AGRUPAMENTO DE DADOS DE MARCADORES MOLECULARES DOMINANTES UTILIZANDO O MÉTODO DE CLASSIFICAÇÃO DE PADRÕES SUPERVISIONADO

PEDRO H. R. CERQUEIRA[1], ROSELI APARECIDA LEANDRO[2], JULIANA GARCIA CESPEDES[3], CLÁUDIO LOPES DE SOUZA JR [4] e ANETE PEREIRA DE SOUZA[5]

1 Introdução

Os estudos de relações filogenéticas e de divergência genética têm sido uma das contribuições mais concretas de marcadores moleculares para a organização e melhoramento genético. O método de classificação de padrões pode ser utilizado neste tipo de estudo pois há o interesse em agrupar os indivíduos semelhantes, de forma que as maiores diferenças ocorram entre os grupos formados. (Duarte et.al, 1999)

Os marcadores moleculares fornecem informações sobre o material genético (DNA) dos indivíduos em estudo, com o emprego de métodos bioquímicos que identificam diferenças no DNA, revelando para uma dada região, uma marca ou uma banda que permite comparar os indivíduos em estudo quanto à sua presença ou ausência. Tais diferenças são codificadas na forma de uma matriz, cujas colunas identificam os genótipos, que podem ser linhagens, híbridos, etc. As linhas indicam as regiões do DNA onde foram avaliadas as diferenças nas características de interesse, sendo as bandas reveladas codificadas pelo número 1 e as não reveladas pelo número 0. Deste modo, os dados correspondem à uma matriz de zeros e uns. (Meyer, 2002)

O método de classificação de padrões pode ser utilizado visto que o interesse está em agrupar indivíduos em classes semelhantes. Este método pode ser: (a) supervisonado, no qual admite-se que a classe que gerou os padrões é conhecida; sabe-se, portanto, o número de classes utilizadas, e (b) não supervisionado, em que os padrões não se encon-

[1]ESALQ-USP, phr.cerqueira@usp.br.
[2]ESALQ-USP, raleandr@esalq.usp.br
[3]UNIFEI, jucespedes@hotmail.com
[4]ESALQ-USP
[5]UNICAMP

tram classificados; não se sabe qual o número de classes utilizadas. (Marques, 2005)

O objetivo deste trabalho é empregar o método de classificação de padrões supervisionado na classificação de diferentes linhagens de milho. Para tanto foram utilizados dados provenientes de uma análise com marcadores moleculares dominantes do tipo RAPD de 18 linhagens de milho provenientes de duas populações diferentes, BR-105 e BR-106.

2 Material e Métodos

Para a realização deste trabalho foram utilizadas 18 linhagens desenvolvidas pelo programa de melhoramento de milho do Departamento de Genética-ESALQ/USP, pelo professor Dr. Cláudio Lopes de Souza Jr. As linhagens são provenientes das populações dos milhos BR-105 e BR-106, desenvolvidas pelo Centro Nacional de Milho e Sogro (Embrapa Milho e Sogro). Ambas apresentam ciclo precoce e baixa altura da planta. A origem das linhagens utilizadas nesse trabalho é apresentada pela Tabela 1, sendo que 8 linhagens se originam da população BR-105 e 10 da população BR-106.

Tabela 1: Origem das 18 linhagens usadas no presente trabalho

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

Como observado, as linhagens pertencem a duas populações diferentes, sendo esperado que essas formem dois grupos quando aplicada o método de classificação de padrões.

O conjunto de dados observado são de variáveis binárias, ou seja, foi obtida uma matriz de zeros e uns. Desta maneira, para a classificação foi necessário utilizar técnicas de análise discreta para dados binários. A aplicação do método de classificação de padrões utiliza o Teorema de Bayes. Para os dados discretos têm-se:

$$P(w_j|\mathrm{x}) = \frac{P(\mathrm{x}|w_j)P(w_j)}{P(\mathrm{x})} \quad j = 1, \ldots, c, \tag{1}$$

sendo que c representa o número de classes ou populações, w_j a j-ésima classe e $P(\mathtt{x}) = \sum_{j=1}^{c} P(\mathtt{x}|w_j)P(w_j)$. Neste caso tem-se duas categorias e as bandas são independentes. Seja $\mathtt{x} = (x_1, \ldots, x_d)^t$, em que d corresponde ao número de bandas e x_i pode receber os valores 0 e 1, com:

$$p_i = P(x_i = 1|w_1) \tag{2}$$

$$q_i = P(x_i = 1|w_2) \tag{3}$$

Assumindo a independência condicional podemos escrever $P(x|w_i)$ como o produto das probabilidades dos componentes de x, assim temos que:

$$P(x|w_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i} \tag{4}$$

$$P(x|w_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}. \tag{5}$$

Considerando-se $g(x) = \ln \frac{P(x|w_1)P(w_1)}{P(x|w_2)P(w_2)}$ tem-se:

$$g(x) = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(w_1)}{P(w_2)} \tag{6}$$

A função $g(x)$ é chamada função discriminante é através dela que é possível definir se o vetor x pertence à classe w_1 ou w_2 Para valores de $g(x) > 0$ o vetor pertence à classe w_1, para valores de $g(x) < 0$ o vetor pertence à classe w_2.

Para o calcular a probabilidade de erro geral foi utilizado:

$$P_e(x) = 1 - \sum_{j=1}^{c} P_{jj}P(w_j) \tag{7}$$

sendo que $P_{jj} = P\{\hat{w} = w_j|w = w_j\}$.

3 Resultados e Discussões

Foram utilizados os dados de RAPD das 18 linhagens. Estes dados correspondem à uma matriz de zeros e uns, com 18 colunas, que correspondem às linhagens, e com 262 linhas que correspondem às bandas. Foram selecionadas ao acaso 5 linhagens da população 1 (w_1) e 5 linhagens da população 2 (w_2), para realizar o treino do classificador.

Após o treino, foram analisadas todas as linhagens para verificar se a linhagem observada corresponde a w_1 ou w_2, utilizando o método de classificação de padrões para dados binários. Após o procedimento foram obtidos os seguintes resultados para as linhagens analisadas:

Tabela 2: Classificação das Linhagens

Linhagem	g(x)	População verdadeira	classificação
1	436,02735	1	1
2	365,33391	1	1
3	215,30898	1	1
4	233,67641	1	1
5	267,29889	1	1
6	350,85123	1	1
7	318,14392	1	1
8	333,96847	1	1
9	-266,25086	2	2
10	-34,32254	2	2
11	-249,79482	2	2
12	-105,96194	2	2
13	-297,72638	2	2
14	-148,17630	2	2
15	-6,11976	2	2
16	201,04062	2	1
17	-267,94796	2	2
18	-313,64743	2	2

Na Tabela 2 observa-se a formação de dois grupos: as linhagens de um a nove

foram classificadas no grupo 1 (População 1) e as linhagens de 10 a 18, com exceção da 16, foram classificadas no grupo 2. Somente a linhagem 16 foi classificada erroneamente. A probabilidade de erro geral para esta classificações foi de 0.05. O método de classificação de padrões fornece bons resultados e é de fácil aplicação.

4 Referências

REFERÊNCIAS BIBLIOGRÁFICAS

DUARTE, M.C.; SANTOS, J.B; MELO, L.C Comparison of similarity coefficients based on RAPD markers in common bean. **Genetics an Molecular Biology**, v.22, n.3, p.427-432, 1999

DUDA, R.O.; HART, P.E.;STORK, D.G. Pattern Classification. 2.ed. New York, 2001, 654p.

MARQUES, JORGE SALVADOR Reconhecimento de padrões, métodos estatísticos neuronais. 2.ed Lisboa, 2005, 284p.

MEYER, ANDRÉIA DA SILVA Comparação de coeficientes de similaridade usados em análises de agrupamentos com dados de marcadores moleculares dominantes. Dissertação de Mestrado em Agronomia - ESALQ - USP - Piracicaba, São Paulo, 2002.