) baseados no estudo de 11 populações de cevada coletadas em diversos locais da Turquia. Os autores trabalharam com 17 sistemas isoenzimáticos condicionados por um mínimo de 2 e um máximo de 6, totalizando 30 alelos nas populações estudadas. Para maiores detalhes sobre os critérios adotados na identificação das isoenzimas, técnicas de eletroforese adotadas, propriedades dos loci e base genética envolvida na sua determinação, consultar Brown *et al*, 1978 b.

Na Tabela 12, adaptada de Nevo *et al* (1986), são encontradas as freqüências alélicas para os 30 loci nas 11 populações estudadas. A tabela contém, ainda, os valores de **h** calculados para cada locus e população, segundo a fórmula já apresentada.

$$H = 1 - \sum_{i=1}^{n} x_i^2$$

Desta tabela pode-se calcular o valor de H da diversidade gênica média para cada população. No caso da população 1 (Gaziantep) o valor de H é obtido como:

$$H = \frac{0{,}00 + 0{,}235 + \ldots + 0{,}213}{30} = 0{,}1067$$

TABELA 12. Freqüência alélica dos 30 locis nas onze populações de cevada na Turquia

Locus	Alelo	Localidade (M)											
		1 (33)	3 (39)	5 (41)	7 (40)	8 (40)	9 (40)	12 (41)	14 (41)	17 (41)	18 (40)	20 (41)	Médias (437)
Aat-1	b	1,000	1,000	1,000	0,410	0,887	1,000	0,634	1,000	1,000	0,975	0,667	0,869
	c	0,0	0,0	0,0	0,590	0,112	0,0	0,366	0,0	0,0	0,025	0,333	0,131
	h	0,00	0,000	0,00	0,484	0,201	0,000	0,464	0,000	0,000	0,005	0,4444	
Acph-2	a	0,136	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,010
	b	0,0	0,487	0,125	0,050	0,350	0,692	0,073	0,3411	0,0	0,0	0,0	0,193
	f	0,864	0,487	0,875	0,950	0,650	0,308	0,927	0,659	0,024	1,000	2,000	0,702
	g	0,0	0,026	0,0	0,0	0,0	0,0	0,0	0,0	0,976	0,0	0,0	0,094
	h	0,265	0,525	0,219	0,095	0,455	0,426	0,135	0,449	0,047	0,000	0,000	
Acgh-3	a	0,548	0,462	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,081
	d	0,032	0,205	0,049	1,000	0,737	0,975	0,300	0,951	0,976	0,050	0,0	0,488
	e	0,065	0,128	0,0	0,0	0,0	0,0	0,400	0,0	0,0	0,0	0,0	0,053
	f	0,355	0,205	0,951	0,0	0,262	0,025	0,300	0,049	0,024	0,700	1,000	0,355
	h	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,250	0,0	0,023
	h	0,165	0,000	0,000	0,223	0,223	0,223	0,509	0,580	0,000	0,410	0,369	
Adh-1	a	0,091	0,0	0,0	0,872	0,128	0,128	0,610	0,400	0,0	0,289	0,244	0,255
	b	0,909	1,000	1,000	0,128	0,872	0,872	0,341	0,500	0,0	0,711	0,756	0,635
	c	0,0	0,0	0,0	0,0	0,0	0,0	0,049	0,100	1,000	0,0	0,0	0,110
	h	0,165	0,000	0,000	0,223	0,223	0,223	0,509	0,580	0,000	0,410	0,369	
Adh-2	a	0,0	0,026	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,002
	c	1,000	0,974	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,998
	h	0,000	0,051	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	
Cat	a	1,000	1,000	1,000	1,000	1,000	1,000	1,0000	,917	1,000	1,000	1,000	0,993
	b	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,083	0,0	0,0	0,0	0,007
	h	0,000	0,051	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	

continua...

TABELA 12. Continuação

Locus	Alelo	Localidade (M)											
		1 (33)	3 (39)	5 (41)	7 (40)	8 (40)	9 (40)	12 (41)	14 (41)	17 (41)	18 (40)	20 (41)	Médias (437)
Est-1	a	0,0	0,077	0,756	0,100	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,087
	b	0,0	0,0	0,0	0,125	0,025	0,0	0,0	0,0	0,0	0,0	0,0	0,014
	e	1,000	0,923	0,244	0,775	0,975	1,000	1,000	1,000	1,000	1,000	1,000	0,899
	h	0,000	0,051	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	
Est-2	a	0,0	0,205	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,018
	e	0,0	0,0	0,0	0,0	0,0	0,0	0,512	0,0	0,0	0,0	0,0	0,048
	f	0,733	0,590	0,049	0,075	0,050	0,625	0,415	0,512	0,439	0,350	0,0	0,339
	h	0,0-	0,0	0,049	0,050	0,887	0,375	0,073	0,488	0,561	0,650	0,927	0,379
	j	0,267	0,128	0,073	0,675	0,012	0,0	0,0	0,0	0,0	0,0	0,073	0,107
	q	0,0	0,077	0,829	0,200	0,050	0,0	0,0	0,0	0,0	0,0	0,0	0,108
	h	0,391	0,587	0,305	0,496	0,209	0,469	0,560	0,500	0,493	0,455	0,135	
Est-4	b	0,0	0,0	0,050	0,050	0,375	0,250	0,075	0,488	1,000	0,700	1,000	0,379
	c	0,0	0,0	0,025	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,002
	f	0,0	0,250	0,150	0,0775	0,075	0,0	0,500	0,073	0,0	0,0	0,0	0,169
	g	1,000	0,750	0,775	0,175	0,550	0,750	0,425	0,439	0,0	0,300	0,0	0,450
	h	0,000	0,375	0,368	0,366	0,551	0,375	0,564	0,564	0,000	0,420	0,000	
Est-5	a	0,0	0,359	0,659	0,250	0,0	0,075	0,220	0,541	0,0	0,079	0,229	0,202
	b	0,467	0,615	0,049	0,0	0,050	0,800	0,098	0,054	1,000	0,0	0,0	0,285
	c	0,533	0,0	0,244	0,975	0,925	0,125	0,610	0,405	0,0	0,921	0,686	0,489
	e	0,0	0,026	0,049	0,0	0,0	0,0	0,073	0,0	0,0	0,0	0,086	0,021
	f	0,0	0,0	0,0	0,0	0,025	0,0	0,0	0,0	0,0	0,0	0,0	0,002
	h	0,498	0,492	0,501	0,049	0,141	0,339	0,565	0,540	0,000	0,152	0,470	
Ipo	a	0,0	0,0	0,0829	0,0	0,0	0,0	0,0	0,049	0,0	0,0	0,0	0,083
	b	0,758	0,333	0,0	0,0	0,564	0,0	0,512	0,732	0,024	0,0	0,537	0,309
	c	0,242	0,667	0,171	1,000	0,436	1,000	0,488	0,220	0,976	1,000	0,463	0,608

Continua...

TABELA 12. Continuação

Locus	Alelo	Localidade (M)											
		1 (33)	3 (39)	5 (41)	7 (40)	8 (40)	9 (40)	12 (41)	14 (41)	17 (41)	18 (40)	20 (41)	Médias (437)
Mdh-2	a	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,073	0,007
	e	1,000	21,000	1,000	1,000	1,000	0,975	1,000	1,000	1,000	1,000	0,927	0,991
	g	0,0	0,0	0,0	0,0	0,0	0,025	0,0	0,0	0,0	0,0	0,0	0,002
	h	0,000	0,000	0,000	0,000	0,000	0,049	0,000	0,000	0,000	0,000	0,135	
Nadhad-1	a	0,563	`0,692	0,805	0,075	0,050	0,250	0,439	0,902	0,0	0,0	0,537	0,390
	c	0,438	0,308	0,195	0,925	0,950	0,750	0,561	0,098	1,000	1,000	0,463	0,610
	h	0,491	0,426	0,314	0,139	0,095	0,375	0,685	0,177	0,000	0,000	0,497	
Nadhad-2	b	1,000	0,974	0,951	0,375	0,912	1,000	0,951	0,949	1,000	0,975	1,000	0,916
	d	0,0	0,026	0,049	0,550	0,037	0,0	0,0	0,51	0,0	0,025	0,0	0,068
	e	0,0	0,0	0,0	0,075	0,050	0,0	0,049	0,0	0,0	0,0	0,0	0,016
	h	0,000	0,051	0,094	0,551	0,164	0,000	0,093	0,097	0,000	0,048	0,000	
Pept-1	b	1,000	1,000	0,317	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,935
	c	0,0	0,0	0,683	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,065
Pgi	a	0,161	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,012
	b	0,839	0,872	1,000	0,900	1,000	1,000	0,902	0,463	1,000	1,000	1,000	0,908
	c	0,0	0,0	0,0	0,0	0,0	0,0	0,098	0,537	0,0	0,0	0,0	0,060
	d	0,0	0,128	0,0	0,100	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,021
	h	0,270	0,223	0,00	0,000	0,000	0,000	0,177	0,497	0,000	0,000	0,000	
Pgm	b	0,879	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	1,000	0,991
	c	0,121	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,0	0,009
	h	0,213	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	0,000	

FONTE: NEVO, E; Z0HARY, D; BAILES, A; KAPLAN, D E STORCH, N. 1986

isto é, somando-se os valores de **h** da coluna correspondente à população em questão e dividindo-se o resultado por 30, que é o número de alelos na população total. Adotando-se o mesmo procedimento podem ser obtidos os valores de **H** para as demais populações.

O próximo passo é a partição da diversidade genética total em suas componentes entre e dentro. Para isto vai ser escolhido o locus Cat (catalase) cujas freqüências estão contidas na tabela já referida.

Com a finalidade de simplificar os cálculos, os valores das freqüências gênicas são dispostos segundo a Tabela 13 a seguir, na qual os valores de J são calculados para cada linha pela expressão

$$\sum_{i=1}^{2} x^2_i$$

e P(a) e P(b) são as médias ponderadas das freqüências gênicas, obtidas multiplicando-se o valor de n de cada população pelas respectivas freqüências dos alelos.

A decomposição de H_t em seus componentes é feita calculando-se:

$$J_t = (0,993)^2 + (0,007)^2 \text{ ou}$$
$$J_t = 0,986$$

onde 0,993 corresponde a p(a) e 0,007 a p(b) constantes da Tabela 13
Daí, estima-se H_t que, como já foi referido, equivale a:

$$H_t = 1 - J_t \text{ ou } H_t = 1 - 0,986 = 0,014$$

Este valor corresponde à diversidade gênica total para o locus Cat estudado.

Posto isto, obtém-se J_S ou a identidade gênica média dentro, que equivale a:

$$J_S = \frac{1}{C} C_i J_i$$

onde $C = \sum C_i$ é o número total de indivíduos e C_i os valores da tabela anterior correspondentes a n ou o número de indivíduos dentro de cada população e $C_i J_i$ o produto de n pelas identidades gênicas respectivas.

Aplicando-se a fórmula correspondente a J_S aos dados da Tabela 13, obtém-se:

$$J_S = \frac{1}{437} (1,00 \times 33 + \ldots + 1,00 \times 41) \text{ ou } J_S = 0,987$$

A partir daí, calcula-se Hs, que equivale a:

$$H_S = 1 - J_S \text{ ou } H_S = 1 - 0,987 \; H_S - 0,013$$

que é a diversidade gênica dentro. Portanto, D_{st} ou diversidade entre obtém-se por diferença

$$D_{st} = 0,014 - 0,013 = D_{st} = 0,001$$

Pode-se observar que H_t foi decomposto em:

TABELA 13. Freqüências alélicas para o locus Cat com as respectivas identidades gênicas J

Populações	n	Freqüências Alélicas		Identidades Gênica (J)
		a	b	
1	33	1,000	0,000	1,000
3	39	1,000	0,000	1,000
5	41	1,000	0,000	1,000
7	40	1,000	0,000	1,000
8	40	1,000	0,000	1,000
9	40	1,000	0,000	1,000
12	41	1,000	0,000	1,000
14	41	0,917	0,083	0,847
17	41	1,000	0,000	1,000
18	40	1,000	0,000	1,000
20	41	1,000	0,000	1,000
Médias		p(a) = 0,993	p(b) = 0,007	

$$H_t = H_S + D_{st} \text{ ou } 0,014 = 0,013 + 0,001$$

isto é, na diversidade gênica média dentro (H_S) e entre (D_{st}).

Com estes valores pode-se estimar, ainda, o valor de G_{st}, isto é, o coeficiente da diferenciação para o locus em questão, que no caso equivale a:

$$G_{st} = \frac{H_t - H_s}{H_t} \text{ ou } G_{st} = 0{,}071$$

No trabalho de Nevo *et al* (1986) foram calculados os valores de H, H_t e G_{st}, para os demais locus dos sistemas isoenzimáticos estudados. Verificou-se, a partir dos dados relativos ao primeiro parâmetro que, à exceção da população 17, todas as outras foram caracterizadas por altos níveis de diversidade genética (H). Com base na diversidade gênica total observou-se que os H_t's variaram de 0,00 a 0,72 com média de 0,18. A diferenciação relativa média entre as populações estudadas foi de $G_{st} = 0{,}47$, com variação entre os valores de 0,02 a 0,66. Verificou-se, ainda, que os loci Pept-1, Acph-3, Est-1 e Adn-1 foram os que exibiram a maior diferenciação interpopulações.

5.4. DISTÂNCIA GENÉTICA DE NEI

A distância genética de Nei (1972) é outra medida que, via regra, acompanha os trabalhos dirigidos para o estudo da diversidade genética. É usada quando se dispõe das freqüências dos genes envolvidos na determinação das isoenzimas, com a finalidade de medir a similaridade e dissimilaridade entre diferentes grupos taxonômicos.

A definição desta distância baseia-se no conceito da identidade dos genes em duas populações X e Y. Considere que estas populações são cruzadas ao acaso e nas quais os alelos múltiplos segregam para um locus. Representando-se por X_i e Y_i, respectivamente, as freqüências dos alelos em X e Y, então a probabilidade de identidade de dois genes escolhidos ao acaso é $J_X = \sum x^2_i$ na primeira e $J_Y = \sum Y^2_i$ na segunda população.

A probabilidade de identidade de um gen de X e outro de Y é, neste caso, a soma dos produtos das probabilidades, isto é, $J_{XY} = \sum X_i Y_i$.

Segundo Nei (1972) quando não há seleção e cada alelo é derivado de uma simples mutação em uma geração ancestral, então os valores esperados para J_X e J_Y são iguais aos coeficientes de endogamia de Wrigth em x e y, respectivamente, enquanto J_{XY} equivale ao coeficiente de parentesco de Malecot (1967).

Definidas estas quantidades pode-se calcular a identidade normalizada dos genes de x e y com respeito a este locus, a qual é definida pela expressão:

$$I_j = \frac{j_{xy}}{\sqrt{j_x \cdot j_y}}$$

A quantidade obtida de I tem valor 1 quando as duas populações têm os mesmos alelos com freqüências idênticas e 0 no caso em que os mesmos alelos não são comuns.

O I pode ser definido, também, com respeito a todos os loci presentes nas populações mencionadas.

Nesta situação é definido como:

$$I = \frac{J_{xy}}{\sqrt{J_x \cdot J_y}}$$

onde, J_x, J_y e J_{xy} são as médias aritméticas de J_x, J_y e J_{xy}, respectivamente, sobre todos os loci, incluindo os monomórficos.

A distância genética de Nei (1972) é definida a partir da expressão:

$$D = - \text{Log}_e I$$

onde I tem o significado expresso anteriormente.

O valor de D expresso desta maneira é apropriado quando a razão da substituição gênica por locus é a mesma em todos os loci. No entanto, quando esta condição não é satisfeita e, ainda, quando os valores de I são elevados, usa-se a expressão D' para a distância, que equivale a:

$$D' = \text{Log}_e I'$$

onde

$$I' = \frac{J'_{xy}}{\sqrt{J'_x \cdot J'_y}}$$

e J'_x, J'_y e J'_{xy} são as médias geométricas de J'_x, J'_y e J'_{xy}, respectivamente. Para maiores detalhes sobre esta distância consultar Nei (1972) citado na bibliografia.

Sabrah e El Metainy (1985) determinaram as distâncias genéticas de Nei (1972) entre cultivares locais e exóticas de feijão fava (*Vicia faba* L.) com vistas a medir as suas similaridades e dissimilaridades. Os autores utilizaram a isoenzima esterase cujos padrões isoenzimáticos foram revelados pela técnica da eletroforese gel em Agar-amido PVP. Dois padrões de bandas foram encontrados: o primeiro constituído de um grupo de bandas que migrava para o ânodo e outro através do cátodo.

As bandas que migravam para o ânodo eram controladas por dois genes A_1 e A_2, de acordo com a distância de migração para este polo. As dirigidas para o cátodo, também controlados por dois genes C_1 e C_2, e assim designados de acordo com as distâncias para este polo. Cada um dos quatro loci identificados constou de três alelos: um nulo (N), outro rápido (F) e o último lento (S). O primeiro apresentou-se como dominante em relação a F e S em todos os loci examinados, porém estes dois últimos mostraram entre si dominância incompleta. Na Tabela 14 são encontradas as freqüências alélicas para os quatro loci nas cinco cultivares, segundo Sabrah e El Metainy (1985).

5.4.1. EXEMPLO

Para mostrar como se calcula a distância genética de Nei (1972), considere as cultivares do Paquistão e Abissínia, nas quais as freqüências alélicas para os quatro loci constam da tabela citada anteriormente.

O primeiro passo para o cálculo deste parâmetro é a determinação das probabilidades de identidades, de dois genes escolhidos ao acaso em cada população.

Estas probabilidades são:

1. População do Paquistão:

Para esta população os valores de $j_x = \Sigma\, x^2_i$ para cada gen são de acordo com a Tabela citada

$$j_xA_1 = 0{,}038^2 + 0{,}615^2 + 0{,}347^2 = 0{,}500$$
$$j_xA_2 = 0{,}183^2 + 0{,}568^2 + 0{,}249^2 = 0{,}418$$
$$j_xC_1 = 0{,}183^2 + 0{,}363^2 + 0{,}454^2 = 0{,}371$$
$$j_xC_2 = 0{,}000^2 + 0{,}481^2 + 0{,}519^2 = 0{,}500$$

TABELA 14. Freqüências alélicas dos quatro loci da isoenzima esterase para cinco cultivares de Feijão Fava

Genes		Paquistaniana	Abissiniana	Ciperiana Romi	Giza-1	Rebaya-40
Ânodo						
A_1	N	0,038	0,018	-	-	0,044
	F	0,615	0,456	0,621	0,537	0,501
	S	0,347	0,526	0,379	0,463	0,455
A_2	N	0,183	0,109	0,149	0,280	0,309
	F	0,568	0,387	0,588	0,412	0,378
	S	0,249	0,504	0,263	0,308	0,313
Cátodo						
C_1	N	0,183	0,090	0,129	0,057	0,140
	F	0,363	0,351	0,455	0,530	0,531
	S	0,454	0,379	0,416	0,413	0,329
C_2	N	-	0,035	-	-	0,091
	F	0,481	0,483	0,500	0,445	0,240
	S	0,519	0,482	0,500	0,555	0,669

FONTE: SABRAH, N; EL-METAINY, A.Y. (1985)

A média desses valores é $j_X = 0,447$ e diz respeito à identidade correspondente a todos os loci da população.

2. População da Abissinia:

Para esta população os valores de $j_Y = \sum y^2_i$, de acordo com a mesma tabela, correspondem a:

$$j_yA_1 = 0,018^2 + 0,456^2 + 0,526^2 = 0,485$$
$$j_yA_2 = 0,109^2 + 0,387^2 + 0,504 = 0,416$$
$$j_yC_1 = 0,090^2 + 0,351^2 + 0,379^2 = 0,275$$
$$j_yC_2 = 0,035^2 + 0,483^2 + 0,482^2 = 0,467$$

cuja média corresponde a $j_Y = 0,410$ referente à identidade gênica para todos os loci da segunda população.

Feito isto, calculam-se os valores para as identidades gênicas de um gen da primeira e outro da segunda população. Para o gene A, com as bandas M, F e S das 2 populações (Tabela 14), estes valores são calculados do modo a seguir:

$$j_{xy}A_1 = 0,038 \times 0,018 + 0,615 \times 0,456 + 0,347 \times 0,526 = 0,464$$

isto é, pela soma dos produtos das freqüências dos dois genes nas primeira e segunda populações. Procedendo desta maneira, obtém-se os demais valores das identidades. Calcula-se em seguida a média dos $j_{xy} = 0,406$ que corresponde à identidade para todos os pares de loci tomados ao acaso nas primeira e segunda populações.

Com estes valores estima-se:

$$I = \frac{0,406}{\sqrt{0,447 \times 0,410}}$$

$$I = \frac{0,406}{0,428}$$

$$I = 0,948$$

O valor de I refere-se à identidade normalizada dos genes entre as duas populações.

Portanto, substituindo-se este valor de I na expressão correspondente a D, tem-se o valor:

$D = - \text{Log}_e I$ ou
$D = - \text{Log}_e (0,948)$
$D = 0,053$

que é a distância genética entre as duas populações estudadas. Este valor, comparativamente aos encontrados para os outros grupos, indica que as cultivares do Paquistão são distantes geneticamente daquelas da Abissínia.

Nestes estudos pode-se calcular, ainda, a média da homozigosidade que se obtém no caso da população do Paquistão, tomando-se a média dos valores de J_x. Esta quantidade no exemplo apresentado corresponde a:

$$\frac{0,500 + 0,418 + 0,371 + 0,500}{4} = 0,447$$

Adotando-se este mesmo procedimento para os outros pares envolvendo as demais populações, obtém-se a matriz com as distâncias de Nei (1972) onde a diagonal é ocupada pelas médias de homozigosidade, a parte acima da diagonal pelos valores de D e abaixo as identidades gênicas respectivas. Para detalhes sobre esta matriz consultar Sabrah e El Metainy (1985), p. 304.

Os estudos relacionados com a diversidade genética vêm sendo conduzidos por diversos autores em plantas dos mais variados tipos. Na Tabela 15 acham-se relacionados os autores, técnicas, utilizadas, fontes e tipos de plantas onde esses estudos já foram realizados até o presente.

TABELA 15. Relação dos autores, técnicas utilizadas e fontes para consulta de trabalhos dirigidos para o estudo da diversidade genética

Autores	Técnicas Utilizadas	Fontes	Plantas
Jain *et al* (1975)	Índice de Shannon e Weaver	Crop Sci v.15, p.700-704	Trigo
Halcomb *et al* (1977)	Idem	Euphytica v.26, p.441-50	Arroz
Toltent *et al* (1979)	Idem	Crop Sci 19, p.789-94	Cevada
Bekele (1984)	Idem	Heriditas v.100, p.131-54	Trigo
Sabrah & El Metainy (1985)	Distância Genética de Ney	Egypt. J. Genet. v.14, p.301-7	Fava
Negassa (1985)	Índice de Shannon e Weaver	Hereditas v.102, p.139-50	Cevada
Nevo *et al* (1986)	Análise da diversidade de Ney	Genética v.68, p.203-13	Cevada
Perri & McIntosh (1991)	Índice de Shannon e Weaver	Crop Sci v.21, p.1350-55	Soja
Perry & McIntosh (1991)	Análise da diversidade genética de Ney	Crop Sci v.31, p.1356-60	Soja
Pecetti (1992)	Índice de Shannon e Weaver e análise da variável canônica	Euphytica, v.60, p.229-38	Trigo

6. METODOLOGIA PARA O ESTUDO DA DIVERGÊNCIA GENÉTICA

6.1. INTRODUÇÃO

O estudo da divergência genética das coleções de germoplasma é realizado adotando-se as técnicas já repassadas nos itens anteriores. Assim sendo, pode-se utilizar as análises das componentes e coordenadas principais (item 2 e 3) e de agrupamento (item 4) e as distâncias Euclidianas (item 2) ou de Mahalannobis (1936) (item 4). Nos itens a seguir serão discutidos os passos principais envolvidos nos estudos sobre a divergência genética com apresentação, inclusive, de exemplo de como as técnicas citadas funcionam na prática.

6.2. PASSOS DA METODOLOGIA

A maioria dos autores adota uma seqüência para estudo da divergência genética que obedece às seguintes etapas: 1) obtenção da matriz com as distâncias de Mahalanobis ou Euclidianas; 2) agrupamento dos genótipos, pelo método de Tocher; 3) cálculo das distâncias médias intra e interconglomerados; 4) identificação dos genótipos de maior divergência genética e 5) determinação do caráter com maior contribuição para a divergência genética.

Cumpridas estas etapas, passa-se então, a estudar as relações entre a divergência genética dos pais e a heterose nos descendentes, quando estes dados são disponíveis, como ocorre nos estudos envolvendo os cruzamentos dialélicos. Esta é a parte mais importante dos estudos relacionados com a divergência genética. A medida desta relação é baseada no coeficiente de correlação existente entre os valores da heterose e a divergência genética média dos progenitores.

6.3. EXEMPLO

Para ilustrar como é procedido o estudo da divergência genética vão ser considerados os dados da avaliação de sete germoplasmas de sisal, a

seguir: dois híbridos (11648 e da Paraíba), duas cultivares (IAC 034 e IAC 069), dois mutantes (um da Paraíba e outro da Bahia) e um representante do *A. Sisalana*. Os acessos fazem parte do Banco Ativo de Germoplasma mantido em Monteiro, PB, pelo Centro Nacional de Pesquisa de Algodão. Os dados basearam-se nas médias de três colheitas realizadas nos anos de 1987, 1988 e 1989, em quatro plantas tomadas ao acaso de cada fileira. As análises foram procedidas usando-se as médias de peso da fibra seca (PFS), peso de uma folha (PIF) e rendimento médio de fibra (RF), medidos em cada acesso.

A Tabela 16 encerra a matriz com as distâncias Euclidianas médias entre os sete germoplasma de sisal estudados. Para o cálculo dessas distâncias foram usados os escores dos primeiros componentes principais que corresponderam a um percentual da variação total superior a 90%. Como esta transformação elimina a correlação entre as variáveis, então estas distâncias são equivalentes às de Mahalanobis (1936) obtidas a partir dos dados originais.

TABELA 16. Distâncias Euclidianas Médias (d^2) entre o Sete Germoplasma de Sisal. Dados de 1987/89

Híbrido ou cultivar	1 H-11648	2 H-PB	3 M-Ba	4 Si	5 IAC-034	6 IAC-069	7 M-PB
1 H-11648	0						
2 H-PB	0,398	0					
3 M-Ba	0,965	0,798	0				
4 Si	1,775	1,604	1,208	0			
5 IAC-034	1,151	1,174	0,652	1,116	0		
6 IAC-069	1,062	0,932	0,858	0,745	0,867	0	
7 M-PB	0,707	0,447	0,452	1,212	0,900	0,643	0

H-11648 = híbrido 11648; H-PB = híbrido da Paraíba; M-Ba = mutante da Bahia; Si = sisalana; M-PB = mutante da Paraíba

Aplicando-se nesta matriz as operações descritas no item 4.3.2.1 correspondentes ao método de Tocher, obtém-se os cinco grupos de germoplasma constantes da Tabela 17.

A partir daí procede-se ao cálculo das distâncias intra e inter-conglomerados. Estas distâncias são calculadas utilizando-se os valores de d^2 entre as combinações dos acessos dentro dos grupos considerados. Por exemplo, para o grupo V constituído pelos germoplasmas IAC 069 (6) e Si (4) pode-se, para facilitar os cálculos, combinar os acessos do seguinte modo:

Acesso	6	4
6	-	
4	6 X 4	-

Como as distâncias entre os acessos com eles mesmos é nula resta, então, considerar a que envolve os germoplasma 6 e 4, cuja distância é, conforme a Tabela 16, 0,745. Logo, este valor corresponde à distância dentro do conglomerado V. Procedendo de forma semelhante obtém-se os valores de d^2 intraconglomerados correspondentes aos outros grupos. No caso dos valores de d^2 interconglomerados determina-se a distância média dos pares de acessos formados pela combinação dos elementos de um grupo contra os dos outros, de modo semelhante ao que é feito nos tabuleiros em xadrez nos estudos de genética. Por exemplo, o cálculo de d^2 entre os grupos IV e V é facilitado combinando-se os acessos IAC 034(5) do primeiro contra os germoplasma IAC 069(6) e Sisalana(4) do segundo, conforme representado a seguir:

Acesso	4	6
5	5 x 4	5 x 6

TABELA 17. Grupos de similaridade entre o sete germoplasma de Sisal formados a partir das distâncias Euclidianas médias (d^2)

Grupo	Nº de componentes	Cultivares ou híbridos
I	1	H-11648
II	2	H-PB e M-PB
III	1	M-Ba
IV	1	IAC 034
V	2	IAC 069 e Si

O valor de d^2 entre os conglomerados IV e V é obtido calculando-se:

$$\bar{d}^2 = \frac{1,116 + 0,867}{2} = 0,991$$

isto é, a média das distâncias entre as cultivares 5 e 4 e 5 e 6, pertencentes a cada um dos grupos. Adotando-se procedimento semelhante obtém-se as distâncias intraconglomerados envolvendo os demais grupos.

Na Tabela 18 são encontrados os valores de d^2 intra e inter-conglomerados calculados pelo processo descrito anteriormente.

No tocante às primeiras verifica-se que elas variaram de 0,000 para os grupos I, III e IV, constituídos de um só membro a 0,745 referente ao conglomerado V encerrando o IAC 069(6) e o Sisalana (4). Para as distâncias entre os conglomerados o valor máximo encontrado foi de 1,418 observado entre os conglomerados I e V, isto é, envolvendo de um lado o híbrido 11648(1) e do outro o IAC 069(6) e o Sisalana(4). Portanto, foram estes os conglomerados mais divergentes e os que deveriam ser objeto de mais atenção dos programas de melhoramento visando à hibridação.

A Tabela 19 contém os valores médios para cada conglomerado com respeito aos caracteres usados na medida da divergência genética. Observa-se

TABELA 18. Distâncias médias intra e inter-conglomerados entre os grupos formados com os sete germoplasma de Sisal

Grupo	I	II	III	IV	V
I	0,000				
II	0,552	0,447			
III	0,965	0,625	0,000		
IV	1,151	1,037	0,652	0,000	
V	1,418	1,097	1,033	0,991	0,745

TABELA 19. Médias das características dentro de cada conglomerado formado com os sete germoplasma de sisal

Conglomerado Nº	Nº de Acessos	Características			
		Nº folha planta	Peso de 1 folha	Peso de fibra sece	Rendimento fibra
I	1	109,25	588,97	2,29	4,02
II	2	90,25	554,51	1,67	4,13
III	1	69,32	645,90	1,64	4,03
IV	1	72,37	716,50	1,81	3,28
V	2	63,25	512,56	0,97	3,07

desta tabela que os conglomerados I e V apresentaram-se não só como os mais divergentes entre si como, ainda, mostraram desempenho contrastante para todas as características analisadas. Este fato não só confirma a classificação destes germoplasma quanto a divergência genética como, ainda, aponta para a importância dos mesmos nos programas de melhoramento baseados na hibridação.

No caso em questão, dever-se-ia tentar o cruzamento entre os membros desses grupos, objetivando explorar a heterose ou a busca de combinações desejáveis nas gerações avançadas depois da F_1. A condição de complementaridade entre os mesmos aponta, assim, para a possibilidade de que nos cruzamentos realizados se possa, inclusive, obter indivíduos transgressivos em relação aos pais.

Nesses estudos pode-se determinar, ainda, o caráter com maior contribuição para a divergência genética. A metodologia mais extensamente utilizada para a determinação deste caráter é por intermédio da freqüência de "rank" 1 que cada variável recebe quando proporciona a maior magnitude do quadrado da diferença entre as médias Y_i de cada par de genótipos (Cruz, 1987).

No emprego da metodologia parte-se das variáveis Y_i padronizados e não correlacionadas ou transformados pelos escores dos primeiros componentes principais, conforme já referido anteriormente. Para exemplificar como funciona a metodologia, considere os escores dos componentes principais a seguir:

Genótipos	y_1	y_2	y_3	y_4
1	-0,1655	0,6322	-0,3787	-2,0343
2	0,2267	-0,2274	-0,9579	-1,1839
3	-0,0311	-0,8558	0,7560	-0,3650
4	-0,0719	0,0113	-0,2775	2,9548
5	0,0766	0,3849	1,9840	-0,1414
6	0,0891	0,5050	-0,6269	1,1220
7	-0,1239	-0,4503	-0,4990	-0,3520

O cálculo de $D_{ij'}$, quando feito a partir de tais variáveis, é realizado adotando-se a expressão:

$$D^2_{ii'} = (Y_{i1} - Y_{I'1})^2 + (Y_{I2} - Y_{I'2})^2 + \ldots (Y_{IJ} - Y_{I'J})^2$$

isto é, subtraindo-se a linha J' da i' e elevando-se ao quadrado cada diferença portanto, no caso da distância D_{12} envolvendo os genótipos 1 e 2, esta expressão equivale a:

$$D^2_{12} = 0,1539 + 0,7389 + 0,3354 + 0,7232$$

Vê-se, desta expressão, que a magnitude de D^2_{12} é o resultado da contribuição de cada termo que participa no cálculo da distância citada. A primeira parcela em D^2_{12} corresponde à contribuição da variável Y_1, a segunda a Y_2, etc, e a última à variável Y_4.

Para a distância considerada, a parcela de mais alto valor é a última, correspondente à variável Y_4; logo, o "rank" 1 é atribuído a Y_4 nesta distância.

O cálculo das demais distâncias permite medir estas contribuições e com isto verificar os números de "ranks" 1 que recebem as diversas variáveis para cada valor de $D^2_{ii'}$.

No exemplo considerado são possíveis 21 combinações, duas a duas, envolvendo as distâncias entre os genótipos. Nestas, a variável Y_1 recebeu o rank 1 nove vezes, a Y2 2, a Y3 6 e Y4 3.

A contribuição da variável a_j para a divergência é dada calculando-se:

$$C_j = \frac{\text{Nº de "rank" 1 da j-ésima variável}}{I \times \frac{(I-1)}{2}} \times 100$$

onde I número de genótipos avaliados.

Aplicando-se a fórmula anterior obtém-se as seguintes contribuições para as variáveis estudadas:

Variável	Contribuição (%)
Y_1	0,00
Y_2	9,52
Y_3	28,57
Y_4	61,90

Logo, a maior contribuição para a divergência genética foi a proporcionada pela variável Y_4.

Deve ser lembrado que a metodologia apresentada, apesar de ser muito utilizada tem, contudo, falhas, segundo Sing (1981). Por esta razão o autor apresenta metodologia alternativa que permite estimar, também, a contribuição de cada variável para o total da divergência genética.

6.4. RELAÇÃO ENTRE DIVERGÊNCIA GENÉTICA E HETEROSE

Na verdade, a relação que se busca é a da distância genética entre os pais e a heterose. A distância genética baseada na composição das populações pode ser apreciada tomando-se a freqüência dos diferentes genótipos, isto é, a distância genotípica, ou a freqüência dos vários alelos para determinado locus, ou seja, a distância genética. É esta distância que, de fato, pode se achar ou não relacionada com a heterose.

A relação é medida através do coeficiente de correlação envolvendo os valores das distâncias (D^2) e os da heterose na F_1 ou F_2, expressa como o desvio dessas gerações em relação à média dos pais.

A associação entre a heterose e a distância tem sido intensamente investigada por diversos pesquisadores. Na Tabela 20 são encontradas a

relação dos autores, fontes para consulta e tipos de plantas nos quais estes estudos vêm sendo conduzidos até hoje.

Os estudos realizados apontam, de modo geral, para uma relação positiva entre a divergência genética (distância genética) e a heterose. Assim sendo, entre os melhoristas já se tem por norma que a magnitude da heterose é proporcional à distância genética entre os progenitores.

No entanto, este tipo de relação pode ser positiva, negativa ou até mesmo nula, conforme evidenciado em outros estudos. Para uma discussão mais detalhada do assunto consultar Ghaderi *et al* (1984).

Um outro aspecto importante desta questão diz respeito ao nível de divergência genética que seria mais adequado para a expressão da heterose. Neste particular, os resultados dos estudos também divergem. Moll *et al* (19862) em trabalho com o milho encontraram que a heterose aumentava na mesma proporção da divergência genética. Todavia, no amendoim foi mostrado por Arunachalan *et al* (1984) que as freqüências e magnitudes da heterose foram maiores nos cruzamentos entre pais pertencentes às classes de divergência intermediária do que entre as extremas.

TABELA 20. Relação dos autores, fontes para consulta de trabalhos envolvidos com a divergência genética nas coleções de Germoplasma de diferentes plantas

Autores	Fontes	Plantas
Moll *et al* (1962)	Crop Sci v.2: p.197-8	Milho
Joshi & Dhawan (1966)	Crop Indian Journal Genetics	Não especificada
Murty & Arunachalan (1996)	Idem V. 26, P.188-9	Não especificada
Goodman (1967)	Fitotecnia Latinoamericana	Milho
Singh & Gupta (1968)	Indian Journal Genetics and Plant Breeding v.28, p. 151-7	Algodão
Goodman & Patterniani (1969)	Economy Botany v.23, p 265-73	Milheto
Bath (1970)	Aust. J. Agric. Res. v.21, p.1-7	Trigo
Jeswani *et al* (1970)	Indian Journal Genetics & Plant Breeding v.30, p.11-25	Linho
Gupta & Singh (1970)	Idem v.30, p.212-21	Grão de Bico
Ram & Panwar (1970)	Idem vol.30, p.1-10	Arroz
Upadhya & Murty (1971)	Idem v.31, p. 63-71	Milheto
Sachan & Sharma (1971)	Idem v.31, p.86-93	Tomate
Govil & Murty (1973)	Idem v.33, p.253-260	Sorgo
Chandhary & Singh (1971)	Idem v.35, p.409-413	Cevada
Peter & Rai (1976)	Idem v.36, p.379-383	Tomate
Chandra (1977)	Enphytica v.26, p.141-48	Linho

Continua...

TABELA 20. Continuação

Autores	Fontes	Plantas
Katyar & Singh (1979)	Indian Journal Genetic & Plant Breeding v.39, p.354-8	Grão de Bico
Partap & Drankhar (1980)	Genet. Agr. v.34, p.323-330	Quiabo
Sindhy & Pandita (1980)	Genet. Agr. v.34, p.235-44	Batata
Sastry *et al* (1980)	Indian Journal Genetics & Plant Breeding v.40, p.140-148	Cevada
Rao *et al* (1980)	Idem v.40, p.73-85	Girassol
Rao *et al* (1981)	Idem v.41, p.179-85	Arroz
Singh *et al* (1981)	Idem v.41, p.168-90	Milheto
Jain *et al* (1981)	Idem v.41, p.220-5	Grão de Bico
Almaraz (1982)	Gen. Agr. v.36, p.23-30	Algodão
Kaldo & Sidhu (1982)	Idem v.36, p.1-7	Melão
Arunachalan *et al* (1982)	Oleagineux v.37, p.415-418	Amendoim
Kanwal *et al* (1983)	Theor. Appl Genet. v.65, p.263-267.	Arroz
Martinez Wilches *et al* (1983)	Crop. Sci. v.23, p.775-781	Milho
Maluf *et al* (1983)	Rev. Bras. Genet. v.3, p.443-460	Tomate
Jastara & Paroda (1983)	Indian Journal Genetic & Plant Breeding v.43, p.63-7	Trigo
Dhagat & Singh (1983)	Indian J. Genet. v. 43, p. 168-172	Milheto
Isleib & Whyne (1983)	Crop Sci. v.23, p.832-41	Amendoim
Cuartero *et al* (1983)	Annales de Edafologia y Agrobiologia v.43, p.1209-19	Pimentão
Ghaderi *et al* (1984)	Crop Sci. v.24, p.37-42.	Feijão e fava
Smith (1984)	Crop Sci. v.24, p.1041-5	Milho
Arunachalan & Bandiopadhyay (1984)	Indian Journal Genetics & Plant Breeding v.44, p.548-54	Não especificada
Arunachalan *et al* (1984)	Indian Journal Genetics & Plant Breeding v.24, p.226-34.	Mustarda
Singh & Gill (1984)	Idem v.44, p.506-13	Algodão
Das & Gupta (1984)	Idem v.44, p.243-7	Vigna mungo
Singh *et al* (1985)	Idem v.45, p.531-38	Soja
Shamsuddimia (1985)	Theor Appl. Genet. v. 70, p.306-8	Trigo
Prassad *et al* (1985)	Trop. Agric. v.62, p.237-42	Amendoim
Rao *et al* (1985)	Genet. Agr. v.39, p.237-48	Cana
Smith *et al* (1985)	Crop Sci. v.25, p.681-5	Milho
Nath *et al* (1985)	Euphytica v.24, p.441-47	Sorgo
Cox *et al* (1985)	Crop Sci. v.25, p.529-32	Soja

Continua...

TABELA 20. Continuação

Autores	Fontes	Plantas
Betl *et al* (1985)	Indian Journal Genetics & Plant Breeding v.45, p.368-375.	Triticale
Sharka & Rama (1986)	Trop. Agric. v.63, p.293-296	Soja
Bhagyalaksimi *et al* (1986)	Indian Journal Genetics & Plant Breeding v. 56, p.15-9	Cana
Kuruvadi (1988)	turrialba v. 38, p.267-81	Trigo
Miranda *et al* (1988)	Rev. Brasil. Genet. v.12, p.881-92	Batata
Miranda *et al* (1988)	Idem v.11, p.929-37	Pimentão

7. PROGRAMAS DE COMPUTADOR PARA O EMPREGO DAS TÉCNICAS DE ANÁLISE MULTIVARIADA NOS ESTUDOS DOS GERMOPLASMAS

7.1. INTRODUÇÃO

O estudo das coleções de germplasma baseia-se, comumente, no exame simultâneo de um grande número de populações avaliadas para diversas características de interesse agronômico ou fisiológico. Nestas condições, a análise da massa de dados obtidos é extremamente trabalhosa ou, senão, impossível por meio das calculadoras convencionais.

A saída, então, é utilizar os recursos computacionais com os quais se pode operar, sem maiores dificuldades, a multidimensionalidade inerente a tais tipos de dados. Nos itens a seguir serão descritos alguns programas de computador com as respectivas rotinas, os quais poderão servir de apoio aos interessados no emprego de técnicas multivariadas nos estudos com os germoplasma.

7.2. DESCRIÇÃO DOS PROGRAMAS

Os programas acham-se descritos em ambiente de software SOC/NTIA desenvolvido pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA). Portanto, não haverá problema no tocante ao seu emprego por parte dos responsáveis pelos Bancos Ativos de Germoplasma (BAG) coordenados pelo Centro Nacional de Recursos Genéticos e Biotecnologia. (CENARGEN).

7.2.1. PROGRAMA: Estandar.cm

OBJETIVO: Produzir uma transformação (padronização) em um conjunto de dados dispostos em formato matricial,. sempre que o conjunto de dados multivariados envolver variáveis com unidades distintas de medida.

DESCRIÇÃO: A leitura dos dados é feita através de um arquivo gerado pelo módulo genese ou aproveitamento de arquivo SOC gerado, internamente, por sub-rotina dentro de um arquivo programa. Este programa somente é operado com o SOFTWARE SOC/NTIA ativado e copiado em um subdiretório corrente ou em disquete. A transformação usada é $Z_i = (X_i - \overline{X})/S(X_i)$

LISTAGEM:

```
/*PADRONIZA AS COLUNAS DE UMA MATRIZ*/

Z1 = Padrao;
nc = ncol(Padrao);
nl = nlin(Padrao);
mc = csoma(Padrao)/nl;
S1 = csoma(Padrao);
I = 1;
enquanto (I <= nc)
{  J = 1;
   sq = 0;
   enquanto (J <= nl)
   {  sq = sq + (Padrao[J,I]^2);
      J = J + 1;
   }
   S1[,I] = sqrt((sq - ((S1[,I]^2)/nl))/(nl - 1));
   I = I + 1;
}
I = 1;
enquanto (I <= nc)
{  J = 1;
   enquanto (J <= nl)
   {  Z1[J,I] = (Padrao[J,I] - mc[,I])/S1[,I];
      J = J + 1;
   }
   I = I + 1;
}
grave Z1;
imprime Z1 $;
```

7.2.2. PROGRAMA: Varcovar.cm

OBJETIVO: Gerar uma matriz de variâncias e covariâncias de um conjunto de dados dispostos em forma matricial.

DESCRIÇÃO: A preparação inicial dos dados é feita através de um arquivo externo, editado em um formato matricial e utilizando opções que geram arquivos não documento, ou pela utilização de arquivo SOC/NTIA, já existente. A saída dos dados é uma matriz simétrica, real, onde os elementos a_{ii} são as variâncias e os demais (a_{ij}) as covariâncias. Este programa somente é operado com o SOC/NTIA ativo e copiado em um subdiretório corrente ou em disquete.

LISTAGEM:

```
/*GERA MATRIZ DE VARIÂNCIA E COVARIÂNCIAS*/

nc = ncol(D);
nl = nlin(D);
sc = csoma(D);
sqc = csquad(D);
M = D';
M = M*D;
I = 1;
enquanto (I <= nc) {
 J = 1;
 enquanto (J <= nc) {
  SE (I == J)  M[I,J] = ((sqc[,I] - (sc[,I]^2)/nl))/(nl - 1);
  SE (J > I)
    {
      K = 1;
      somap = 0;
      enquanto (K <= nl) {
        somap = somap + D[K,I]*!D[K,J];
        K = K + 1;
      }
      M[I,J] =(1/(nl - 1))*(somap - (sc[,I]*sc[,J])/nl);
      M[J,I] = M[I,J];
    }
```

```
   J = J + 1;
  }
  I = I + 1;
}
grave M;
imprime M $;
```

7.2.3. PROGRAMA: Corre.cm

OBJETIVO: Gerar uma matriz de correlações de um conjunto de dados dispostos em formato matricial.

DESCRIÇÃO: As informações geradas por este programa são utilizadas por outros módulos do SOC/NTIA. Nenhum teste de hipóteses é feito sobre as correlações e o programa é executado, geralmente, após a padronização dos dados com o programa estandar.cm. É operado somente com SOFTWARE SOC/NTIA ativado.

LISTAGEM:

```
/*GERA MATRIZ DE CORRELAÇÃO*/

nc = ncol(A);
nl = nlin(A);
sc = csoma(A);
sqc = csquad(A);
R = A';
R = R*A;
I = 1;
enquanto (I <= nl) {
 J = 1;
 enquanto (J <= nc) {
  SE (I == J)  R[I,J] = 1;
  SE (J > I)
    {
      K = 1;
      somap = 0;
      enquanto (K <= nl) {
```

```
            somap = somap + A[K,I]*!A[K,J];
            K = K + 1;
          }
          R[I,J] = ( somap - sc[,I]*sc[,J]/nl )/
                sqrt( (sqc[,I] - (sc[,I]#2/nl))*(sqc[,J] - (sc[,J]#2/nl)));
          R[J,I] = R[I,J];
        }
      J = J + 1;
    }
    I = I + 1;
}
grave R;
imprime R $;
```

7.2.4. PROGRAMA: Conver.cm

OBJETIVO: Gerar os escores dos componentes principais, operando com os dados padronizados e autovetores associadas às matrizes de correlação ou de variâncias e covariâncias.

DESCRIÇÃO: Os escores são obtidos internamente, através de operações com as matrizes de dados padronizados e autovetores. Os dados obtidos podem ser utilizados no cálculo de distâncias ou para o estudo da dispersão das componentes ou coordenadas principais. O programa somente é operado com o SOC/NTIA ativado e copiado em um subdiretório corrente.

LISTAGEM:

```
/* GERA ESCORES DOS COMPONENTES PRINCIPAIS */

NumLin = nlin(D);
NumCol = ncol(D);
Escore = criamat(NumLin,NumCol,0);
Escore1 = criamat(NumLin,NumCol,0);

I = 1;
enquanto (I <= NumLin)
{  J = 1;
```

```
  enquanto (J <= NumCol)
  { K = 1;
    enquanto (K <= NumCol)
    { Escore[I,J] = Escore[I,J] + D[I,K]*!C2[K,J];
      k = k + 1;
    }
    J = J + 1;
  }
  I = I + 1;
}
Escore1 = D*C2;
grave Escore1;
imprime Escore1 $;
grave Escore;
imprime Escore $;
```

7.2.5. PROGRAMA: Cpp.cm

OBJETIVO: Selecionar os escores associados aos dois maiores autovalores de uma matriz de correlação ou de variâncias e covariâncias, com vistas à preparação do gráfico com os componentes ou coordenadas principais.

DESCRIÇÃO: O arquivo é obtido através de operações com as matrizes de dados padronizados e autovetores. O programa somente é operado com o SOC/NTIA ativado e copiado em um diretório corrente.

LISTAGEM:

```
/* SELECIONA ESCORES PARA DIAGRAMA DE DISPERSÃO */

if (ncol(D) != nlin(c2)) {
  anote " ";
   anote "Nr. DE LINHAS DE C2 e' DIFERENTE DO Nr. DE COLUNAS
DE D";
  anote " ";
  fim;
}
ptos=nlin(D);
```

```
colc2=ncol(c2);
pontos=D[,1 2];
i=1;
enquanto (i<=ptos) {
 j=colc2-1;
 enquanto (j<=colc2) {
  pontos[i,j-colc2+2]=D[i,]*c2[,j];
  j=j+1;
 }
 i=i+1;
}
grave pontos;
anote " ";
imprime pontos $;
```

7.2.6. PROGRAMA: Matriz.cm

OBJETIVO: Produzir a transformação da matriz de distância para o estudo da análise das coordenadas principais.

DESCRIÇÃO: A leitura dos dados é feita através de um arquivo gerado pelo módulo gênese ou pelo aproveitamento de um arquivo SOC. Este programa somente é operado com o SOC/NTIA ativado e copiado em um subdiretório corrente.

LISTAGEM:

```
/*COORDENADA PRINCIPAL*/

a=-1/2*d;
om=ncol(a);
ml=lsoma(a)/om;
mc=csoma(a)/om;
mt=soma(a)/(om^2);
b=a;
i=1;
enquanto (i <= om)
 {
  j=1;
```

```
  enquanto (j <= om)
  {
  b[i,j]=a[i,j]-ml[i,1]-mc[1,j]+mt;
  b[j,i]=b[i,j];
  j=j+1;
  }
i=i+1;
}
grave b;
```

7.2.7. PROGRAMA: Dme2.cm

OBJETIVO: Gerar a distância Euclidiana média a partir de um conjunto de dados dispostos em forma matricial.

DESCRIÇÃO: A leitura dos dados é feita, diretamente pelo módulo gênese, através de um arquivo externo, ou pelo aproveitamento de um arquivo SOC, gerado internamente.

LINGUAGEM:

```
/*DISTÂNCIA EUCLIDIANA MÉDIA*/

nl=nlin(mat);
nc=ncol(mat);
ml=lsoma(mat);
matt=mat;
dme2=diag(ml)*0;
i=1;
enquanto (i<=(nl-1)) {
 j=i+1;
 enquanto (j<=nl) {
  k=1;
  enquanto (k<=nc) {
   dme2[i,j]=dme2[i,j]+(((matt[i,k]-matt[j,k])#2)/nc);
   k=k+1;
  }
  dme2[i,j]=sqrt(dme2[i,j]);
  dme2[j,i]=dme2[i,j];
```

```
   j=j+1;
  }
  i=i+1;
}
grave dme2;
imprime dme2 $;
```

7.2.8. PROGRAMA: dme4.cm

OBJETIVO: Gerar a matriz generalizada de Mahalanobis, através dos escores dos componentes principais, ou com dados obtidos pela técnica da condensação piotal.

DESCRIÇÃO: A preparação dos dados é feita através da geração dos escores pela multiplicação dos dados padronizados com a matriz dos autovetores normalizados. A operação deste programa só é possível com o SOFTWARE SOC/NTIA ativado.

LISTAGEM:

```
/*DISTÂNCIA EUCLIDIANA, MAHALANOBIS (D2) A PARTIR DOS
ESCORES DOS COMPONETES PRINCIPAIS, COND. PIVOTAL ETC*/

mat=leia "escore";
nl=nlin(mat);
nc=ncol(mat);
ml=lsoma(mat);
matt=mat;
dme4=diag(ml)*0;
i=1;
enquanto (i<=(nl-1)) {
 j=i+1;
 enquanto (j<=nl) {
  k=1;
  enquanto (k<=nc) {
   dme4[i,j]=dme4[i,j]+((matt[i,k]-matt[j,k])#2);
   k=k+1;
  }
  dme4[j,i]=dme4[i,j];
```

```
    j=j+1;
  }
  i=i+1;
}
grave dme4;
imprime dme4 $;
```

7.2.9. PROGRAMA: Binário.cm

OBJETIVO: Gerar a matriz de similaridade (simétrica) em um conjunto de dados binários (0 e 1) quando se trabalha com os padrões isoenzimáticos, visando ao cálculo do "Simple Matching Coeficient".

DESCRIÇÃO: As informações geradas por este programa são utilizadas para alimentar outros que executam análise de agrupamento (Cluster Analysis). A preparação dos dados é feita pelo módulo gênese do SOC/NTIA e, a partir daí, utilizadas a expressão $S_{ij} = (a + d)/(a + b + c + d)$ e a geração da matriz de similaridade. É operado somente com o SOFTWARE SOC/NTIA ativado e no diretório corrente.

LISTAGEM:

```
/*Gera matriz de similaridade para dados binÁrios*/

NumLin  = nlin(MatriBin);
NumCol  = ncol(MatriBin);
MatFreq = criamat((NumLin/2)*(NumLin-1),4,0);

/* Calculo da Matriz de Frequencias (a, b, c, d) */

i = 1;
k = 1;
enquanto (i <= NumLin - 1)  {
  j = i + 1;
  enquanto (j <= NumLin)  {
    w = 1;
    enquanto (w <= NumCol)  {
```

```
        se (MatriBin[i,w] == 0)  {
          se (MatriBin[j,w] == 0) MatFreq[k,1] = MatFreq[k,1] + 1;
          cc MatFreq[k,2] = MatFreq[k,2] + 1;
        }
        cc {  se (MatriBin[j,w] == 0) MatFreq[k,3] = MatFreq[k,3] + 1;
            cc MatFreq[k,4] = MatFreq[k,4] + 1;
        }
        w = w + 1;
      }
      k = k + 1;
      j = j + 1;
    }
    i = i + 1;
}

/* Calculo dos Coeficientes Binarios (CoefiBin) */

CoefiBin = criamat((NumLin/2)*(NumLin-1),1,0);
SomaFreq = lsoma(MatFreq);

i = 1;
enquanto (i <= (NumLin/2)*(NumLin-1))  {
  CoefiBin[i,] = (MatFreq[i,1] + MatFreq[i,4])/SomaFreq[i,];
  i = i + 1;
}

/* Geracao da Matriz de Similaridade (Similar) */

Similar = criamat(NumLin,NumLin,1);

i = 1;
k = 1;
enquanto (i <= NumLin)  {
  j = 1;
  enquanto (j <= NumLin)  {
    se (i < j)  {
      Similar[i,j] = CoefiBin[k,];
      Similar[j,i] = CoefiBin[k,];
      k = k + 1;
```

```
    }
    j = j + 1;
  }
  i = i + 1;
}
grave Similar;
imprime Similar $;
```

7.3. DESCRIÇÃO DAS ROTINAS

Na seqüência a seguir são apresentadas as rotinas para o cálculo dos componentes principais, coordenadas principais, cálculo de distâncias e geração da matriz de similaridade de dados binários para os programas descritos no item anterior. Nas Figuras 3 a 5 são encontrados os fluxogramas que facilitam o uso dessas rotinas para o emprego dos métodos multivariados nos estudos das coleções de germoplasma.

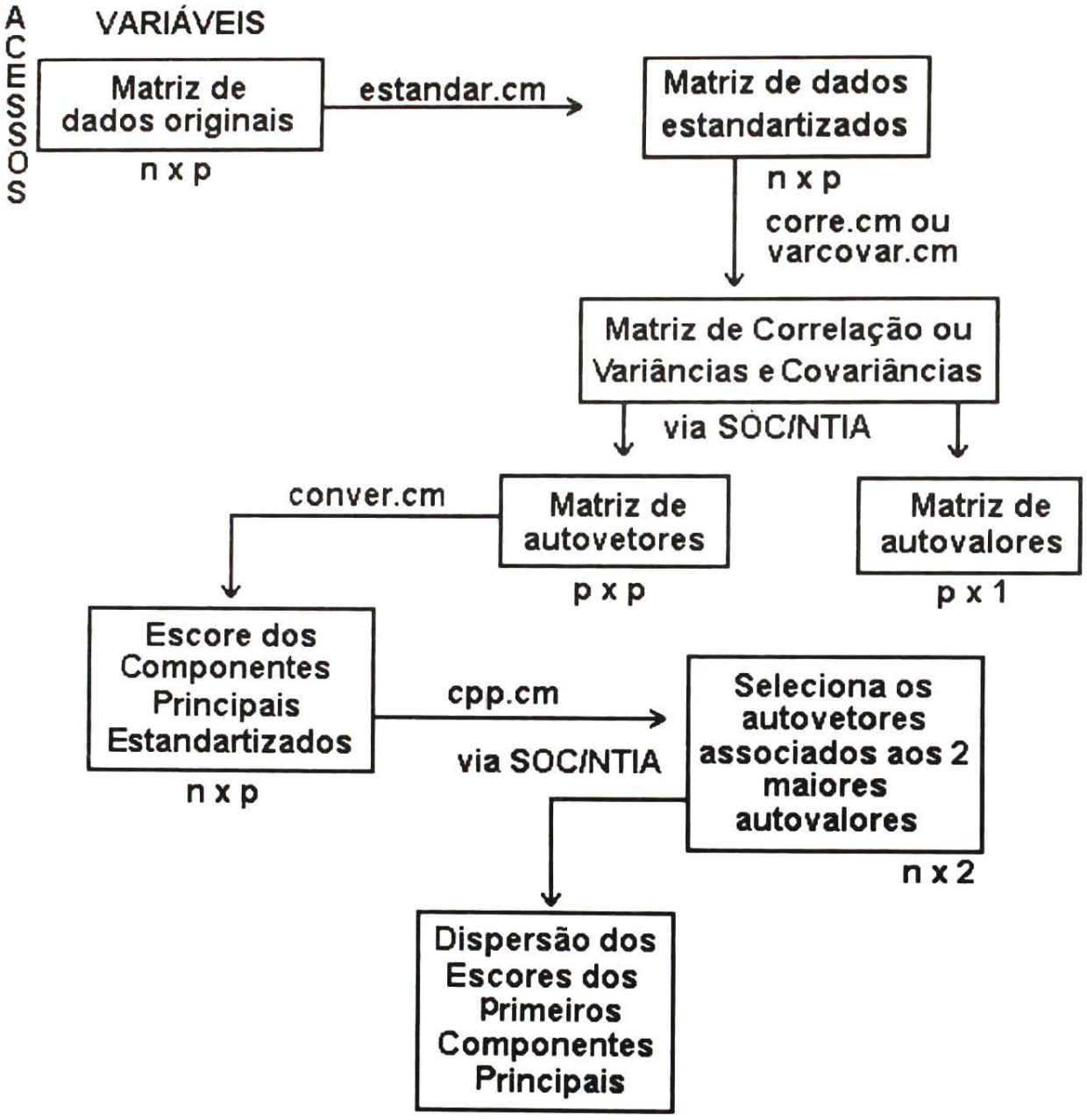


FIG. 3. Fluxograma das etapas da análise dos Componentes Principais

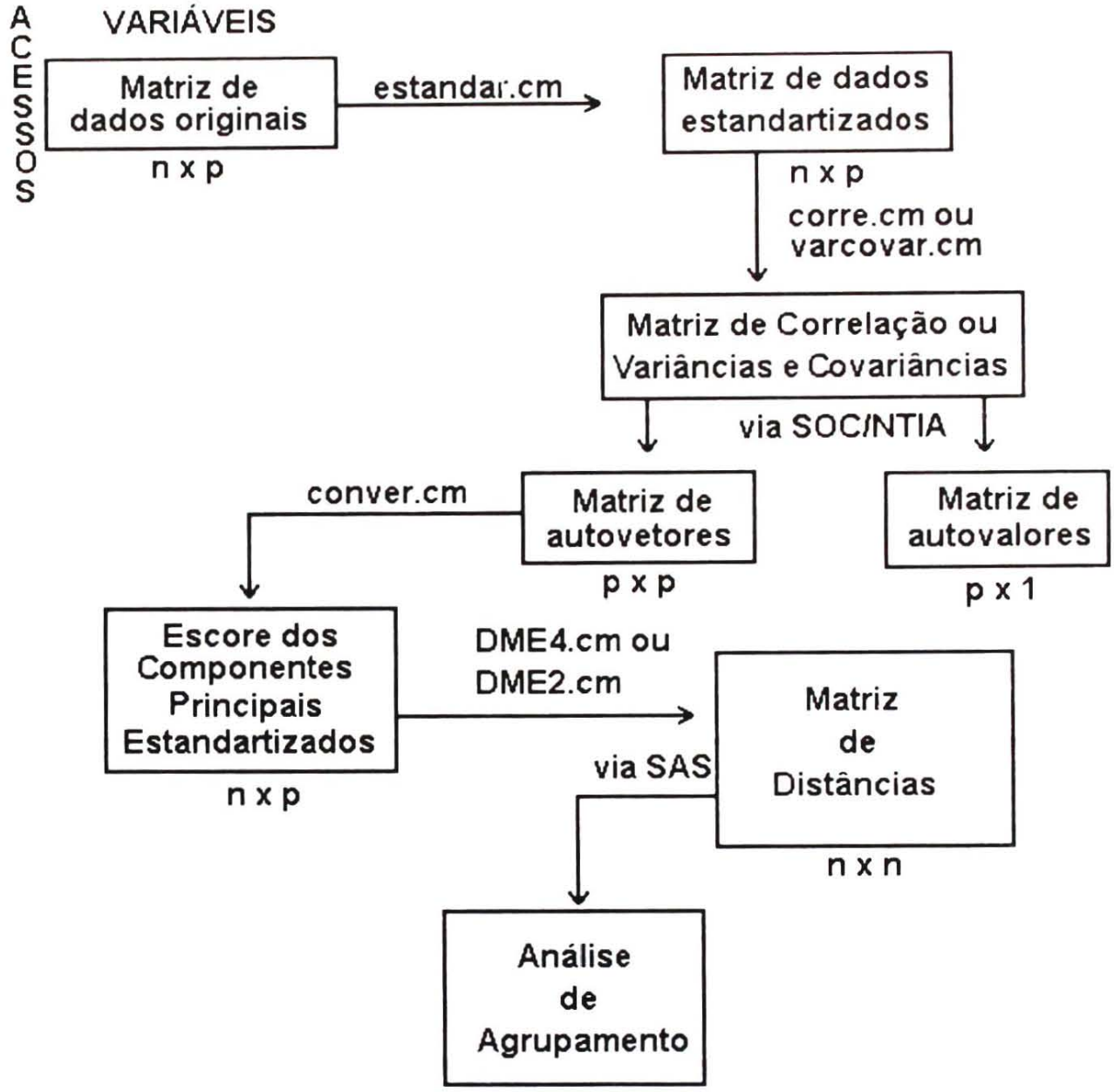


FIG. 4. Fluxograma das etapas para cálculo das distâncias

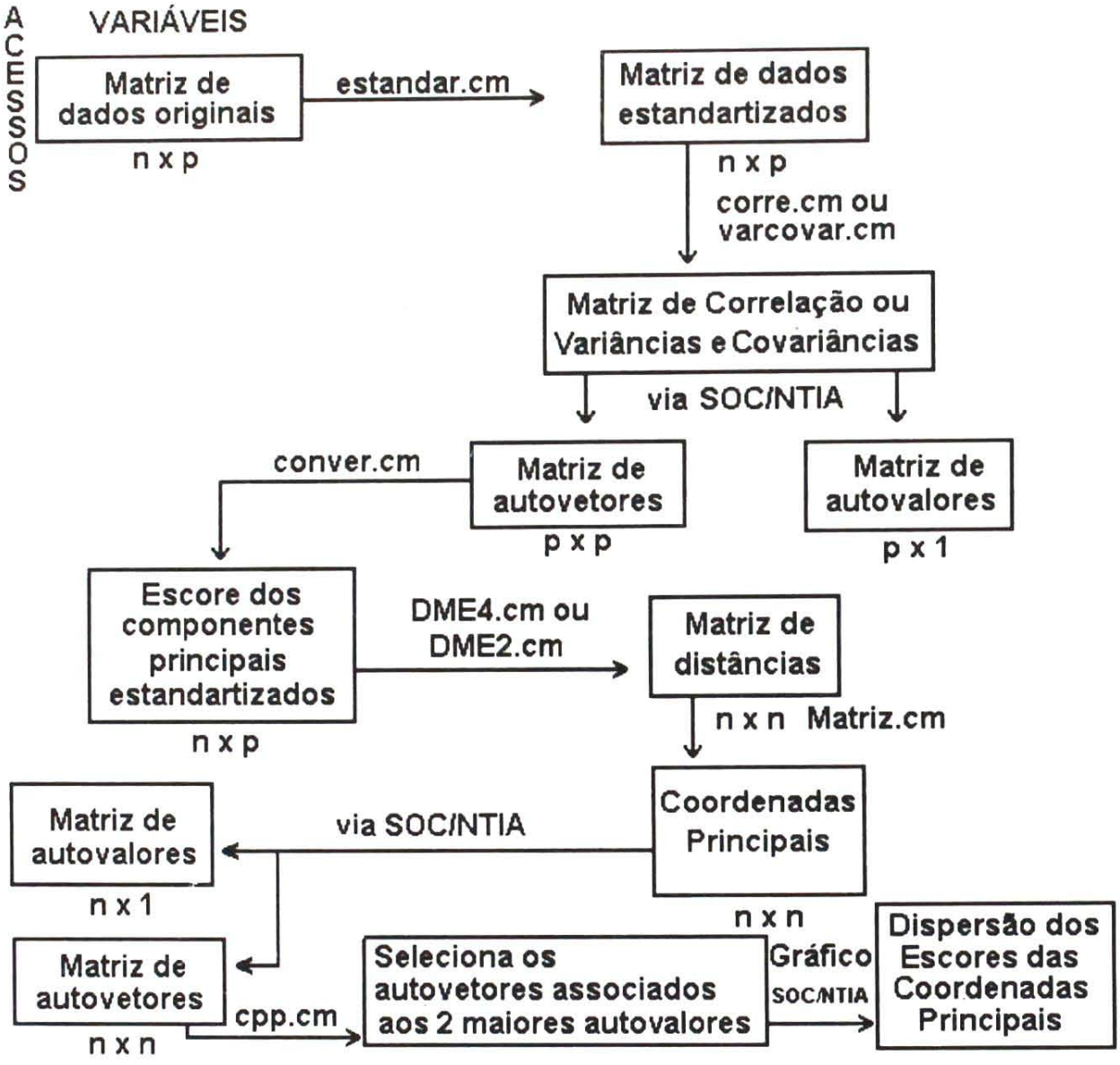


FIG. 5. Fluxograma das etapas da análise das Coordenadas Principais

7.3.1. EXEMPLO DE ROTINA PARA CÁLCULO DOS COMPONENTES PRINCIPAIS

```
genese Padrao
num x1 x2 x3 x4;
arquivo a=abref(dados.soc)x1 x2 x3 x4;
{leiaf(a);}
imprime -l80 Padrao
tl dados para calculo dos componentes principais;
}
estat Padrao
var x1 x2 x3 x4;
est n m dp ep cv;
}
cm
Padrao=leia"Padrao";
exec "estandar.cm";
fim;
cm
D=leia "Z1";
exec "varcovar.cm";
fim;
cm
M=leia "M";
c1,c2=auto(M);
c1;
c2;
grave c2;
fim;
cm
D=leia "Z1";
c2=leia "c2";
exec "conver.cm";
fim;
cm
D=leia "Z1";
c2=leia "c2";
```

```
exec "cpp.cm";
fim;
grafico -s3 -h0 -v0 pontos
graf col2*col1;
rot1 primeiro eixo;
rot2 segundo eixo;
}
```

7.3.2. EXEMPLO DE ROTINA PARA CÁLCULO DE DISTÂNCIAS

```
genese Padrao
num x1 x2 x3 x4;
arquivo a=abref(dados.soc) x1 x2 x3 x4;
{leiaf(a);}
imprime Padrao
tl dados para calculo de distancias;
}
cm
Padrao=leia"Padrao";
exec "estandar.cm";
fim;
cm
D=leia "Z1";
exec "varcovar.cm";
fim;
cm
M=leia "M";
c1,c2=auto(M);
c1;
c2;
grave c2;
fim;
cm
D=leia"Z1";
c2=leia"c2";
```

```
exec"conver.cm";
fim;
cm
mat=leia"escore";
grave mat;
exec "dme4.cm";
fim;
```

7.3.3. EXEMPLO DE ROTINA PARA CÁLCULO DAS COORDENADAS PRINCIPAIS

```
genese Padrao
num x1 x2 x3 x4;
arquivo a=abref(dados.soc) x1 x2 x3 x4;
{leiaf(a);}
imprime Padrao
tl dados para calculo das coordenadas principais;
}
cm
Padrao=leia"Padrao";
exec "estandar.cm";
fim;
cm
D=leia "Z1";
exec "varcovar.cm";
fim;
cm
M=leia "M";
c1,c2=auto(M);
c1;
c2;
grave c2;
fim;
cm
D=leia"Z1";
c2=leia"c2";
```

```
exec"conver.cm";
fim;
cm
mat=leia"escore";
grave mat;
exec "dme2.cm";
fim;
cm
d= leia "dme2";
grave d;
exec "matriz.cm";
M = leia "b";
c1,c2=auto(M);
c1;
grave c2;
D=leia "b";
grave D;
D1 = b*C2;
D1;
exec "cpp.cm";
fim;
grafico -s3 -h0 -v0 pontos
graf col2*col1;
eixo hor=-14 ate 18 2.0;
eixo ver =-5 ate 5 0.5;
rot1 primeiro eixo;
rot2 segundo eixo;
tl1 ANALISE DAS COORDENADAS PRINCIPAIS;
}
```

7.3.4. ROTINA PARA GERAÇÃO DA MATRIZ DE SIMILARIDADE A PARTIR DE DADOS BINÁRIOS

```
genese binario
num x1 x2 x3 x4;
arquivo arq1 = abref(dados.soc) x1 x2 x3 x4;
```

```
   leiaf(arq1);  }

imprime -l70 binario
tl listagem;
}

cm
MatriBin = leia "binario";
exec "binario.cm";
fim;
```

8. REFERÊNCIAS BIBLIOGRÁFICAS

AKINOLA, I.O.; WILLIANS, P.C. A numerical classification of Cajanus cajan (1.) Millsp. accessions based of morphological and agronomic attributes. Aust. Agric. Res., v.23, p.995-1005, 1973.

ALMARAJ, S.F.A.. Genetic divergence in Gossypium barbadense L. Genet. Agr.. v. 36, p.23-30, 1982.

ARNAND, I.I.; RAWAT, D.S. Genetic diversity, combining ability and heterosis in brown mustard. Indian J. genet. v. 44, p.226-234. 1984.

ARUNACHALAM, V.; BANDYOPADHYAY, A.; NIGHAN; S.H.; GIBBONS, R.W. Heterosis in relation to genetic divergence and specific combining ability in groundnut (Arachis hupogaea L.). Euphitica. v.33, p.33-39, 1984.

ARUNACHALAM, V.; BANDYOPADHYAYA, A. Limits of genetic divergence for occorrence of Heterosis - Experimental evidence from crop plants. Indian J. Genet. v.44, p.548-554, 1984.

ARUNACHALAN, V.; BANDYOPADHYAYA, A.; NIGAN, S.M.; GIBBONS, R.V. Heterotic potencial of single crosses in groundnut (Arachis hypogaea L.) Oleagineux. v.37, p.415-418, 1982.

BARTUAL, R.; CARBONELL, E.A.; GREEN, D.E. Multivariate analysis of a collection of soy bean cultivars for southwestern Spain. Euphytica. 34, p.113-123, 1985.

BEKELE, Endashaw. Analysis of regional pattern of phenotypic diversity in the ethiopian tetraploids and hexaploids wheats Heteditas. v. 100, p.131-154, 1984.

BEHL, R.K.; SINGH, V.P.; PARODA, R.S Genetic divergence in relation to heterosis and specific combining ability in triticale. Indian J. Genet., v. 45, p. 368-375, 1985.

BHAGYALAKSHMI, K.V.; NAGARAJAN, R.; NATARAJAN, B.V. Heterosis in some divergent sugar cane clones. Indian Journal Genetics & Plant Breeding. v. 56, p.15-9. 1986.

BHATT, G.M. Multivariate analysis approach to selection of parents for hybridization aiming at yield improvement in self pollinate crop. Australian Journal of Agricultural Research. v.21, p.1-7, 1970.

BUSEY, Philip.; BROSCHAT, TIMOTHY, K.; CENTER, BARBARA J. Classification of St. Augustinegrass. Crop Sci. v. 22, p.469-473, 1982.

CARADUS, I.R.; MACKAY, A.R.; WOODFIELD, D.R.; BOSCH, I. VANDEN; WEWLA, S. Classification of a world collection of white clovers cultivars. Euphytica. v.42, p.183-196, 1989.

CHANDRA, S. Comparison of Mahalanobis's method and metrogliph technique in the study of genetic divergence in Linum usitatissinum L. germplasm collection. Euphytica. v.26, p.141-148, 1977.

CHAUDHARY, B.D.; SINGH, V.P. Genetic divergence in some indian and exotic barley varieties and their hybrids. Indian Journal of Genetics & Plant Breeding. v.35, p. 409-413, 1975.

COX, T.S.; KIANG, Y.T.; GORMAAN, M.B.; RODGERS, D.M. Relationship between coeficient of parentage and genetic similarity indices in the soybean. Crop Sci. v. 25, p.529-32, 1985.

CUARTERO, I.; GOMES-GUILLAMON, M.L.; DIAZ, G.; SIMON, J.J.; CORBONELL, E. Agrupacion intraespecifica en variedades de pimiento. Anales de Edafologia y Agrobiologia, v.42, p. 1209-19, 1983.

DAS, P.K.; GUPTA, T.D.. Multivariate analysis in black gram (Vigna mungo (L.) Hepper). Indian J. Genet. & Plant Breeding, v. 44, p. 243-247, 1984.

DHAGAT, N.K.; SINGH, S.P. Genetic divergence in kodo millet. Indian J. Genet., v.43, p.168-172, 1983.

EDYE, L.A.; WILLIANS, W.T.; PRITCHARD, A.J. A numerical analysis of variation patterns in australians introductions of Glicine withtii (G. javanica). Aust. J. Agric. Res., v.21, p.57-69, 1970.

GHADERI, A.; ADAMS, M.W.; NASSIB, A.M. Relationship between genetic distance and heterosis for yield and morphological traits in dry edible bean and fava bean. Crop Sci. 24, p.37-42, 1984.

GOODMAN, M.M. The races of maize; I. The use of Mahalanobis generalized distances to measure morphological similarity. Fitotecnia Latino Americana, v. 4, p.1-23, 1967.

GOODMAN, M.M. PATTERNIANI, E. the races of maize. II. Choices of the appropriate characters for racial classification. Economy Botany. v.23, p.265-73, 1969.

GOVIL, J.M. MURTY, B.R. Genetic divergence and nature of heterosis in grain sorghum. Indian Journal of Genetics & Plant Breeding. v.33, p.253-260, 1973.

GUPTA, M.P. SINGH, R.B. Genetic divergence for yield and its components in green gram. Indian Journal Genetics Y Plant Breeding. v. 30, p.212-21, 1970.

HAMON, PERLA; TOURÉ, BARAKI. The classification of the cultivated yams (Discorea caynensis rotundata complex) of west África. Euphytica. v.47, p. 179-187, 1990.

HOLCOMB, I.; TOLBERT; D.M.; JAIM, S.K. A diversity analysis of genetic resources in rice. Euphytica. v. 26, p. 441--450, 1977.

HUSSAINI, S.H.; GOODMAN, M.M. TIMOTHY, D.H. Multivariate and geographical distribution of the world collection of finger millet. Crop Sci. v. 17, p. 257-263, 1977.

ISLEIB, T.G.; SWHYNE, J.C. Heterosis in test crosses of 27 exotic peanut cultivars. Crop Sci. v.23, p. 832-41, 1983.

JAIN, K.C.; PANDYA, B.P.; PANDE, K. Genetic divergence in chickpea. Indian Journal Genetics & Plant Breeding. v.41, p.220-5, 1981.

JAIN, S.K.; QUALSET, C.O.; BHATT, G.M.; WUV., K.K. Geographical patterns of phenotypic diversity in a world collection of durum wheats. Crop Scii. v.15, p. 700-704, 1975.

JASTARA, D.S.; PARODA, R.S. Genetic divergence in wheat. Indian Journal Genetics & Plant Breeding. v.43, p.63-7, 1983.

JESWANI, L.M.; MURTY, B.R.; MEHRA, R.B. Divergence in relation to geographical origins in a wold collection of linseed. Indian Journal genetics & Plant Breeding. v.30, p.11-25, 1970.

JOSHI, M.G.; DHAWAN, M.L. Genetic improvement in yeld with special reference to self-fertilizing. Crop Indian Journal Genetics. v.26, p.101-13, 1966.

KALDO, I.D.; SIDHU, A.S. Genetic divergence in muskmelon (Cucumis melo L.), Genetica Agraria. v.36, p.1-7, 1982.

KANWAL, K.S.; SINGH, R.M.; SINGH, I.; SINGH, R.B. Divergent gene pools in rice improvement. Theor Appl. Genet. v.65, p.263-267, 1983.

KATIYAR, R.P.; SINGH, S.P. Genetic divergence in chickpea. Indian Journal Genetics & Plant Breeding. v.39, p.354-8, 1979.

KURUVADI, S. Multivariate analysis of genetic divergence in wheat. Turrialba, v.38, p.267-271, 1988.

MALUF; W.R.; FERREIRA, P.E.; MIRANDA, J.E.C. Genetic divergence in tomatoes and its relationship with heterosis for yield in F_1 hybrids. Rev. Bras. Genet., v.3, p.453-460, 1983.

MARTIN, F.W.; RHODES; A.M. Subespecific grouping of eggplant cultivars. Euphytica. v.28, p.367-383, 1979.

MARTIN, F.W., RHODES; A.M. Intra-specific classification of Discorea alafa. Trop. Agric. v.54 m.l, p. 1-13, 1977.

MARTIN, F.W.; RHODES; A.M. The relationship of Dioscorea cayensis and D. rotundafa. Trop. Agric., v.55, p.193-206, 1978.

MARTINEZ WILCHES, O.J.; GOODMAN, M.M.; TIMOTHY, D.H. Measuring racial diferentiation in maize using multivariate distance measured stantardized by variation in F_2 population. Crop Sci, v.23, p. 775-781, 1983.

MIRAND, J.E.C.; CRUZ, C.D.; COSTA, C.P. Predição do comportamento de híbridos de pimentão (Capsicum annum L.) pela divergência genética dos progenitores. Rev. Bras. Genet., v. 11, p.929-937, 1988.

MIRANDA, J.E.C.; CRUZ, C.D. PEREIRA, A.S. Análise de trilha e divergência genética de cultivares de batata doce. Rev. Bras. Genet., v.12, p.881-892, 1988.

MOLL, R.H.; SALHUANA, W.S.; ROBINSON, H.F. Heterosis and genetic diversity in variety crosses in maize. Crop Sci. v.2, p.197-8, 1962.

MURTY, B.R.; ARUNACHALAN, L. The nature of genetic divergence in relation to breeding system in crop plants. Indian Journal Genetics & Plant Breeding. v. 26, p.188-9, 1966.

NATH, BHALA; ONURAN, ABRAS O.; HOUSE L. Genetic divergence among nonrestorer collection of sorghum (Sorghum bicolor (L.) Moench) and its relationship with heterosis. Euphytica. v.24, p.441-447, 1985.

NEGASSA, MULUGYETA. Patterns of phenotypic diversity in an ethiopian barley collection, and the Arussi-Bale highland as a center of origin of barley. Hereditas, v.102, p. 139-150, 1985.

NEI, M. Genetic distance between populations. Am Nat. v.106, p. 283-292, 1972.

NEI, M. Analysis of gene diversity in subdivided populations. Proc. Nat. Acad. Sci. v.70, p. 3321-3323, 1973.

NEVO, E.;ZOHARY, D.; BEILES, A.; KIAPLAN, D.; STORCH, M. Genetic diversity and environmental of wild barley, Hordeum spontaneum in Tukey. Genética, v.68, p. 203-213, 1986.

PARTAP, P.S.; DHANKHAR, B.S.; PANDITA, M.L.; DUDI, B.S. Genetic divergence in parents and their hybrids in okra (Abelmoschus sculentus L. Moench.). genet. Agr., v.34, p.323-330, 1980.

PECETTI LUCIANO; ANNICCHIARICO, PAOLO; DAMANIA, A.B. Diodiversity in germplasm collection of durum wheat. Euphytica, v. 60, p. 229-238, 1992.

PETER, K.V.; RAI, B. Genetic divergence in tomato. Indian Journal of Genetics & Plant Breeding. v.36, p.379-383, 1976.

PERRY, M.C.; Mc INTOSH, M.S. Geographical patterns of variation in the USDA a Soybean germplasm collection. I. Morphological traits. Crop Sci. v.31, p.1350-1355, 1991.

PERRY, M.C.; Mc INTOSH, M.S.; STONER, A.K. Geographical patterns of variation in the U.S.D.A. soybean germoplasm collection II. allozyme frequencies. Crop. Sci. v.31, p.1356-11360, 1991.

RAM, I. PANWAR, D.V.S. Intra-especific divergence in rice. Indian Journal Genetics & Plant Breeding. v.30, p.1-10, 1970.

RAO, C.P.; RAHMAN, M.A. RAO, P. NAGESWARA; REDDY, J.R. Genetic divergence analysis in sugarcane. Genet. Agr., v.39, p.237-248, 1985.

RAO, A.V. PRASSAD, A.S.R.; SAI KNISNHMA, T.; SIDHU, D.V.; SRINIVASON, T.E. Genetic divergence among some brown planthopper resistant rice varieties. Indian Journal Genetics & Plant Breeding. v.41, p.179-85, 1981.

RAO, V. RANGA; RAMACHANDRAN, M.; SHARMA, J.R. Multivariate analysis of genetic divergence in safflower. Indian Journal of Genetics & Plant Breeding, v. 40, p.73-85, 1980.

REZAI, A.; FREY, K.J. Multivariate analysis of variation among wild acession--seed traits. euphytica. v.49, p.111-119. 1990.

SABRAH, M.S.; EL METAINY, A.Y. Genetic distances between local and exotic cultivars of Vicia faba based on esterase isozyme variation. Egypt. I. Genetic. Cyto., v.14, p.301-307, 1985.

SASTRY, L.V.S. ISUNDARSEAN, S.; RAO, G.S.P.; RAO, U.M.B.; MURTY, B.R. Genetic divergence in two-row barley varieties for maltin quality characters. Indian Journal of Genetics & Plant Breeding. v.40, p.140-148, 1980.

SHAMSUDDIN, A.K.M.; Genetic diversity in relation to heterosis combining ability in sring wheat. Thear. Appl. Genet., v.70, p.306-308, 1985.

SHARMA, S.K.; RANA, N.D. A study of genetic divergence in a collection ofg small - seeded soybean acession. Trop. Agric. v.63, p.293-296, 1986.

SIDHU, A.S.; PANDITA, M.L. Genetic divergence for yield and its components in pofato. (Solanum tuberosum L.) Genetica Agraria, v.34, p.235-44, 1980.

SINGH, KARENDRA; RAM, HARI HAR. Genetic divergence in new breeding lines of soybean. Indian J. Genet. v.45, p.531-538, 1985.

SINGH, R.B.; GUPTA, M.P.. Multivariate analysis of divergence in upland cotton. Indian Journal of Genetics & Plant Breeding. v.28, p.151-7, 1968.

SINGH, R.R.; AWADHESH, K.; CHOUHAN, P.S. Genetic divergence in pearl millect. Indian Journal Genetics & Plant Breeding. v.41, p.168-90, 1981.

SINGH, T.H.; GIEL, S.S. Genetic diversity in Upland cotton under differents environments. Indian J. Genet. v.44, p.506-513, 1984.

SMITH, J.C.S. Genetic variability withim U.S. hybrid maize, multivariate analysis of isoenzyme data. Crop. Sci. v.24, p. 1041-5, 1984.

SMITH, J.C.S.; GOODMAN, M.M.;STUBER, C.N. Genetic variability with maize germplasm. II. Widely used inbred lines 1970 to 1979. Crop. Sci. vi25, p.681-5, 1985.

SMITH, S.E.; DOSS-AL, A.; WARBUTON. M. Morfhological and agronomic variation in north african and arabian alfafas. Crop. Sci., v.31, p.1159-1163, 1991.

TOLBERT, D.M.; QUALSET, C.O.; JAIN, S.K. Craddock, J.C. A diversity analysis of collection of barley. Crop Sci. v. 19, p. 798-794, 1979.

UPADHYAYA, M.K.; MURTY, B.R. Genetic diversity and combining ability in pearl millet. Indian Journal of Genetics & Plant Breeding. v.31, p.63-71, 1971.

UPADHYAYA, M.K.; MURTY, B.R. Genetic divergence in relation to geographical distribution in pearl millet. Indian Journal Genetics & Plant Breeding. v.30, p.704-15, 1970.

VERONESI, F.; FALCINELLI, M. Evaluation of an italian germplasm collection of Festyca arundinacea Scharad trhough a multivariate analysis. Euphytica. v.38, p.211-220, 1988.

ZEVEN, A.C.; SCHACHL, R. Grups of bread sheat landrasces in australian alps. Euphytica, v.41, p.235-246, 1989.

ZEVEN, ANTON, C.; VAN HINTUM, THEO J.L. Classification of landraces and improved cultivars of hexaploide wheats (Triticum aestivum, T. compactum and T. spelta) grow in the USA and discribed in 1922. Euphytica, v.59, p.33-47, 1992.

Impressão: EMBRAPA-SPI