ISABELA DE CASTRO SANT'ANNA

REDES NEURAIS ARTIFICIAIS NA DISCRIMINAÇÃO DE POPULAÇOES DE RETROCRUZAMENTO COM DIFERENTES GRAUS DE SIMILARIDADE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

VIÇOSA

MINAS GERAIS – BRASIL 2014

Ficha catalográfica preparada pela Seção de Catalogação e Classificação da Biblioteca Central da UFV

T
S231r
2014

Sant'Anna, Isabela de Castro, 1989-
Redes neurais artificiais na discriminação de populações de retrocruzamento com diferentes graus de similaridade / Isabela de Castro Sant'Anna. – Viçosa, MG, 2014.
xiii, 115f. : il. ; 29 cm.

Orientador: Cosme Damião Cruz.
Dissertação (mestrado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Melhoramento genético. 2. Análise discriminante. 3. Inteligência artificial. 4. Redes neurais. I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Programa de Pós-graduação em Genética e Melhoramento. II. Título.

CDD 22. ed. 576.5

ISABELA DE CASTRO SANT'ANNA

REDES NEURAIS ARTIFICIAIS NA DISCRIMINAÇÃO DE POPULAÇOES DE RETROCRUZAMENTO COM DIFERENTES GRAUS DE SIMILARIDADE

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Genética e Melhoramento, para obtenção do título de *Magister Scientiae*.

APROVADA: 26 de fevereiro de 2014

Moyses Nascimento

Leonardo Lopes Bhering
(Coorientador)

Cosme Damião Cruz
(Orientador)

A Deus,

Que sempre ilumina minha vida, me fazendo crescer e vencer as dificuldades encontradas,

OFEREÇO

Aos meus pais Sílvia e Marcus,
Que me criaram para vencer e se realizam com a concretização dos meus sonhos,

DEDICO

"O mal de quase todos nós é que preferimos ser arruinados pelo elogio a ser salvos pela crítica."

Norman Vincent Peale

“Ganhe o respeito dos demais tendo a ousadia de ser você mesmo”.

Dr House

“O que fizemos apenas por nós mesmos morre conosco. O que fizemos pelos outros e pelo mundo permanece e é imortal.”

Albert Pike

AGRADECIMENTOS

A Deus e ao Divino Espírito Santo que iluminam minhas escolhas e apontam meus caminhos.

À Universidade Federal de Viçosa pelas oportunidades em minha vida.

Ao Conselho Nacional do Desenvolvimento Cientifico e Tecnológico (Cnpq), pela concessão da bolsa de estudos.

Ao professor Cosme Damião Cruz pela sua orientação, disponibilidade, amizade, compreensão e pelo exemplo. A você professor, minha consideração, meu respeito e minha gratidão.

Ao professor Leonardo Lopes Bhering, pela sua co-orientação, dedicação e ensinamentos.

Ao professor Pedro Crescêncio Souza Carneiro, pela sua co-orientação, pelo exemplo e pela insistência em conscientização dos alunos sobre seus ensinamentos.

Ao professor Moyses Nascimento, pela amizade e disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

A todos os professores e colegas do Programa de Pós-Graduação em genética e Melhoramento, em especial aqueles que contribuíram de alguma forma para a execução deste trabalho e para minha formação acadêmica.

Aos professores do curso de Ciências Biológicas da UFV que desde cedo exigiram um nível de excelência, um agradecimento especial a Professora Mara e ao professor Everaldo pelo exemplo e contribuições para a minha vida acadêmica.

Aos amigos do laboratório de Biometria pelos momentos compartilhados em especial ao Leo Peixoto pelas contribuições para realização desse trabalho.

Aos funcionários do BIOAGRO pelo auxílio e contribuição indireta para o desenvolvimento desse trabalho.

Aos colegas do BIOINFO, por tornar agradáveis as horas de trabalho, pelos conselhos, pela amizade e aprendizado. Um agradecimento especial a João Filipi, Vinicius, Gabi, Haroldo, Jackeline, Angélica, Danielli, Felipe, Digner, Amálio, Lívia Tomé, Luiza, Caio, Fábio, Marciane, Gyslaine, Daniel e principalmente a Rafael pelas valiosas contribuições para realização desse trabalho.

Aos professores do curso do Programa de Pós Graduação em Genética pelo conhecimento e pelo exemplo a ser seguido.

Aos meus amigos da academia, do salão da Sol, de Ervália, de Viçosa, do COLUNI 04/06, da Biologia, da UFV, pelos infinitos conselhos e torcida em especial a Bianca, Danielle, Fabiana, Renato, Marcos, Josicelli, Kíssia, Júlia, Carla, Aline, Sara, Lidiane, Geverson, Isabelly e Isadora.

Aos meus pais Marcus e Sílvia pelo amor incondicional, apoio e incentivo durante toda a minha vida.

A minha irmã Lívia pela amizade, apoio, e as boas risadas!

Aos meus primos, avos e tios pela constante torcida, em especial a minha tia Cássia que sempre foi um exemplo pra mim e sua família que me acolheu no início da minha caminhada, e aos meus padrinhos: Sandra, Arthur e Elvira e Wilson.

Ao meu avô Quim que sempre me incentivou a ser “grande” e sentiria um imenso orgulho de mim, minha eterna saudade *(in memoriam)*.

Ao meu namorado Henrique por todo companheirismo, amor, apoio por me acalmar e descontrair nas horas difíceis.

Agradeço também a todos que de alguma forma contribuíram e torceram para essa grande conquista!

BIOGRAFIA

ISABELA DE CASTRO SANT'ANNA, filha de Sílvia Maria de Castro Sant'Anna e Marcus Vinícius Silva Sant'Anna, nasceu em Ervália, Minas Gerais, no dia 28 de junho de 1989.

No Município de Ervália, cursou o ensino primário na Escola Municipal do Casca, de 1996 a 1999. Na Escola Estadual de Professor David Procópio cursou parte do ensino fundamental de 2000 a 2001. Em 2002 continuou o ensino fundamental no Colégio Cener até 2003.

Em 2003, iniciou o ensino médio no Colégio de Aplicação-COLUNI (UFV) na cidade de Viçosa-mg que foi concluído em 2006.

Em 2008, iniciou a graduação em Ciências Biológicas pela Universidade Federal de Viçosa (UFV), colando grau em Setembro de 2012.

Em setembro de 2012, iniciou a pós-graduação em Genética e Melhoramento pela Universidade Federal de Viçosa.

SUMÁRIO

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

Capítulo 1

Capítulo 2

RESUMO

SANT'ANNA, Isabela de Castro, M.Sc., Universidade Federal de Viçosa, fevereiro de 2014. **Redes neurais artificiais na discriminação de populações de retrocruzamento com diferentes graus de similaridade.** Orientador: Cosme Damião Cruz. Co-orientadores: Leonardo Lopes Bhering e Pedro Crescêncio Souza Carneiro.

A correta classificação de indivíduos é de extrema importância para fins de preservação da variabilidade genética existente bem como para a maximização dos ganhos. As técnicas de estatística multivariada comumente utilizada nessas situações são as funções discriminantes de Fisher e de Anderson, que permitem alocar um indivíduo inicialmente desconhecido em uma das g populações prováveis ou grupos pré-definidos. Entretanto, para altos níveis de similaridade como é o caso de populações de retrocruzamentos esses métodos tem se mostrado pouco eficientes. Atualmente, muito se fala de um novo paradigma de computação, as redes neurais artificiais, que podem ser utilizadas para resolver diversos problemas da Estatística, como agrupamento de indivíduos similares, previsão de séries temporais e em especial, os problemas de classificação. O objetivo desse trabalho foi realizar um estudo comparativo entre as funções discriminantes de Fisher e de Anderson e as redes neurais artificiais quanto ao número de classificações incorretas de indivíduos sabidamente pertencentes a diferentes populações simuladas de retrocruzamento, com crescentes níveis de similaridade. A dissimilaridade, medida pela distância de Mahalanobis, foi um conceito de fundamental importância na utilização das técnicas de discriminação, pois quantificou o quanto as populações eram divergentes. A obtenção dos dados foi feita através de simulação utilizando o programa computacional Genes. Cada população, gerada por simulação, foi caracterizada por um conjunto de elementos mensurados por características de natureza contínua. Foram geradas considerados 50 locos independentes, cada qual com dois alelos. As relações de parentescos e a estruturação hierárquica foram estabelecidas considerando populações genitoras geneticamente divergentes, híbrido F_1 e cinco gerações de retrocruzamento em relação a cada um dos genitores, permitindo estabelecer parâmetros de eficácia das metodologias testadas. Os dados fenotípicos das populações foram utilizados para estabelecimento da função discriminante de Fisher e Anderson e para o cálculo da taxa de erro aparente (TEA), que mede o

número de classificações incorretas. As estimativas de TEA foram comparadas com as obtida por meio das Redes Neurais Artificiais. As redes neurais artificiais mostraram-se uma técnica promissora no que diz respeito a problemas de classificação, uma vez que apresentaram um número de classificações incorretas de indivíduos menor que os dados obtidos pelas funções discriminantes.

ABSTRACT

SANT'ANNA, Isabela de Castro, M.Sc.,Universidade Federal de Viçosa, february, 2014. **Artificial neural networks to discriminate backcross populations with different degrees of similarity.** Adviser: Cosme Damião Cruz; Co-advisers: Leonardo Lopes Bhering and Pedro Crescêncio Souza Carneiro.

The correct classification of individuals has a top importance for the genetic variability preservation as well as to maximize gains. The multivariate statistical techniques commonly used in these situations are the Fisher and Anderson discriminant functions, allowing to allocate an initially unknown individual in a probably g population or predefined groups. However, for higher levels of similarity such as backcross populations these methods has proved to be inefficient. Currently, much has been Said about a new paradigm of computing, artificial neural networks, which can be used to solve many statistical problems as similar subjects grouping, time-series forecasting and in particular, the classification problems. The aim of this study was to conduct a comparative study between the Fisher and Anderson discriminant functions and artificial neural networks through the number of incorrect classifications of individuals known to belong to different simulated backcross with increasing levels of populations similarity. The dissimilarity, measured by Mahalanobis distance, was a concept of fundamental importance in the use of discrimination techniques, due the quantification of how much populations were divergent. Data collection was done through simulation using the software Genes. Each population generated was characterized by a set of elements measured by characteristics of a continuous distribution. The relations of relatives and hierarchical structuring were established considering genetically divergent populations, F1 hybrid and five generations of backcrossing in relation to each of the relatives, establishing measures of effectiveness of the tested methodologies. The phenotypic data of populations were used to establish the Fisher and Anderson discriminant function and the calculation of the apparent error rate (AER), which measures the number of incorrect classifications. The ERA Estimations were compared with those obtained by means of neural networks. The artificial neural network is shown as a promising technique to solve classification problems, once it had a number of incorrect individuals classifications smaller than the data obtained by the discriminant functions.

1. INTRODUÇÃO

Estudos que visam à discriminação de populações têm sido de grande importância para o desenvolvimento de programas de melhoramento genético e para conservação de biodiversidade. Análises da diversidade genética, por meio de características fenotípicas, têm orientado a escolha de genitores apropriados, em etapas iniciais de programas de melhoramento, levando à otimização dos ganhos seletivos, devido à variabilidade encontrada nos grupos divergentes. Além disso, as análises de diversidade genética têm permitido a quantificação da variabilidade existente e facilitado o gerenciamento dos bancos de germoplasma, poupando tempo e recursos (Cruz et al., 2011).

Dessa forma, a classificação incorreta de populações, que apresentam risco de extinção ou detentores de genes de interesse agronômico, representa um grande prejuízo para programas de conservação da biodiversidade *in* ou *ex-situ*, para programas de preservação ambiental e para definição de cruzamentos em programas de melhoramento quando da utilização de germoplasma "exótico" ou quando da definição de grupos heteróticos. Neste sentido, os estudos de estruturação de populações são de grande valia, uma vez que permitem entendimento das melhores estratégias para incrementar e preservar a diversidade das espécies e ou dos indivíduos dentro das espécies ou populações (Cruz et al., 2011).

Atualmente, há várias metodologias disponíveis para a quantificação e a avaliação da diversidade em estudos populacionais, seja a partir de informações fenotípicas ou a partir de dados genotípicos. No entanto, devido à grande variedade de informações a serem avaliadas e as particularidades de cada material biológico, a escolha e a correta aplicação da metodologia mais adequada são de grande importância para obtenção de resultados confiáveis. Nesse sentido, técnicas de bioinformática e de análise estatística têm se mostrado muito úteis nos programas de melhoramento genético, principalmente nos estudos de diversidade genética (Barbosa et al., 2011).

Comumente, os métodos que se baseiam em análises estatísticas multivariadas tem sido alternativa eficaz nos estudos de diversidade e em problemas classificatórios, entre eles as análises discriminantes são importantes, pois permite estudar as diferenças entre dois ou mais grupos ou populações, em função de um conjunto de informações conhecidas para todos os elementos dos grupos. Estas análises discriminantes são, assim, utilizadas

para classificar um indivíduo ou um grupo de indivíduos em diferentes populações conhecidas e permitem também a elaboração de regras de classificação ou discriminação, que serão então, utilizadas para classificar novos elementos nos grupos já existentes (Cruz et al., 2011, Mingoti 2005).

Entretanto, em diversas situações os pesquisadores apesar de se dispor de dados experimentais adequados para a realização das análises multivariadas os resultados obtidos não são satisfatórios em razão da incapacidade desta técnica detectar diferenças entre populações não linearmente separáveis. Existem situações em que é necessário captar o máximo de informações, além de estatísticas como médias e covariâncias, de amostras do experimento para alcançar um modelo mais ajustado, quando existe um volume muito grande de dados que é afetado por grandes ruídos, e outras em que não foi possível a mensuração de um volume razoável de indivíduos, pois, dependendo das aplicações, a obtenção dos dados pode ter um alto custo. Nesse contexto, a opção por uma nova metodologia que permita fazer generalizações e extrair informações a partir de dados incompletos mostra-se muito promissora para a classificação de populações, como é o caso da utilização de Redes Neurais Artificiais (Cruz et al., 2011).

As RNAs são baseadas em modelos matemáticos inspirados no sistema neural de organismos inteligentes e possuem regras de treinamento em que os pesos de suas conexões são ajustados de acordo com os padrões apresentados, ou seja, adquirem conhecimento através da experiência. Por meio de processos de simulação é possível emular uma situação real, normalmente encontrada em um programa de melhoramento, avaliar e comparar técnicas tradicionalmente utilizadas para discriminação de populações com técnicas fundamentadas em inteligência computacional tal como a estratégia baseada em RNAs.

Pelo exposto foi realizado este estudo, com os seguintes objetivos:

- Estudar a viabilidade da discriminação de populações com diferentes graus de dissimilaridade por Redes Neurais Artificiais.

- Comparar as técnicas de discriminação por redes neurais artificiais com as técnicas de análise discriminante propostas por Anderson e por Fisher, afim de minimizar a probabilidade de má classificação, ou seja, a probabilidade de classificar erroneamente um indivíduo em uma população π_i quando ele realmente pertence a uma população π_j, com $i \neq j$.

- Observar a eficiência das técnicas de análise multivariada na discriminação de populações simuladas de retrocruzamento com diferentes graus de similaridade.

2. REVISÃO DE LITERATURA

2.1. DIVERSIDADE GENÉTICA NO MELHORAMENTO

Os problemas ambientais sempre foram preocupações no melhoramento e, atualmente, as mudanças climáticas e as adversidades geradas pelo aquecimento global, têm sido fator adicional que propulsiona e motiva descobertas na ciência com o objetivo final de permitir, de forma mais acurada, a escolha de genótipos mais produtivos e resistentes e o desenvolvimento de meios que permitam a preservação dos recursos biológicos existentes.

A mensuração e a preservação da biodiversidade existente se mostra essencial para a contínua evolução e, até mesmo, para a sobrevivência de muitas espécies. Dessa maneira, estudos de diversidade genética que permitam a discriminação de populações são essenciais para garantir o direcionamento de recursos e a preservação em bancos de germoplasma, assegurando a disponibilidade de informações sobre os acessos do banco e facilitando o trabalho de seleção de possíveis doadores de genes, podendo também ser usada para eliminação de duplicatas e prevenir a perdas de recursos genéticos (Cruz et al., 2011).

A avaliação de diversidade genética é frequentemente realizada pelos melhoristas e permite a otimização da seleção de genitores de forma que as melhores combinações híbridas sejam preditas. Os esforços são focados em combinações mais promissoras, ou seja, aquelas entre linhagens pertencentes a grupos heteróticos diferentes com bom desempenho e que devam apresentar considerável complementação deforma que a deficiência genética de uma possa ser suplantada pela superioridade de outra(Cruz et al., 2011).

Para isso, a preocupação tem sido em quantificar a dissimilaridade genética existente entre os indivíduos, pois a mesma permite informações sobre o grau de semelhança ou de diferença entre os genótipos, permitindo a formação dos grupos heteróticos pelos métodos de agrupamento, que são

essenciais na escolha de genitores que possuam boa complementaridade gênica (Cruz et al., 2011).

Os coeficientes de similaridade, ou o seu complemento aritmético, têm sido utilizados por diferentes autores com o intuito de quantificar o grau de semelhança (associação entre duas entidades) ou a divergência, a partir das variáveis disponíveis para o estudo. Neste sentido, os métodos de estatística multivariada, que permitem que as diversas características possam ser analisadas conjuntamente, têm contribuído efetivamente na identificação dos genótipos mais promissores nos programas de melhoramento genético de várias culturas (Oliveira et al., 2007).

Estudos de diversidade genética propiciam ao pesquisador a capacidade de avaliar padrões de agrupamento, formular e testar hipóteses sobre a dissimilaridade ou diversidade obtida. Entretanto, devido ao número de estimativas de dissimilaridade ser relativamente grande torna-se impraticável o reconhecimento de grupos homogêneos apenas pelo exame visual das estimativas (Cruz et al., 2011). Portanto, faz se necessária a utilização de métodos de agrupamento ou de projeções de distância em gráficos bi ou tridimensionais em que cada coordenada é obtida a partir da medida de dissimilaridade escolhida (Cruz & Carneiro, 2006).

Os métodos hierárquicos constituem uma categoria de metodologias da estatística multivariada que podem ser utilizados para agrupamento de genótipos por um processo iterativo que culmina no estabelecimento de um dendrograma ou o diagrama em árvore. Nesse caso, não há preocupação com o número ótimo de grupos, uma vez que o interesse maior está na “árvore” e nas ramificações que são obtidas. As delimitações podem ser estabelecidas por um exame visual do dendrograma, em que se avaliam pontos de alta mudança de nível, tomando-os em geral como delimitadores do número de genótipos para determinado grupo (Cruz et al., 2011).

Diversos estudos utilizam análises de agrupamento na visualização e interpretação da diversidade genética, com base em caracteres morfológicos e agronômicos em plantas como a gabiroba (Rezende et al., 2009), coentro (Melo et al., 2011), a batata doce (Martins et al., 2013), e o alho (Viana, 2013).

Populações de retrocruzamento

No melhoramento genético, o método de retrocruzamento constitui uma estratégia bastante utilizada para transferir um ou poucos alelos controladores

de determinada característica. As populações de retrocruzamento são confeccionadas com o intuito de melhorar a expressão fenotípica de uma característica de um dado cultivar, geralmente de bom desempenho, mas com pequenas deficiências proporcionada pela falta de um ou poucos alelos favoráveis que são, geralmente, encontrados em um fonte denominada genitor não-recorrente (Borém, 2009).

O termo retrocruzamento se refere aos repetidos cruzamentos dos indivíduos da população segregante com uma das linhagens genitoras. O genitor que contém o alelo que confere o fenótipo desejado é denominado de não recorrente ou doador, ou seja, é utilizado apenas uma vez nos cruzamentos. No retrocruzamento a geração F_1 é cruzada com um dos genitores. O genitor que é submetido aos sucessivos cruzamentos com os indivíduos da população segregante é denominado de recorrente. O genitor recorrente deve ser cuidadosamente escolhido, pois deve apresentar bons atributos para todas as características, exceto aquelas que serão doadas pelo não recorrente (Borém, 2009).

O método de retrocruzamento é muito indicado para transferência de pequenas proporções genômicas de genótipos não adaptados ou de espécies selvagens para genótipos elite (Lorencetti et al., 2006). Também tem sido utilizado na adaptação de germoplasma exótico (Nass, 2001). Entretanto, o método se torna complexo para transferência de genes controladores de características quantitativas (Borém, 2009).

A similaridade genética em populações avançadas por retrocruzamentos e seu genitor recorrente aumenta de forma previsível dentro do conceito meiótico, fundamentado da contribuição equitativas dos gametas masculinos e femininos, pois proporção de genes do genitor doador é reduzida à metade após cada geração de retrocruzamentos. Consequentemente, a recuperação do genoma do genitor recorrente aumenta proporcionalmente a cada novo retrocruzamento. Espera-se, considerando a contribuição gamética equitativa, que a recuperação do genitor recorrente no RC_x ocorra na proporção de $(2^{x+1} - 1) / 2^{x+1}$, em que x é número de retrocruzamentos com o genitor recorrente (Cruz, 2005). Assim, após cinco gerações de retrocruzamento a proporção esperada de similaridade entre a RC_5 e o genitor recorrente, seja de aproximadamente 98,44% (Borém, 2009). Esse grau de similaridade torna-se bastante difícil a discriminação das diferentes populações de retrocruzamento

entre si e com o respectivo genitor recorrente que, na verdade, é o propósito da estratégia de melhoramento por retrocruzamento.

Populações de retrocruzamentos podem ser vistas, ou utilizadas, em outro contexto em análises biométricas. Assim, quando se dispões de um conjunto de populações derivadas por retrocruzamentos a relação de similaridade pode ser estabelecida previamente a partir de modelos que levam em consideração a contribuição meiótica em cada geração. Modelos biométricos, que buscam identificar o padrão de dissimilaridade entre populações, podem ser utilizados e suas eficácias podem ser quantificadas comparando a medida biométrica da diferenciação entre pares de populações e a diferenciação esperada tendo em vista o número e o tipo de cruzamento realizado.

Existem vários métodos de agrupamento que se baseiam na dissimilaridade entre as populações que podem realizar essa separação como os agrupamentos hierárquicos, de otimização e de projeção no plano bidimensional. Distinguir biometricamente a diferenciação entre quaisquer pares de populações não tem sido tarefa trivial, principalmente quando estas populações apresentam baixo nível de diferenciação. Assim, é necessário avaliar a potencialidade da técnica para este propósito considerando populações com maior e menor similaridade. Na prática podemos ter informações a priori sobre a dissimilaridade entre populações tomando por base sua origem, adaptação, histórico de uso, dentre outros. Entretanto, uma medida exata é difícil de ser estabelecida, por isto o recurso biométrico de se estabelecer conjunto de populações com graus variados e conhecidos de dissimilaridade é importante e facilmente estabelecido em delineamento genético envolvendo genitores, população híbrida e gerações de retrocruzamento. Independente do nível de similaridade, o mais apropriado para estudar o padrão de dissimilaridade é a utilização de métodos de estatísticos multivariados coma a Análise Discriminante de Fisher e Anderson, ou ainda a utilização de Redes Neurais Artificiais.

2.2. ANÁLISES DISCRIMINANTES

A estatística multivariada é um ramo da Estatística que se preocupa com o estudo, a relação e a interpretação de dados. Ela se baseia na mensuração de diversas variáveis simultâneas, o que a torna essencial para os processos

de experimentação agronômica, pois a análise isolada de cada uma das variáveis, pode não conseguir caracterizar, de maneira adequada, a variabilidade dos dados (Cruz et al., 2011).

Uma das técnicas de análise multivariada que permite alocar um novo indivíduo a uma das varias populações distintas, previamente conhecidas, é a análise discriminante. Esta consiste na obtenção de funções que permitam diferenciar um determinado "indivíduo", com base em medidas de várias características, em uma entre várias populações distintas, buscando minimizar a probabilidade de uma classificação errônea. A utilização dessa técnica é bastante frequente, uma vez que é simples e possui alta eficiência para uma ampla variedade de estruturas populacionais. Ainda, de acordo com Cruz et al., (2012), a sua aplicação busca minimizar a probabilidade de uma classificação errônea, isto é, de se classificar o referido "indivíduo" em uma população, quando este na realidade pertence a outra.

A análise discriminante é utilizada para classificar um determinado elemento (E), num determinado grupo de variáveis, entre os diversos grupos existentes π_1, π_2, π_3,.... π_i. Para tal é necessário que o elemento (E) a ser classificado pertença realmente a um dos π_i grupos, e que sejam conhecidas as características dos elementos dos diversos grupos. Essas características são especificadas a partir de v variáveis aleatórias (X_1 , X_2 , X_3 ,..., X_v). No processo de classificação consideram-se os custos decorrentes de eventuais erros de classificação, bem como as probabilidades "a priori" de que o elemento pertença a cada um dos grupos.

Contudo, a mesma função discriminante que separa objetos pode também servir para alocar, e o inverso, regras que alocam objetos podem ser usadas para separar. Normalmente, discriminação e classificação se sobrepõem na análise, e a distinção entre separação e alocação é confusa. Uma boa classificação deve resultar em pequenos erros, isto é, deve haver pouca probabilidade de má classificação. Segundo Johnson & Wichern (1999) para que isso ocorra a regra de classificação deve considerar as probabilidades a priori e os custos de má classificação. Outro fator que uma regra de classificação deve considerar é se as variâncias das populações são homogêneas ou não.

Análise Discriminante de Fisher

A Função Discriminante Linear de Fisher, uma combinação linear das características observadas que apresenta melhor poder de discriminação entre os grupos, constitui a base de todo o estudo na análise discriminante. Esta função tem a propriedade de minimizar a probabilidade de má classificação, quando as populações apresentam média e variância conhecidas. Contudo, tal situação pode não ocorrer na prática, necessitando-se, portanto de estimativas e métodos de estimação dessas probabilidades ótimas (CRUZ et al., 2012).

Considerando duas populações (π_i e π_i') com vetor de médias v-variado μ_i e $\mu_{i'}$ e matriz de covariâncias comuns $\sum$, de ordem v, define-se a função discriminante linear de Fisher pela expressão 1.

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = \left(\mu_i - \mu_{i'}\right)' \Sigma^{-1}\tilde{x} \qquad (1)$$

Assim, a função discriminante $D_{ii'}(\tilde{x})$ é uma combinação linear do conjunto de caracteres que possibilita alocar um determinado indivíduo, com vetor de observações $\tilde{x}$, em uma população π_i, ou π_i', com máxima probabilidade de acerto. Define-se também o ponto médio entre duas populações π_i e π_i' pelo valor m, expresso pela equação 2 ou 3.

$$m_{ii'} = \frac{1}{2}\left(\mu_i - \mu_{i'}\right)' \Sigma^{-1}\left(\mu_i + \mu_{i'}\right) = \alpha' u = \frac{1}{2}\left(\alpha'\mu_1 + \alpha'\mu_2\right) \qquad (2)$$

ou

$$m_{ii'} = \frac{1}{2}\left[D(\mu_1) + D(\mu_2)\right] \qquad (3)$$

Com a função discriminante estimada, adota-se a regra de classificação conforme as expressões 4 e 5.

- Aloca-se $\tilde{x}$ em π_i se:

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = \left(\mu_1 - \mu_2\right)' \Sigma^{-1}\tilde{x} \geq m_{ii'} \qquad (4)$$

- Aloca-se $\tilde{x}$ em $\pi_{i'}$ se:

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = \left(\mu_1 - \mu_2\right)' \Sigma^{-1}\tilde{x} < m_{ii'} \qquad (5)$$

A ideia básica da Análise Discriminante de Fisher foi transformar observações multivariadas X em observações univariadas Y derivadas das populações π_1 e π_2 em que estas apresentassem o maior grau de separação possível. Fisher sugere tomar combinações lineares de X para criar as

combinações Y's, pois tais combinações podem ser facilmente manipuladas.

Análise discriminante de Anderson

De acordo com Anderson (1958), quando se dispõe de várias populações e se deseja alocar um novo indivíduo a cada uma delas, um procedimento importante é que os indivíduos estejam divididos em populações distintas, e que seja estabelecida as probabilidades "a priori" para as várias populações, pois há casos nos quais a probabilidade de um determinado indivíduo pertencer a uma dada população pode ser muito distinta da dele pertencer a outra, de forma que a experiência do pesquisador torna-se de extrema importância.

Com estas informações, são geradas funções, que são combinações lineares das características avaliadas, e que tem por finalidade obter a melhor discriminação entre os indivíduos, alocando-os em suas devidas populações.

Além disso, as funções permitem também a classificação de novos genótipos, de comportamento desconhecido, nas populações já conhecidas. A eficácia das variáveis utilizadas em promover a discriminação também é avaliada, permitindo conhecer a adequação da função estimada.

Para o estabelecimento da função discriminante de Anderson, considera-se que, para uma população π_j (j = 1, 2, ..., g), o vetor da variável aleatória $\tilde{x}$ tem distribuição $N_v(\mu_j, \Sigma)$, com a seguinte função densidade de probabilidade (equação 6).

$$f_j(\tilde{x}) = |2\pi\Sigma|^{-1/2} \; e^{-\frac{1}{2}\left[\left(\tilde{x}-\mu_j\right)'\Sigma^{-1}\left(\tilde{x}-\mu_j\right)\right]} \qquad (6)$$

Também é admitido que a probabilidade de uma observação pertencer a uma determinada população é p_j ($\sum_{j=1}^{g} p_j = 1$), conhecida *a priori*. Assim, pode-se estabelecer a função discriminante, dada pela probabilidade de $\tilde{X}$ pertencer a π_j, por meio do logaritmo da função densidade de probabilidade de $\tilde{x}$ e da probabilidade a *priori*, de forma que se tenha:

$$D_j(\tilde{x}) = -\frac{1}{2}\left[\ln(2\pi) + \ln|\Sigma_j|\right] - \frac{1}{2}\left[\left(\tilde{x}-\mu_j\right)'\Sigma_j^{-1}\left(\tilde{x}-\mu_j\right)\right] + \ln(p_j) \qquad (7)$$

ou, de forma simplificada:

$$D_j(\tilde{x}) = \ln(p_j) + \left(\tilde{x} - \frac{1}{2}\mu_j\right)' \Sigma^{-1}\mu_j \quad (8)$$

Com base nas médias de cada população e na matriz de variância e covariância entre as médias das populações, obtiveram-se as respectivas funções discriminantes. Cada função é uma combinação linear das v características avaliadas, existindo tantas funções quanto for o número de populações avaliadas. A partir das funções discriminantes, estima-se, para cada genótipo, o valor discriminante, permitindo, classificar o i-ésimo indivíduo, com vetor de média $\tilde{x}_i$, na população π_j se e somente se $D_j(\tilde{x}_i)$ for o maior entre os elementos do conjunto $\{D_1(\tilde{x}_i), D_2(\tilde{x}_i),..., D_g(\tilde{x}_i)\}$.

2.3. REDES NEURAIS ARTIFICIAIS (RNAs)

A inteligência artificial tem permitido uma nova abordagem no processo de tomada de decisão em diversas áreas da ciência com grande potencial no melhoramento genético animal e vegetal (Ventura et al., 2013). Um novo paradigma pode ser empregado no melhoramento genético para fins de seleção que não envolve modelagem estocástica, mas princípios de aprendizado em abordagem de inteligência computacional.

Em diversas situações o pesquisador dispõe de dados experimentais apropriados para realização das analises biométricas necessárias para avaliação dos experimentos com objetivo de discriminar indivíduos ou populações, porém as análises biométricas nem sempre são capazes de produzir resultados satisfatórios, pois o modelo adotado e as estatísticas requeridas (médias, variâncias e covariâncias) podem ser insuficientes para descrever e caracterizar convenientemente as particularidades de cada população. Neste contexto, a realização de análises por meio de métodos computacionais que sejam capazes de aprendizagem e generalização a partir de toda a informação disponível, sendo tolerantes a ruídos, representa um grande avanço para os estudos envolvendo procedimentos estatísticos e para o melhoramento genético.

Esse novo paradigma trata da inteligência artificial que vem sendo cada vez mais utilizada e tem permitido, com o avanço da tecnologia e do entendimento da neurociência, a criação de modelos de neurônios artificiais muito próximo dos neurônios biológicos que conseguem se conectar formando

as Redes Neurais Artificiais. Essas conexões as tornam capazes de análises de diversas situações, aprendizagem, reconhecimento de padrões e generalização (Braga et al., 2007).

O desempenho da rede é determinado pelas conexões entre os seus elementos e por isso, pode-se treinar uma rede neural para executar uma função particular ajustando-se os valores das conexões entre os elementos (Haykin, 2001). Esse processo permite uma adaptação da RNA às particularidades de um problema, que a torna capaz de generalizações, permitindo as mesmas respostas para estímulos similares.

As RNAs caracterizam-se pela sua arquitetura e pelo ajustamento de seus pesos às conexões durante o processo de aprendizado. A arquitetura de uma rede neural é definida pelo número de camadas (camada única ou múltiplas camadas), pelas conexões entre camadas, pelo número de neurônios em cada camada, pelo tipo de conexão entre eles (feedforward ou feedback) e pelo algoritmo de aprendizado (Haykin, 2001).

O número de camadas é um fator crucial na determinação da capacidade da rede de solucionar problemas. Geralmente, nas redes, as camadas são classificadas em três tipos: Camada de Entrada: onde os padrões são apresentados à rede; Camadas Intermediárias ou Ocultas: destinadas a realizar grande parte do processamento dos dados e atribuir pesos através das conexões ponderadas, ou seja, são extratoras das características; Camada de Saída: onde o resultado final é concluído (saída de rede) e apresentado (saída desejada).

Na Figura 1 está ilustrada uma arquitetura de rede neural na qual podem ser identificadas as camadas de entrada, as camadas intermediárias e a camada de saída que deve retornar valores preditos para as variáveis de interesse, cuja resposta pode, de forma semelhante aos dados de entrada, ser uni ou multivariado. Essa figura representa o primeiro modelo de redes de múltipla camada, o Perceptron multicamadas que surgiu e tornou as redes capazes de resolver problemas não linearmente separáveis.

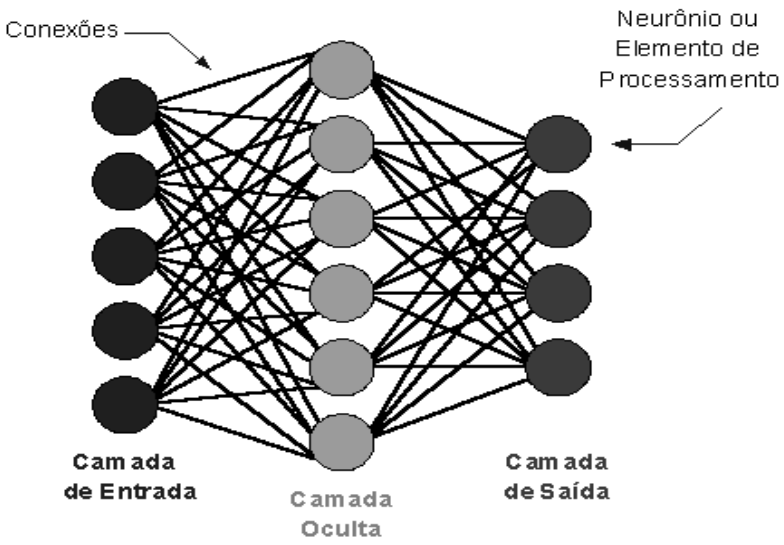


Figura 1. Representação dos três tipos de camadas existentes em redes neurais de múltiplas camadas, característica do modelo Perceptron Múltiplas Camadas (MultiLayer Perceptron - MLP).

As RNAs vêm sendo desenvolvidas há aproximadamente cinco décadas, sendo, atualmente, muito utilizadas em diversas áreas da ciência. Foram inspiradas pela percepção de que o cérebro humano processa as informações de uma maneira mais rápida e distinta dos computadores digitais convencionais devido a sua complexidade de operação e ao fato de se processar em paralelo (Haykin, 2001).

Devido à complexidade com que opera ainda se faz necessário muitos estudos para o esclarecimento do seu funcionamento e aumento da sua confiabilidade. Seu desenvolvimento é marcado por muitas descobertas e evolução de modelos que foram abandonados ou aprimorados. Consequentemente é uma área de estudo com muitos problemas abertos à pesquisa teórica.

Em 1943, o primeiro modelo artificial de um neurônio biológico foi apresentado pelo neuroanatomista e psiquiatra Warren McCulloch e do matemático Walter Pitts (McCulloch&Pitts, 1943). Eles mostraram que uma coleção de neurônios era capaz de calcular certas funções lógicas.

Em 1949, Hebb propôs uma regra de aprendizagem básica que se baseava em sinapses com pesos diferentes.

Em 1959, Rosenblatt desenvolveu o 1º modelo de Rede Neural, o Perceptron, reunindo as idéias de Hebb, McCulloch e Pitts. O Perceptron

poderia aprender funções lógicas, pois apresentava seus neurônios arranjados em uma rede com uma topologia particular.

Em 1962, Widrow desenvolveu um tipo diferente de processador para Redes Neurais, denominado ADALINE, o qual dispunha de uma poderosa estratégia de aprendizado.

Em 1969 Minsky e Papert expuseram as limitações do Perceptron o que levou ao abandono dos investimentos das pesquisas em Redes Neurais.

Em 1974, Werbbs lançou as bases do algoritmo Back-Propagation, que permitiu que Redes Neurais com múltiplas camadas apresentassem capacidade de aprendizado.

Alguns anos depois o modelo Perceptron Múltiplas Camadas (MultiLayerPerceptron - MLP) e o algoritmo backpropagation tornaram as redes neurais artificiais uma metodologia amplamente utilizada em várias áreas da ciência (Braga et al., 2007).

Fundamentos Biológicos das Redes Neurais Artificiais

Uma característica única do sistema nervoso é a sua plasticidade, característica que tem servido de motivação para estudos de inteligência artificial. Segundo DeGroot (1994), "a plasticidade neural é a propriedade do sistema nervoso que permite o desenvolvimento de alterações estruturais em resposta à experiência, e como adaptação a condições mutantes e a estímulos repetidos".

A cada nova experiência do indivíduo, ocorre rearranjos nas redes de neurônios, um conjunto de sinapses são reforçadas e múltiplas possibilidades de respostas ao ambiente tornam-se possíveis. Portanto, "o mapa cortical de um individuo está sujeito a constantes modificações com base no uso ou atividade de seus caminhos sensoriais periféricos" (Kandel 1991).

As RNAs foram inspiradas no Sistema Nervoso Biológico. Na Figura 2 está representado a estrutura fundamental do sistema nervoso que é o neurônio constituído pelo corpo celular, axônio e os dendritos. Os dendritos são os elementos de entrada, que conduzem os sinais das extremidades para o corpo celular. O corpo celular combina os sinais dos dendritos formando um sinal excitante ou inibitório. Esses sinais são transmitidos pelos axônios. As

extremidades do axônio de um neurônio são conectadas com os dendritos de outros neurônios através das sinapses, formando uma rede (Guyton, 1981).

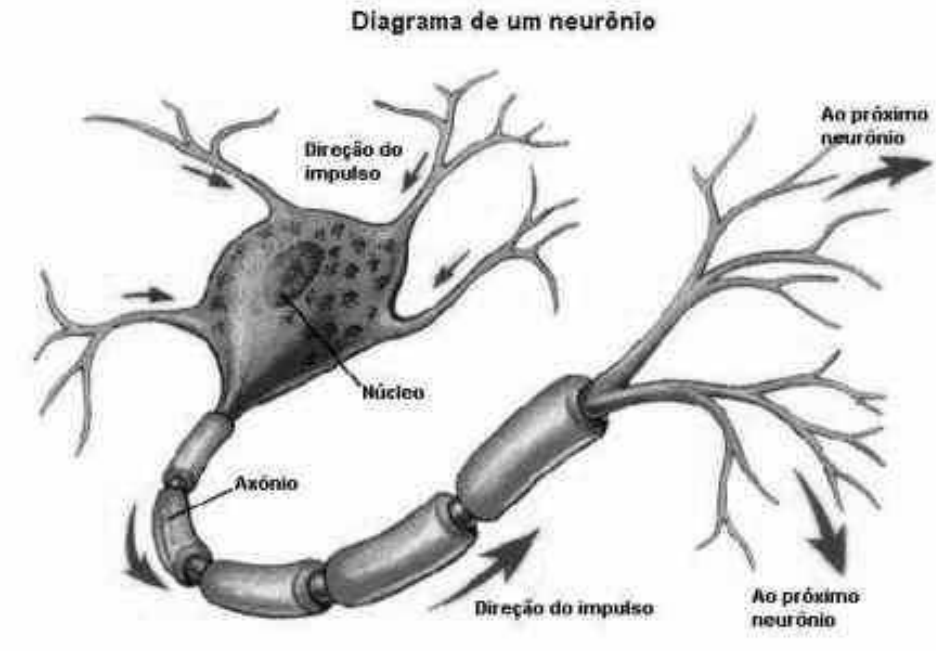


Figura 2: Representação de neurônios biológicos (Adaptada de http://cdoze.wikispaces.com/file/detail/sistema-nervoso-60.jpg. Detalhe para as parte que compõem o neurônio e o sentido da sinapse.

Os Neurônios Artificiais

Durante o processo de treinamentos das RNAs os pesos entre as conexões dos neurônios artificiais são os parâmetros ajustáveis que variam à medida que o conjunto de treinamento é apresentado à rede. Dessa maneira, são responsáveis pelo conhecimento adquirido. Na Figura 3 é representado um modelo de neurônio artificial que representa primeiro modelo artificial de um neurônio biológico apresentado em1943 por Warren McCulloch e Walter Pitts (McCulloch & Pitts, 1943).

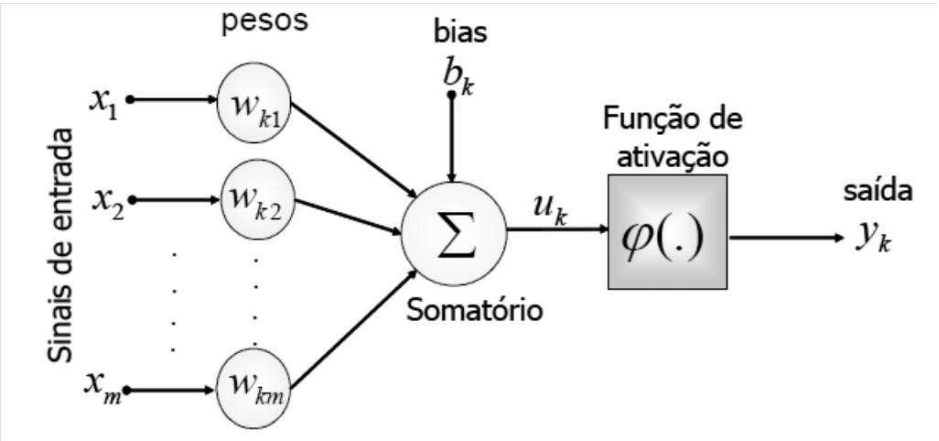


Figura 3. Modelo não linear de um neurônio artificial (Adaptado de Haykin, 2001). Onde x_1, x_2, ..., x_m são as entradas da rede; w_{k1}, w_{k2}, ..., w_{km} são

os pesos, ou pesos sinápticos, associados a cada entrada; b_k é o termo bias; u_k é a combinação linear dos sinais de entrada; φ(.) é a função de ativação e y_k é a saída do neurônio.

Assim, o processo de aprendizado supervisionado em uma RNA com pesos, resulta em sucessivos ajustes dos pesos sinápticos, de tal forma que a saída da rede seja a mais próxima possível da resposta desejada. Tipicamente, a ordem de amplitude normalizada da saída do neurônio está no intervalo [0, 1] ou alternativamente [−1, 1]. O modelo neural também inclui um termo chamado de “bias”, aplicado externamente, simbolizado por b_k. O b_k tem o efeito do acréscimo ou decréscimo da função de ativação na entrada da rede, dependendo se é positiva ou negativa, respectivamente (Peixoto, 2012).

A transmissão de impulso nervoso de um neurônio biológico acontece quando a soma dos impulsos que ele recebe ultrapassa o seu limiar de excitação *(threshold)*. No neurônio artificial de McCulloch e Pitts (MCP) a ativação do neurônio é obtida através de uma "função de ativação", que ativa ou não a saída, dependendo do valor das somas ponderadas de suas entradas (Braga et al., 2007).

Funções de ativação

Assim como nos neurônios biológicos os neurônios artificiais precisam receber estímulos que excedam um valor de limiar para que o impulso seja transmitido. A porta de limiar compara a soma ponderada das entradas com um valor limite. Caso a soma exceda o limiar, a saída é ativada, caso isso não ocorra ela permanece desativada (Braga et al., 2007).

As funções de ativação fornecem o valor da saída de um neurônio a partir das somas ponderadas recebidas pelo neurônio, e deve ser escolhida de acordo com o problema em estudo. Elas limitam a saída dos neurônios nos intervalos de [0,1] ou [-1,1]. As principais funções de ativação utilizadas são: função Limiar (Degrau), função sigmoidal e função tangente hiperbólica, que serão ilustradas nas figuras 4, 5 e 6 respectivamente.

a. Função Limiar (Degrau)

A função de limiar utilizada no modelo de McCulloch e Pits modela a característica "tudo-ou-nada" deste neurônio. É expressa da na equação 9.

$$f(v) = \begin{cases} 1, \text{ se } v \geq 0; \\ 0, \text{ se } v \leq 0; \end{cases} \quad (9)$$

Nos neurônios construídos com essa função, a saída *y* será igual a 0, caso o valor de ativação *v* seja negativo e 1 nos casos em que o valor de ativação seja positivo.

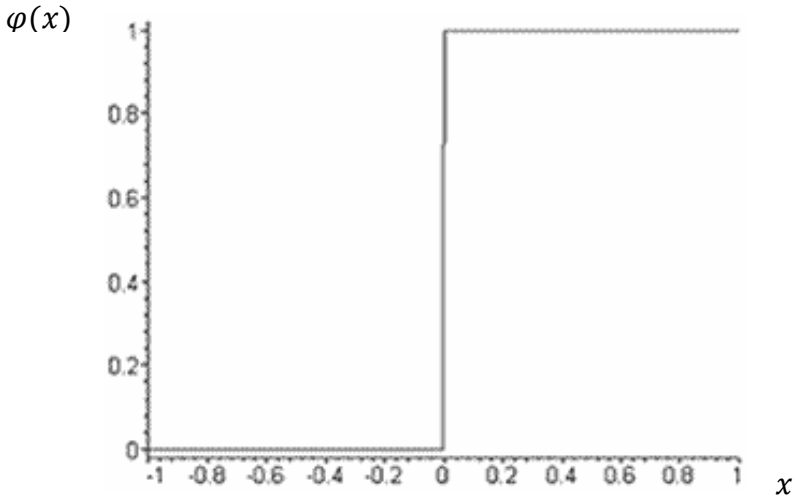


Figura 4: Gráfico da Função Limiar.

b. Função Sigmoidal

Segundo Haykin(2001) A função *sigmoidal* (logsig) é a função de ativação mais utilizada na construção das Redes Neurais Artificiais. Esta função, ao contrário da função limiar, pode assumir todos os valores entre 0 e 1. Por definição, a função sigmoidal é monótona crescente com propriedades assintóticas e de suavidade, e como visto na Figura 5, apresenta gráfico na forma de "s".

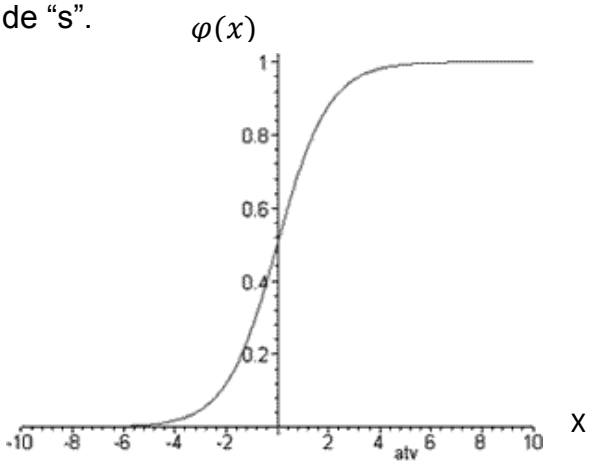


FIGURA 5. Gráfico da função sigmoidal.

A função logística, definida na Equação 10 constitui exemplo de função sigmoidal.

$$\varphi(x) = \frac{1}{1+ e^{-ax}} \tag{10}$$

em que a é o parâmetro de inclinação da função sigmoide. A variação do valor do parâmetro a proporciona funções sigmoides com diferentes inclinações. Quando este parâmetro se aproxima do infinito, a função sigmoide se torna uma função limiar, podendo, no entanto, assumir um intervalo contínuo de valores entre 0 e 1, ao contrário da função limiar que somente assume valor 0 ou 1. Pode ser necessário, porém, que a função de ativação assuma valores entre 1 e -1. Esta característica traz benefícios analíticos. Para obtermos tal intervalo de valores utilizamos as Funções Signum, no caso da Função de Limiar, e a Função Tangente Hiperbólica, no caso da Função Sigmoide.

c. Função Tangente Hiperbólica

Segundo Haykin (2001), em algumas situações se deseja uma função de ativação que se estenda de -1 a +1, assumindo uma forma anti-simétrica em relação à origem. Neste caso, utiliza-se uma forma correspondente à logsig denominada de função tangente hiperbólica (tansig) – figura 6 – definida pela equação 11:

$$\varphi(x) = \tanh\left(\frac{x}{2}\right) = \frac{1- e^{-x}}{1+ e^{x}} \tag{11}$$

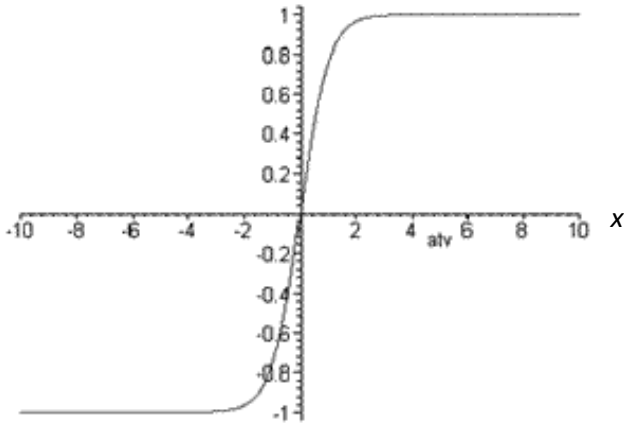


FIGURA 6. Gráfico da função tangente hiperbólica.

Topologias das Redes Neurais Artificiais

A escolha da arquitetura utilizada pelas redes neurais, bem como os valores iniciais dos pesos utilizados no processo de treinamento ainda é feita de maneira empírica, e determina o tempo necessário para o treinamento (Silva et al., 2008). Normalmente, os valores iniciais dos pesos da rede são números aleatórios uniformemente distribuídos, em um intervalo definido. A escolha errada destes pesos pode levar a uma saturação prematura. Nguyen e Widrow (1990) encontraram uma função que pode ser utilizada para determinar valores iniciais melhores que valores puramente aleatórios.

O número de neurônios em cada camada de uma RNA determina a sua capacidade de generalização, e sua precisão na resolução do problema. A determinação do número de neurônios depende da complexidade do problema, do número de exemplos de treinamento, da quantidade de ruído presente nos exemplos, da complexidade da função a ser aprendida pela rede, e da distribuição estatística dos dados de treinamento (Tafner, 1995). As redes neurais artificiais, com um número suficiente de neurônios, podem aproximar qualquer função linearmente contínua. O poder de aproximação dependerá da arquitetura da rede e do número de neurônios em cada camada oculta (Braga et al., 2007; Haykin, 2001).

Segundo Silva et al. (2008) na literatura, não existem estudos que indiquem como deve ser a distribuição de neurônios nas camadas. Entretanto, faz-se necessário uma escolha criteriosa para não utilizar um número de unidades maior nem menor que o suficiente. Já que, um número alto de unidades leva a memorização dos dados de treinamento, e torna a rede incapaz de generalizações e, portanto, incapaz de reconhecer padrões, ou seja extrapolar suas conclusões para dados não treinados (overfitting). Um número de neurônios inferior ao mínimo necessário aumenta o tempo de aprendizagem da rede, que pode não alcançar os pesos adequados, ou seja, a rede pode não encontrar a solução mais adequada. Sendo assim, faz se necessário a reconfiguração da mesma.

Dentre as vantagens da utilização da RNAs, ressaltam-se duas: primeiramente, a sua estrutura não linear, capaz de “captar complexas características entre o conjunto de dados de entrada” (Galvão et al., 1999); e em segundo lugar, a sua capacidade de não requerer informação detalhada

sobre os processos físicos do sistema a ser modelado (Sudheer et al., 2003). Talvez por esses motivos, a utilização das redes neurais tem se mostrado mais promissora devido a possibilidade de um desempenho superior aos modelos convencionais utilizados na solução de problemas (Braga et al., 2007).

Paradigmas de Aprendizagem

O processo de aprendizagem consiste na etapa por meio da qual os parâmetros livres de uma rede são adaptados através dos estímulos fornecidos pelo ambiente de treinamento e se torna capaz de fornecer uma solução generalizada para uma classe de problemas (Haykin, 2001).

As redes neurais se baseiam nos dados para extrair sua capacidade de generalização. Portanto, um treinamento adequado é a etapa mais importante do processamento dos dados. Os pesos atribuídos às conexões entre os neurônios detêm a aprendizagem e a capacidade de melhorar o desempenho.

Segundo Xavier (2003) é necessário que 50 a 90% do total de dados do experimento seja utilizado para o treinamento da rede neural de forma que a ela "aprenda" as regras e não "decore" exemplos. O restante dos dados é apresentado à rede neural em uma etapa posterior de validação do experimento, na qual a rede treinada é utilizada em dados ainda não utilizados a fim de que ela possa classificar corretamente os mesmos.

De acordo com Braga et al. (2007), o conceito de aprendizado está relacionado ainda à melhoria do desempenho da rede segundo algum critério preestabelecido. O erro quadrático médio da resposta da rede em relação ao conjunto de dados fornecido pelo ambiente, por exemplo, é utilizado como critério de desempenho pelos algoritmos de correção de erros.

O processo de aprendizado se dá com a atualização dos pesos sinápticos a cada iteração da rede, conforme a Equação 13.

$$w_{(t+1)} = w_{(t)} + \Delta w_{(t)} \quad (13)$$

em que $w_{(t)}$e $w_{(t+1)}$representam os valores dos pesos nos instantes t e $t+1$, respectivamente, e $\Delta w_{(t)}$ é o ajuste aplicado ao peso a cada iteração.

Segundo Haykin (2001) o tipo de aprendizagem das RNA é determinado pela forma através da qual é efetuada a mudança nos parâmetros. O aprendizado em redes neurais pode ser classificado, de acordo com presença

ou ausência de realimentação explícita do mundo exterior, em supervisionado e não supervisionado (Barreto, 2004).

No aprendizado supervisionado a rede utiliza um agente externo que assinala acertos e erros de acordo com o padrão de entrada e a resposta desejada. Já no aprendizado não supervisionado (auto-organização) não existe um agente externo indicando a resposta desejada para os padrões de entrada, utiliza-se, entretanto, exemplos semelhantes para que a rede calcule as correlações entre eles e responda corretamente.

a. Aprendizado Não Supervisionado

Nesse tipo de aprendizagem, não existe um agente externo (professor ou supervisor) para "acompanhar" o processo de aprendizado. A rede é treinada por meio de excitações ou padrões de entrada para, arbitrariamente, organizar os padrões em categorias. Para uma entrada aplicada a rede, é fornecida uma resposta indicando a classe a qual a entrada pertence. Se o padrão de entrada não corresponde às classes existentes, uma nova classe é gerada. Dessa forma, o aprendizado não supervisionado se aplica às classes de problemas em que se deseja formar grupos de populações semelhantes.

b. Aprendizado Supervisionado

A aprendizagem supervisionada (aprendizagem com professor) é caracterizada pela utilização de um agente externo que indica à rede a resposta desejada para o padrão de entrada. A rede neural é treinada por meio da apresentação de pares de entradas e saídas. Para cada entrada a rede produz uma resposta na saída, que é comparada com a resposta desejada. Por meio da análise de erros, realiza-se o ajuste dos pesos sinápticos, sendo este processo normalmente utilizado para redes de retropropagação. O aprendizado supervisionado se aplica às classes de problemas em que se deseja mapear padrões de entrada e saída, como no caso de problemas de classificação. Os algoritmos mais conhecidos para aprendizado supervisionado são a regra delta (Windrow, 1960) e o algoritmo back-propagation (Rumelhart, 1986), sua generalização para redes múltiplas camadas. Na Figura 8, retirado de Braga et al. (2007), está apresentado o esquema do aprendizado supervisionado.

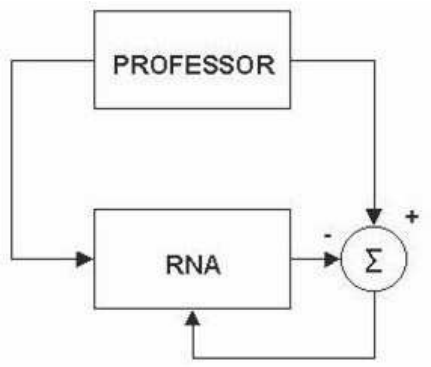


Figura 7: Aprendizado supervisionado, adaptado de Braga et al., 2007.

c. Algoritmo da Retropropagação ("Backpropagation")

O algoritmo de aprendizado é um conjunto de regras bem definido para o treinamento da rede na solução de um problema. Existem muitos tipos de algoritmos de aprendizado específicos para determinados modelos de redes neurais, estes algoritmos diferem entre si principalmente pelo modo como os pesos são modificados. Nas redes neurais de múltiplas camadas o algoritmo mais comum é o de retropropagação ("backpropagation") que poder ser visualizado na Figura 9.

O algoritmo de retropropagação se baseia no aprendizado supervisionado por correção de erros. Basicamente, a aprendizagem por retropropagação de erro consiste em dois passos através das diferentes camadas da rede: um passo para frente, Feed-forward (a propagação), e um passo para trás, Feed-backward (retropropagação) (Haykin, 2001). Primeiro, um padrão é apresentado à camada de entrada da rede. A resposta de uma unidade é propagada como entrada para as unidades na camada seguinte, até a camada de saída, onde é obtida a resposta da rede e o erro é calculado. No segundo passo, o erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados conforme o erro é retropropagado (Pereira, 2009).

O erro de uma rede neural pode ser calculado como a diferença entre a saída real gerada pela rede e a saída desejada fornecida. Os erros são calculados sucessivamente até que sejam minimizados a um valor satisfatório, definido *a priori*. Sendo assim, pode ser visualizada uma curva de erros, a qual está diretamente relacionada à natureza do modelo de neurônio utilizado. Nem sempre é possível alcançar o menor valor de erro ou o mínimo global atingindo

o que chamamos de mínimo local. Caso este erro alcançado seja desfavorável, é necessário recomeçar processo de aprendizado.

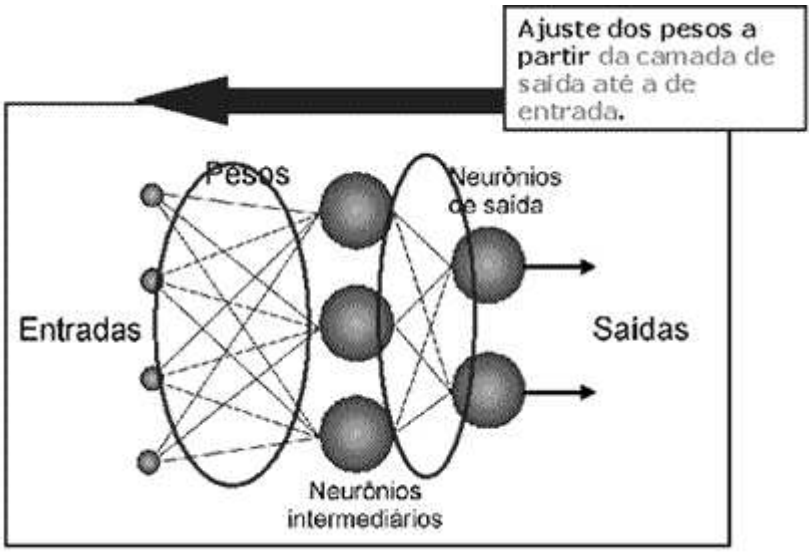


Figura 8: Algoritmo de aprendizado a um conjunto de regras bem definidas para a solução de um problema de aprendizado. Algoritmo "Backpropagation."

$$e_k = d_k - y_k \qquad (14)$$

Em que, para um estímulo k,

e– corresponde a um sinal de erro*;*

d – corresponde a uma saída desejada, apresentada durante o treinamento*;*

y – corresponde a uma saída alcançada pela rede após a apresentação do estímulo de entrada.

d. Modelo Perceptron Múltiplas Camadas (MultiLayerPerceptron – MLP)

Em 1958, Frank Rosenblatt propôs o modelo o Perceptrons, que era composto de uma estrutura de rede de neurônios MCP (McCulloch-Pitts) e uma regra de aprendizado (Braga et al., 2007). Esse modelo possuía apenas uma camada e tinha como saída um valor binário (Ludwig Junior & Montgomery, 2007; Haykin, 2001). Entretanto, por possuir uma única camada, esta RNA podia se aplicada apenas a problemas linearmente separáveis. Essa limitação foi resolvida em 1986 com a criação e aplicação do algoritmo back-propagation as redes de múltipla camada (Braga et al., 2007; Haykin, 2001; Mcclellan & Rumelhart, 1987).

As RNAs de múltiplas camadas com uma ou mais camadas intermediárias, função de ativação não-linear (Logsig ou Tansig) são capazes de resolver problemas de classificação linearmente separáveis ou não (Sarle, 1994). Esse novo modelo é conhecido como MultiLayerPerceptron - MLP.

A rede MLP possui neurônios com função não linear, possuem uma ou mais camadas ocultas ou intermediarias e possui alto grau de conectividade entre seus elementos processadores. Esta conectividade é definida pelos pesos sinápticos. As camadas intermediárias da rede são como detectores de características, as quais serão representadas através dos pesos sinápticos. Uma camada é suficiente para aproximar qualquer função contínua e duas camadas podem aproximar qualquer função matemática (Cybenco, 1989)

2.4. UTILIZAÇÃO DAS REDES NEURAIS ARTIFICIAIS NA CLASSIFICAÇÃO DE POPULAÇÕES

As RNAs têm sido utilizadas em diversas áreas da ciência como finanças, marketing, medicina e estatística como pode ser constato na literatura nos trabalhos: de diagnóstico médico (Mangasarian et al., 1990, Steiner et al., 1994), de predição de falência bancária (Tam et al., 1992), de aplicação à fabricação da pasta e papel industrial (Fadum, 1993), do mundo financeiro (Cipra, 1992), do controle de processos químicos (Nascimento et al., 1993), da obtenção de um modelo organizacional (Almeida, 1995).

As redes Neurais têm ainda contribuindo em inúmeros trabalhos de melhoramento genético e áreas afins. Há relatos de uso das RNAs para classificação de imagens em sensoriamento remoto (Aitkenhead & Aalders, 2008), análise de diversidade genética (Barbosa et al., 2011), identificação de genótipos superiores (Mugnai et al., 2008) e predição do valor genético em animais (Ventura et al., 2012), predição de valores genéticos em plantas (Peixoto, 2012) e classificação de frutos (Nakano,1997; Ramos, 2003; Simões et al., 2003), classificação de genótipos (Nascimento et al., 2013) dentre outros.

Alguns autores têm avaliado o desempenho das RNAs comparativamente com métodos de análises discriminante, aplicado ao processo de classificação populações (Barbosa et al., 2011;Bennett e Mangasarian, 1992; Patuwo et al., 1993; Pereira, 2009;. Steiner et al.,1994;

Tam et al., 1993). Nesses trabalhos, faz-se necessário o desenvolvimento de modelos que permitam estabelecer o relacionamento entre a entrada de padrões de classificação, análise e processamento dessa informação e convergência para uma saída definida. A rede neural deve aprender a reconhecer padrões de entrada e definir a saída segundo classes definidas, ou seja, dado um determinado padrão de entrada, escolher em que categoria ele se enquadra melhor (Ramos, 2003).

As redes neurais se baseiam nos dados para extrair sua capacidade de generalização. Portanto, um treinamento adequado é a etapa mais importante do processamento dos dados. Para isso faz-se necessário um bom conjunto de dados para que os pesos atribuídos às conexões entre os neurônios sejam capazes de validarem a aprendizagem e de melhorar o desempenho. O conjunto de dados pode ser obtido de bancos, quando disponíveis, ou obtido através de experimentações no melhoramento genético o que demanda tempo e recursos.

Em muitas situações, tanto o tempo quanto à disponibilidadede recursos são fatores limitantes. Sendo assim, uma boa alternativa para se obter um conjunto de dados de treinamento adequado é a utilização das técnicas de simulação que permite a obtenção de um grande volume de dados em um curto período de tempo sem os custos de implantação e condução de experimentos (Bhering, 2008 e Corrêa 2001).

2.5. O USO DE SIMULAÇÃO EM ESTUDOS DE DIVERSIDADE GENÉTICA

O avanço crescente na área de informática tem permitido estudos cada vez mais abrangentes nas ciências em geral. Isso ocorre tanto pelo desenvolvimento de tecnologias, que levaram ao sequenciamento do genoma de varias espécies, quanto pela descoberta de mecanismos que permitam o armazenamento e analises dessas sequências, tornando possível estuda-las das mais diversas formas, o que permite um avanço inimaginável.

Nos estudos de diversidade genética as análises biométricas têm avançado muito com o desenvolvimento de aplicativos cada vez mais apropriados para análise de informações de natureza fenotípica quanto genotípica. Além disso, avanços da área de bioinformática–ciência responsável por armazenar e relacionar dados biológicos, com o auxílio de métodos

computacionais e algoritmos matemáticos (Prosdocimi, 2007)– tem possibilitado o estudo de fenômenos biológicos, realizados por meio da simulação de dados. De acordo com Cruz (2006), para a simulação de um fenômeno biológico, devem ser estabelecidos parâmetros e restrições, de forma que o cenário emulado reflita da maneira mais fiel possível o fenômeno biológico estudado.

Segundo Dachs (1988), o processo de simulação consiste em emular, por meio de recursos computacionais, o comportamento de um sistema real englobando certos tipos de modelos lógicos, que permitam descrever o sistema natural (Naylor,1971). Ainda, segundo Banks (2000), a simulação permite a criação de uma história artificial da realidade e que permite a realização das observações e inferências nas características de operação do sistema real representado.

O processo de simulação tem sido utilizado desde 1950 pelo setor de mineração. Entretanto, são diversas as áreas de aplicação da simulação. Banks et al. (1998), Harrell et al. (2000), Law e Kelton (1986) e Lobão (2000), destacam os sistemas computacionais e de telecomunicações, fabricação, negócios, logística, militar, treinamento e científica. A simulação de populações para estudos da estrutura genética é importante e vem sendo utilizada atualmente em vários ramos da genética, como em estudos da filogeografia humana (Novembre et al., 2008), análise discriminante em estruturas de populações (Jombart et al., 2010), estudos de genoma (Price et al., 2006), detecção das interações gene-a-gene e variância genética (Bhattacharya et al., 2010).

Os estudos fundamentados em simulação se baseiam em modelos mais simples que os sistemas reais o que deve ser levado em consideração, pois, a simplicidade do modelo simulado não deve afetar o seu desempenho em comparação ao modelo real. Segundo McNitt (1985), a simulação envolve modelos que representam a entidade a ser investigada e é ainda uma metodologia para avaliação destes modelos.

Segundo Silva et al., (1999), estimativas determinadas sobre um número maior de repetições são mais precisas. Por esse motivo, estimativas baseadas na utilização de bancos de dados são mais devidamente realizadas. Nem sempre, é possível realizar experimentos que com grande numero de observações e repetições, por esse motivo as técnicas de simulação de dados

com medias e variâncias conhecidas bem como a ampliação dos dados conservando essas propriedades tem poupados tempo e dinheiro permitindo estudos mais diversificados e abrangentes.

A simulação de dados por ser uma ferramenta capaz de fornecer resultados para análises mais elaboradas a respeito da dinâmica do sistema permite uma interpretação mais profunda e abrangente do sistema estudado (Harrel, 2000). É importante que, em alguns estudos da área de melhoramento genético, cada característica simulada tenha a propriedade de descrever uma variável com a média, herdabilidade e precisão experimental estabelecidas pelo pesquisador (Cruz, 2013).

O processo de simulação segue o método científico, ou seja, formula as hipóteses, prepara o experimento, testa as hipóteses através do experimento e valida às hipóteses através dos resultados obtidos (Harrel, 2000). Dessa maneira, é fundamental que se tenha um bom planejamento do estudo. Por isso, simular requer não somente o conhecimento de um software específico, mas também recursos humanos especializados em análise de dados.

A simulação contribui também com a formulação de referenciais teóricos e práticos que possam orientar novos autores sobre o uso de softwares e métodos biométricos, para que se possa aproveitar melhor o conjunto de dados e interpretar corretamente os resultados.

2.6. REFERÊNCIAS BIBLIOGRÁFICAS

AITKENHEAD, M.; AALDERS, I. Classification of Landsat Thematic Mapper imagery for land cover using neural networks. **International Journal of Remote Sensing**,v. 29, n. 7, p. 2075-2084, 2008.

ALMEIDA, F. C. Desvendando o uso de Redes Neurais em Problemas de Administração de Empresas. **Revista de Administração de Empresas**, São Paulo, 1995, p. 46-55.

ANDERSON T.W. **An Introduction to Multivariate Statistical Analysis.**New York: John Wiley & Sons, 1958, 345 p.

ALVES, R.M.; GARCIA, A.A.F.; CRUZ, E.D.; FIGUEIRA, A. Seleção de descritores botânico-agronômicos para caracterização de germoplasma de cupuaçuzeiro. **Pesquisa Agropecuária Brasileira**, v.38, p.807-818, 2003.

ARAUJO, D.G. de; CARVALHO, S.P.; ALVES, R.M. Divergência genética entre clones de cupuaçuzeiro (*Theobroma grandiflorum* Willd. ex Spreng. Schum.). **Ciência e Agrotecnologia**, v.26, p.13-21, 2002.

BANKS, J. **Handbook of simulation: principles, methodology, advances, applications, and Practice**. New York: John Wiley & Sons, 1998.

BANKS, J. **Introduction to simulation.** Proceedings of the Winter Simulation Conference. Atlanta, 2000.

BARBOSA, V. C. **Redes Neuronais e "Simulated Annealing" como Ferramentas para Otimização Combinatória, Investigación Operativa**, 1989, vol. 1, n. 2, p.125-142.

BARBOSA, C. D.; VIANA, A. P.; QUINTAL, S. S. R.; PEREIRA, M. G. Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology**,v. 11, n. 3, p. 224-231, 2011.

BARRETO, J. M. **Introdução às Redes Neurais Artificiais**. Santa Catarina: UFSC, 2004.

BENNETT, K. P. & MANGASARIAN, O. L**. Robust Linear Programming Discrimination of Two Linearly Inseparable Sets,** Optimization Methods and Software, 1992, vol.1, p. 23-34.

BHERING,L.L.**Mapeamento genético em famílas simuladas de irmãos completos.** Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, 166p, 2008.

BORÉM, A.; Miranda, G. V. 2009. **Melhoramento de Plantas.** 5ta ed. Editora UFV, Viçosa, 2009, 523p.

BORÉM, A. **Biotecnologia florestal**. Universidade Federal de Viçosa, 2007.

BRAGA, A. P.; FERREIRA, A. C. P. L.; LUDERMIR, T. B. **Redes neurais artificiais: teoria e aplicações.**LTC Editora, 2007.

BHATTACHARYA, S.; HEITMANN, K.; WHITE, M.; LUKIĆ, Z.; WAGNER, C.; HABIB, S. Mass Function Predictions Beyond LCDM. **arXiv preprint arXiv:1005.2239**, 2010.

CORRÊA, F.J.C. **Avaliação de métodos de seleção tradicionais, assistida por marcadores moleculares e por genes candidatos, com dados simulados.** Viçosa, MG: UFV, 2001, 54p. Tese (Mestrado em Zootecnia) Universidade Federal de Viçosa, 2001.

CIPRA, B. A. **A Chaotic Walk on Wall Street,** SIAM News, may 1992.

CRUZ, C.D. 2013. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**. 35: 271-276.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético ed.5**. Viçosa, UFV**.** 480 p. 2012.

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

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos Biométricos Aplicados ao Melhoramento Genético** - Vol 2. 2.ed. Viçosa: UFV, 2006. 585p.

CRUZ, C. D.; **Princípios de Genética Quantitativa** ed.1. Viçosa, UFV. 92 p. 2005.

CYBENKO,G. Continuous Valued Neural Networks with Two Hidden Layers Are Suficient. **Technical Report** (Tufts University, Medford, 1988).

DACHS, J. N. W. **Estatística computacional: uma introdução em turbo pascal**. Rio de Janeiro. LTC, 1988.

DEGROOT, Jack. **Neuroanatomia**. ed. 21. Rio de Janeiro : Guanabara, 1994.

DIAS, L.A. dos S.; KAGEYAMA, P.Y.; CASTRO, G.C.T. Divergência genética multivariada na preservação de germoplasma de cacau (*Theobroma cacao* L.). **Agrotrópica**, v.9, p.29-40, 1997.

FADUM, O. **Artificial Inteligence : expert systems, fuzzy logic and neural network applications in the paper industry**, Pulp & Paper, 1993.

CARVALHO FILHO, E. C. B. C., **Modelagem, Aplicações e Implementações de redes Neurais**. Anais da IV Escola Regional de Informática da SBC Regional Sul, 21 a 27 de abril de 1996. Páginas 36 - 53.

FISHER, R.A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, v.7, p.179-188, 1936.

GALVÃO, C. O.; VALENÇA, M. J. S.; VIEIRA, V. P. P. B.; DINIZ, L. S.; LACERDA, E. G. M.; CARVALHO, A. C. P. L. F.; LUDERMIR, T. B. **Sistemas inteligentes: Aplicações a recursos hídricos e ciências ambientais**. UFRGS: ABRH, 1999.

GORNI, A. A. Redes Neurais Artificiais - **Uma abordagem revolucionária em Inteligência Artificial.**Micro Sistemas, São Paulo, 1993.

GUYTON, A.C. **Fisiologia Humana**. 6ª ed., Rio de Janeiro, Ed. Interamericana, 1988.

HARREL, C. R.; GHOSH, B. K.; BOWDEN, R. **Simulation Using ProModel®**. McGraw-Hill, 2000.

HAYKIN, S. **Neural Networks - A Comprehensive Foundation**. Macmillan College Publishing, inc., 1994.

HEBB, **D.O. Brain Mechanisms and Learning**. London: J. F. Delafresnaye (Ed.), 1961.

HECHT-NIELSEN, R. **Neurocomputing**. Addison-Wesley Publishing Company, New York, 1991.

HOPFIELD, J.J. Neural networks and physical systems with emergent collective properties. Pro. Nat. **Acad. Sci**., 79:2554-8,1982

JOHNSON, R. A. & WICHERN, D. W. **Applied Multivariate Statistical Analysis**. New Jersey, Prentice-Hall, inc., 1988.

JOMBART, T.; DEVILLARD, S. & BALLOUX, F. 2010. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. **BMC Genetics** 11:94 doi:10.1186/1471-2156-11-94.

KANDEL, E. R. Cellular mechanisms of learning and the biological basis of individuality. In Kandel, E. R., Schwartz, J. H., and Jessell, T. M., editors,**Principles of Neural Science**, chapter 65, pages 1009--1031. Appleton & Lange, Norwalk, CT, third edition. 1991

LAW, A. M., KELTON, W. D. **Simulation Modeling and Analysis**, 3rd ed. McGraw-Hill, New York, 2000.

LOBÃO, E. C. **Discussão, sistematização e modelamento do processo de realização de estudos de simulação**. Tese doutorado. São Carlos: USP, 2000.

LORENCETTI, C.; et al. Retrocruzamento como uma estratégia de identificarge nótipos e desenvolver populações segregantes promissoras em aveia. **Ciência Rural.** v.36, n4, jul-ago, 2006.

LUIDWIG JUNIOR, O.; MONTGOMERY, E. Redes Neurais – **Fundamentos**

e Aplicações com programas em C. Editora Ciência Moderna, 2007, 125p.

MANGASARIAN, O. L., SETIONO, R. & WOLBERG, W. H. **Pattern Recognition via Linear Programming : Theory and Application to Medical Diagnosis**, in : Large-Scale Numerical Optimization, Thomas F. Coleman and Yuying Li, (Eds.), SIAM, Philadelphia, p.22-30, 1990.

MARCHIORO, V.S.; CARVALHO, F.I.F. de; OLIVEIRA, A.C. de; CRUZ, P.J.; LORENCETTI, C.; BENIN, G.; SILVA, J.A.G. da; SCHMIDT, D.A.M. Dissimilaridade genética entre genótipos de aveia. **Ciência e Agrotecnologia**, v.27, p.285-294, 2003.

Martins, E. C. A., Peluzio, J. M., Coimbra, R. R., da Silveira, M. A., Oliveira, J. D. D. D., & de Oliveira Junior, W. P. Diversidade genética em batata-doce no Tocantins= Genetic diversity in sweet potato in Tocantins. **Bioscience Journal,** v. 30, n. 2, 2013**.**

MCCULLOCH, W. S.; PITTS, W. A logical calculus of the ideas immanent in nervous activity. **Bull Math Biol**,v. 5, n. 4, p. 115-133, 1943.

MCNITT,L.L., **Simulação em Basic.** Rio de Janeiro, Livros Técnicos e Científicos Editora S.A., 1985, 348p.

Melo, R. A., Resende, L. V., Menezes, D., Beck, A. P. A., Costa, J. C., Coutinho, A. E, & Nascimento, A. V. S. do. Genetic similarity between coriander genotypes using ISSR markers. **Horticultura Brasileira,** *29*(4), 526-530, 2011.

MINGOTI, S. A.; Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada, Editora UFMG, 2005.

MINSKY, M., S. PAPERT. **Perceptrons**: **An introduction to computational geometry** MIT Press, Cambridge, Massachusetts (1969)

MUGNAI, S.; PANDOLFI, C.; AZZARELLO, E.; MASI, E.; MANCUSO, S. Camellia japonica L. genotypes identified by an artificial neural network based on phyllometric and fractal parameters. **Plant systematics and evolution**, v. 270, n. 1, p. 95-108, 2008.

NAKANO, K.. Application of neural networks to the color grading of apples. **Computers and Electronics in Agriculture** 18 (1997) 105-116.

NASCIMENTO, C. A. O. e YAMAMOTO, C. I. **Modelagem de Processos Químicos via Redes Neurais**, 1993, I Workshop em Redes Neurais, IPT, São Paulo.

NASCIMENTO FILHO, F.J. do; ATROCH, A.L.; SOUSA, N.R. de; GARCIA, T.B.; CRAVO, M. da S.; COUTINHO, E.F. Divergência genética entre clones de guaranazeiro. **Pesquisa Agropecuária Brasileira**, v.36, p.501-506, 2001.

NASCIMENTO, M., PETERNELLI, L. A., CRUZ, C. D., NASCIMENTO, A. C. C., FERREIRA, R. D. P., BHERING, L. L., & SALGADO, C. C. (2013). Artificial

neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology**, 13(2), 152-156.

NASS, L. L. et al. (Eds.) **Recursos genéticos e melhoramento** - plantas.Rondonópolis: Fundação MT, 2001. cap. 2, p. 29-55.

NAYLOR,T.H., BALINTFY,J.L., BURDICK,D.S., CHU,K. **Técnicas de simulação em computadores.** São Paulo, Vozes, 1971, 401p.

NGUYEN, D.; WIDROW, B**. Improving the Learning Speed of 2 Networks by Choosing Initial Values of the Adaptatives Weights.** University, Stanford, CA, 1990.

NEI, M. Genetic distance between populations. **American Naturalist**.Chicago, v. 106, p. 238-292, 1972.

OLIVEIRA, Maria do Socorro Padilha de; FERREIRA, Daniel Furtado; SANTOS, JoãoBosco dos. Divergência genética entre acessos de açaizeiro fundamentada em descritores morfoagronômicos. **Pesq. agropec. bras.**, Brasília , v. 42, n. 4, 2007.

OLIVEIRA, E. Q. D., BEZERRA, N. F., Negreiros, M. Z. D., Barros Júnior, A. P., FREITAS, K., SILVEIRA, L., & LIMA, J. Desempenho agroeconômico do bicultivo de alface em sistema solteiro e consorciado com cenoura." **Horticultura Brasileira**, v. 22, n. 04, p. 712-717, 2004.

PATUWO, E., HU, M. Y. & HUNG, M. S. **Two-Group Classification using Neural Networks**, Decision Sciences, 1993, vol 24, n. 4.

PEIXOTO, L. A. **Redes Neurais Artificiais na predição do valor genético.** Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, 97p, 2013.

PEREIRA, T.A. **Discriminação de populações com diferentes graus de similaridade por redes neurais artificiais.** Dissertação (Mestrado em Genética e Melhoramento) – Universidade Federal de Viçosa, 88p, 2009.

PRICE, A. L.; PATTERSON, N. J.; PLENGE, R. M.; WEINBLATT, M. E.; SHADICK, N. A.; REICH, D. Principal components analysis corrects for stratification in genome-wide association studies. **Nat Genet**, v. 38, n. 8, p. 904-909, 2006.

PROSDOCIMI, Francisco. Curso de Bioinformática. Biotecnologia ciência & desenvolvimento, 77f., 2007. Disponível em: <http://biotec.icb.ufmg.br/chicopros/FProsdocimi07_CursoBioinfo.pdf>. cessado em: 04 fevereiro 2014.

RAMOS, J.P.S. **Redes Neurais Artificiais na Classificação de Frutos: CenárioBidimensional**, Ciênc. agrotec., Lavras. V.27, n.2, p.356-362, mar./abr., 2003.

REGAZZI, A.J. Análise Multivariada, Notas de Aula EST 746, Departamento de Informática da Universidade Federal de Viçosa, v.2, 2006.

ROSENBLATT,F. Principles of Neurodynamics: **Perceptrons and the theory of brain mechanisms.** Spartan Books, New York, 1962.

RUMELHART, D. E. & MCCLELLAND, J. L. (1987). Learning the past tenses of english verbs: Implicit rules or par-allel distributed processing. In B. MacWhinney (Ed.), **Mechanisms of Language Acquisition** (pp. 194-248). Mah-wah, NJ: Erlbaum.

RUMELHART,D.E., HINTON,G.E., WILLIAMS,R.J., Learning representa tions by back-propagating errors. **Nature**, 323:533-536,1986.

SARLE, W. S. **Neural networks and statistical models**. Proceedings of the Nineteenth Annual SAS Users Group International Conference, 1994. 13p.

SILVA, C.M.; GONÇALVES-VIDIGAL, M.C.; VIDIGAL FILHO, P.S.; SCAPIM, C.A.; DAROS, E.; SILVÉRIO, L. Genetic diversity among sugarcane clones (Saccharum spp.). **Acta Scientiarum**. Agronomy, v.27, p.315-319, 2005.

SILVA, Eugênio. OLIVEIRA, Anderson Canêdo de. Dicas para a Configuração de Redes Neurais. Disponível em: http://equipe.nce.ufrj.br/thome/ grad/ nn/ mat_didatico /dicas_configuracao_rna.pdf>. Acesso em: Nov 2008.

SIMÕES, A.S. e COSTA, A.H.R., Classificação de Laranjas Baseada em Padrões Visuais. Anais do Simpósio Brasileiro de Automação Inteligente, 2003.

SOUZA, F.F.; QUEIRÓZ, M.A.; DIAS, R.S.C. Divergência genética em linhagens de melancia. **Horticultura Brasileira**, v.23, p.179-183, 2005.

STEINER, M. T. A. Verificação da Eficiência de Métodos Estatísticos no R. de Padrão via Simulação, XXVI SBPO, Florianópolis, dez. 1994. Pk

SUDHEER, K.; GOSAIN, A.; RAMASASTRI, K. Estimating actual evapotranspiration from limited climatic data using neural computing technique. **Journal of Irrigation and Drainage Engineering**,v. 129, n. 3, p. 214-218, 2003.

TAFNER, M., XEREZ, M., e RODRIGUES FILHO, I.**Redes Neurais artificiais : introdução e princípios de neurocomputação.**Blumenau : EKO, 1995.

TAM, K. Y. & KIANG, M. Y. Managerial Applications of Neural Networks : The Case of Bank Failure Predictions, **Management Sciences** 38 n. 7, 1992, p.926-947.

VASCONCELOS, E.S. de; CRUZ, C.D.; BHERING, L.L. FERREIRA, A. Estratégias de amostragem e estabelecimento de coleções nucleares. **Pesquisa Agropecuária Brasileira**, v.42, p.507-514, 2007.

VENTURA, R.V.; SILVA, M.A.; MEDEIROS, T.H.; DIONELLO, N.L.; MADALENA, F.E.; FRIDRICH, A.B.; VALENTE, B.D.; SANTOS, G.G.; FREITAS, L.S.; WENCESLAU, R.R.; FELIPE, V.P.S.; CORRÊA, G.S.S. Uso de redes neurais artificiais na predição de valores genéticos para peso aos 205 dias em bovinos da raça Tabapuã. **Arq.Bras. Med. Vet. Zootec.,** v.64, n.2, p.411-418, 2012.

VIANA, J. P. G. **Diversidade genética em alho (Allium sativum L.)** Universidade Federal do Piauí (2013).

XAVIER, E. **"Aplicação de Inteligência Computacional na Gerência de Redes Através da Automatização do Uso de Agentes Móveis".**(Dissertação de Mestrado). Universidade Federal de Santa Catarina- UFSC. Florianópolis.2003.

WIDROW, B.; HOFF, M.E. **Adaptive Switching Circuits,** In 1960. IRE. 1960.

CAPITULO 1

BIOMETRIA APLICADA AO ESTUDO DA DIVERSIDADE EM POPULAÇÕES ESTRUTURADAS NO DELINEAMENTO GENÉTICO DE RETROCRUZAMENTOS

RESUMO

O objetivo desse trabalho foi avaliar a capacidade de discriminação de técnicas multivariadas adotando um conjunto de populações derivadas por retrocruzamentos, com diferentes graus de similaridade e complexidade de diferenciação**.** Cada população, gerada por simulação, foi caracterizada por um conjunto de elementos mensurados por características de natureza contínua. Para esse estudo dados genotípicos foram simulados para trezes populações, estruturada em delineamento genético de cruzamentos entre genitores contrastantes seguido de retrocruzamentos, com 100 indivíduos cada. Foram considerados 50 locos independentes, cada qual com dois alelos. O conjunto de dados simulados foi utilizado para estimação da dissimilaridade genotípica e subsequente análise da diversidade genética. As relações de parentescos e a estruturação hierárquica foram estabelecidas considerando populações genitoras geneticamente divergentes, híbrido F_1 e cinco gerações de retrocruzamento em relação a cada um dos genitores, permitindo estabelecer parâmetros de eficácia das metodologias testadas. Para fins de comparação de medidas e método de agrupamento do padrão de dissimilaridade foi estabelecido um sistema estruturado de populações em que o parentesco ou os níveis de hierarquia são previamente conhecidos. Assim, a partir do par de genitores (P1 e P2) divergentes foram geradas 11 outras populações. Considerou-se um conjunto de treze características com herdabilidade de 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 e 80. O efeito genético dos caracteres foram estabelecidos por meio da ação de 20 locos, tomados ao acaso entre os 50 locos previamente simulados, com efeito aditivo diferencial, com pesos estabelecidos a partir de uma distribuição binomial, e grau médio de dominância nulo. Foram obtidas medidas de dissimilaridade para realização de método de agrupamentos hierárquicos e de otimização e projeção em plano bidimensional para análise de sua diversidade. Os métodos foram eficientes agrupando nos extremos as populações de genitores P1, P2 seguidas por seus retrocruzamentos, o híbrido F1, no centro**,** para dados genotípicos. Entretanto, devido ao nível de similaridade ser superior a 75%, para dados fenotípicos ficou evidente que faz se necessário a utilização de métodos estatísticos para diferenciação dos mesmos.

ABSTRACT

The aim of this study was to evaluate the discrimination capability of multivariate techniques adopting a set of populations derived by backcrossing, with varying degrees of similarity and differentiation complexity. Each population generated by simulation, was characterized by a set of elements measured by characteristics of a continuous distribution. Genotypic data for this study were simulated for thirteen populations, structured in a genetic design crosses between divergent parents followed by backcrossing with 100 individuals each. Were considered 50 independent loci, each one with two alleles. The set of simulated data was used to estimate the genotypic dissimilarity and subsequent analysis of genetic diversity. For purposes of comparison of measured and method of grouping pattern dissimilarity a structured populations in which the relationship or hierarchy levels are previously known system was established. Thus, from the pair of parents (P1 and P2) 11 other divergent populations were generated. The relations of kinship and hierarchical structuring were established considering genetically divergent populations, F1 hybrid and five generations of backcrossing in relation to each of the parents, establishing measures of effectiveness of the tested methodologies. Were considered a number of traits with thirteen heritability of 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 and 80. The genetic effect of the characters have been established to be ruled by 20 loci, randomly chosen among the 50 loci previously simulated with differential additive effect with weights established from a binomial distribution, and average degree of dominance null. Dissimilarity measures for carrying out the method of hierarchical clustering, optimization and bidimensional projection plane for analysis of diversity were obtained. The methods were efficient clutering extreme populations P1, P2 followed by their parents backcrossing, the F1 hybrid in the center for genotypic data. However, due to the similarity level above of 75 % for phenotypic data it became apparent that it is necessary to use a statistical methods to differentiate them.

1. INTRODUÇÃO

Os fatores ambientais sempre atuaram como agentes perturbadores de processos seletivos e, atualmente, as grandes variações climáticas e as adversidades geradas pelo aquecimento global, têm motivado estudos aprofundados tanto na escolha de genótipos mais produtivos e resistentes quanto no estabelecimento de estratégias eficazes que permitam preservação dos recursos biológicos existentes. Dentre estes estudos, destacam aqueles relacionados aos estudos da diversidade genética.

A mensuração e a preservação da biodiversidade existente se mostra essencial para a contínua evolução e para a sobrevivência de muitas espécies. Portanto, estudos de diversidade genética que permitam a discriminação de populações são essenciais para o gerenciamento dos bancos de germoplasma, assegurando a disponibilidade de informações sobre os acessos e facilitando o trabalho de seleção de possíveis doadores de genes (Cruz et al., 2011).

Deve-se ter em mente que a classificação incorreta de indivíduos, e, ou, populações representa grande prejuízo para programas de conservação da biodiversidade *in* ou *ex-situ*, para programas de preservação ambiental e para definição de cruzamentos em programas de melhoramento quando da utilização de germoplasma “exótico” ou quando da definição de grupos heteróticos. Neste sentido, os estudos de estruturação de populações são de grande valia, uma vez que permite entendimento das melhores estratégias para incrementar e preservar a diversidade das espécies e ou dentro das espécies.

Atualmente, há varias metodologias disponíveis para a quantificação e a avaliação da diversidade em estudos populacionais, a partir de informações fenotípicas e genotípicas. No entanto, devido ao grande variedade de informações a serem avaliadas, dada ainda as particularidades de cada material biológico, a escolha e a correta aplicação da metodologia mais adequada são de grande importância para obtenção de resultados confiáveis. Nesse sentido, técnicas de bioinformática e de análise estatística, tem se mostrado muito úteis nos programas de melhoramento genético, principalmente nos estudos de diversidade genética (Barbosa et al., 2011).

Diversos estudos utilizam análises de agrupamento na visualização e interpretação da diversidade genética, com base em caracteres morfológicos e agronômicos em plantas como a gabiroba (Rezende et al., 2009), coentro (Melo et al., 2011), a batata doce (Martins et al., 2013), e o alho (Viana, 2013).

Entretanto, mesmo nas situações em que os pesquisadores dispõem de dados experimentais adequados, tanto no aspecto quantitativo quanto qualitativo, alguns resultados de discriminação tem se mostrados inapropriados ou de baixa acurácia, quando obtidos por meio de metodologias fundamentadas em análises multivariadas. Outras técnicas mais eficazes, utilizando outros princípios merecem ser investigadas.

A comparação de métodos biométricos alternativos é dificultada pela falta de um conjunto de dados apropriados em especial aqueles resultantes da experimentação tendo em vista que tanto o tempo quanto os de recursos físicos e financeiros são fatores limitantes. Sendo assim, uma boa alternativa para se obter um conjunto de dados que permitam a comparação de eficácia de métodos e modelos é aquela fundamentada na utilização das técnicas de simulação que permite a obtenção de um grande volume de informações, sob um sistema que se conhece valores paramétricos e hipóteses possam ser convenientemente testadas. Estes dados podem ser obtidos em um curto período de tempo sem os custos de implantação e condução de experimentos (Bhering, 2008 e Corrêa, 2001).

Por meio de processos de simulação, é possível emular uma situação real, normalmente encontrada em um programa de melhoramento, e avaliar e comparar técnicas tradicionalmente utilizadas para discriminação de populações. Sendo assim, o objetivo desse trabalho foi evidenciar a similaridade genotípica entre populações, a partir de dados genotípicos e fenotípicos dentro da expectativa do padrão previsto pelo processo meiótico, comparando os resultados obtidos por meio das técnicas de análises multivariadas de diversidade genética.

2. MATERIAL E MÉTODOS

2.1. Simulação dos dados Genotípicos

A simulação e análise dos dados foram realizadas no Laboratório de Bioinformática da Universidade Federal de Viçosa, localizado no Instituto de Biotecnologia aplicada a Agropecuária (BIOAGRO) utilizando aplicativo computacional Genes (Cruz, 2013).

Dados genotípicos foram, originalmente, simulados para dez populações em equilíbrio de Hardy-Weinberg, com 100 indivíduos em cada população. Foram geradas informações relativas a 50 locos manifestando, em cada loco, dois alelos codominantes. Este conjunto prévio de dados foi utilizado para o cálculo de uma medida de dissimilaridade genotípica de Nei (1972) e também para estudos da diversidade genética das populações pela projeção gráfica das distâncias em plano bidimensional. O par de populações mais divergente foi tomado para gerar um sistema hierárquico de retrocruzamentos tal como apresentado na Figura 1.

As relações de parentescos e a estruturação hierárquica foram estabelecidas considerando as populações genitoras geneticamente divergentes, o híbrido F_1 e cinco gerações de retrocruzamentos obtidas em relação a cada um dos genitores, permitindo estabelecer parâmetros para quantificação da eficácia das metodologias testadas.

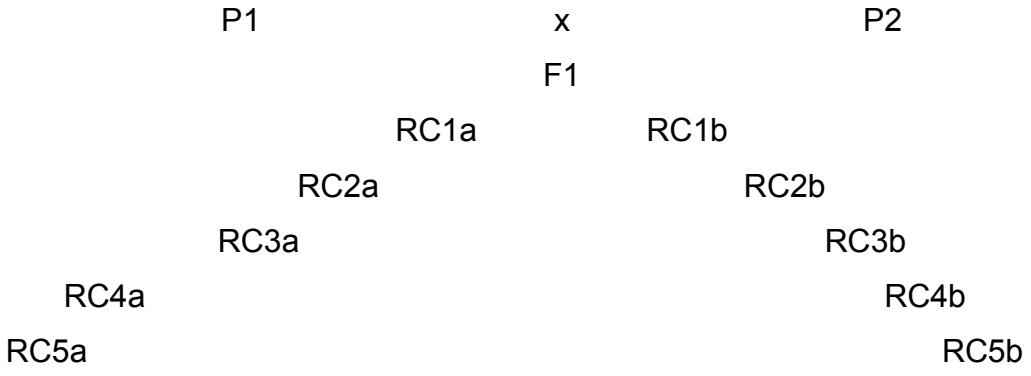


Figura 1: Esquema estruturado dos cruzamentos entre o P_1 e P_2 e suas gerações de retrocruzamentos “a” e “b” como genitores recorrentes.

O conhecimento prévio dos parâmetros genéticos relativos à diferenciação entre cada par de populações assume grande importância neste trabalho para a quantificação e determinação da eficácia do método de discriminação a ser recomendado para uso como critérios de seleção em um programa de melhoramento. Entretanto, na prática a diversidade genética tem sido mensurada a partir de informações fenotípicas que depende em parte da estrutura genética das populações, expressa por frequências genotípicas e alélicas, e em parte por fatores ambientais determinantes da variabilidade ambiental (Falconer & Mackay, 1996). Por isto, para enriquecimento deste trabalho, foi realizado estudo da diversidade genética tanto com base nas informações genotípicas, mimetizando situações em que se dispõe apenas de informações de dados genotípicos, quanto a partir de valores fenotípicos resultantes da inclusão de fator ambiental (ruído) para observar se tais técnicas biométricas seriam apropriadas para realizar as análises de diversidade com a finalidade de diferenciar essas populações à semelhança do que seria esperado com base em princípios meióticos que descrevem a contribuição de cada genitor para a próxima geração.

2.2. Estabelecimento dos valores fenotípicos

Neste estudo foram utilizadas informações genotípicas e fenotípicas de 13 populações π_j, j=1,2,...,p=13, constituídas por indivíduos (n_j=100 para todo j) mensurados em relação a v=13 características quantitativas, contínuas, com distribuição normal, tendo média e variância previamente conhecidas. Para cada população simulada, considerou-se igualdade das matrizes de variância e covariâncias, uma vez que sem essa pressuposição perde-se a linearidade das funções discriminantes. Todas as populações foram simuladas por meio do uso do aplicativo computacional Genes (Cruz, 2013).

Foram simulados 13 variáveis com valores de média e herdabilidade previamente estabelecidos. O delineamento experimental adotado na simulação foi o delineamento inteiramente casualizado em que cada população apresentava 100 genótipos, assumindo-se herdabilidade de 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 e 80%, valores de média semelhante ao valor da herdabilidade apenas para fins didáticos.

Estas características foram estabelecidas pela ação de alelos de 20 locos, tomados ao acaso entre os 50 previamente genotipados, com efeito aditivo diferencial e com pesos da importância do loco, sobre a variabilidade genotípica total do caráter, estabelecidos a partir de uma distribuição binomial e grau médio de dominância nulo.

Para cada variável, estabelecida a partir de valores da média e da herdabilidade conhecidos, foi utilizado o seguinte modelo estatístico determinante do valor fenotípico:

$$Y_{ij} = \mu + G_i + \varepsilon_{ij}$$

Em que:

Y_{ij}: observação simulada de uma dada característica para o i-ésimo indivíduo pertencente à j-ésima população;
μ: média geral da característica, cujo valor é especificado pelo pesquisador;
G_i: efeito associado ao i-ésimo genótipo;
ε_{ij}: erro aleatório, sendo $\varepsilon_{ij} \sim N(0, \sigma^2)$

Como eram conhecidos previamente os valores da herdabilidade e da média e, também, foi estabelecido o controle genético do caráter (número de locos e ação de cada loco), as estimativas da variância genética e ambiental puderam ser calculadas. Por princípios de genética quantitativa, as variâncias genéticas esperadas nas gerações de retrocruzamento podem ser previstas pelas frequências genotípicas e efeitos aditivos e de dominância, considerados nulos neste trabalho. O modelo aditivo foi empregado de forma que o valor genotípico de cada indivíduo foi gerado e, acrescido de um efeito ambiental, simulado admitindo distribuição normal, com média zero e variância σ^2, forneceu o valor fenotípico submetido às análises. Segundo Cruz et al., (2012), esse modelo além de mais simples, tem sido rotineiramente utilizado no melhoramento por fornecer valiosas informações para o êxito de programas de melhoramento.

Para fins didáticos as características utilizadas foram divididas em dois cenários: Cenário A- Alta herdabilidade (v = 55, 60,..., 80) e Cenário B – Baixa herdabilidade (v= 20, 25,..., 50), conforme pode ser visualizado nas Tabelas 1 e 2.

Tabela 1: Médias paramétricas das características simuladas das 13 populações constituindo o Cenário A de alta herdabilidade.

Populações	Características					
	h²=55	h²=60	h²=65	h²=70	h²=75	h²=80
P1	50,00	55,00	60,00	65,00	70,00	75,00
P2	25,00	27,50	30,00	32,50	35,00	37,50
F1	37,50	41,25	45,00	48,75	52,50	56,25
RC1a	43,75	48,13	52,50	56,88	61,25	65,63
RC2a	46,88	51,56	56,25	60,94	65,63	70,31
RC3a	48,44	53,28	58,13	62,97	67,81	72,66
RC4a	49,22	54,14	59,06	63,98	68,91	73,83
RC5a	49,61	54,57	59,53	64,49	69,45	74,41
RC1b	31,25	34,38	37,50	40,63	43,75	46,88
RC2b	28,13	30,94	33,75	36,56	39,38	42,19
RC3b	26,56	29,22	31,88	34,53	37,19	39,84
RC4b	25,78	28,36	30,94	33,52	36,09	38,67
RC5b	25,39	27,93	30,47	33,01	35,55	38,09

Tabela 2: Médias paramétricas das características simuladas das 13 populações constituindo o Cenário B de baixa herdabilidade.

	Características						
Populações	h²=20	h²=25	h²=30	h²=35	h²=40	h²=45	h²=50
P1	20,00	25,00	30,00	35,00	40,00	45,00	20,00
P2	10,00	12,50	15,00	17,50	20,00	22,50	10,00
F1	15,00	18,75	22,50	26,25	30,00	33,75	15,00
RC1a	17,50	21,88	26,25	30,63	35,00	39,38	17,50
RC2a	18,75	23,44	28,13	32,81	37,50	42,19	18,75
RC3a	19,38	24,22	29,06	33,91	38,75	43,59	19,38
RC4a	19,69	24,61	29,53	34,45	39,38	44,30	19,69
RC5a	19,84	24,80	29,77	34,73	39,69	44,65	19,84
RC1b	12,50	15,63	18,75	21,88	25,00	28,13	12,50
RC2b	11,25	14,06	16,88	19,69	22,50	25,31	11,25
RC3b	10,63	13,28	15,94	18,59	21,25	23,91	10,63
RC4b	10,31	12,89	15,47	18,05	20,63	23,20	10,31
RC5b	10,16	12,70	15,23	17,77	20,31	22,85	10,16

2.3. Técnicas multivariadas utilizadas no estudo da diversidade genética a partir de informações genotípicas e fenotípicas

Para estudo da diversidade genética entre populações e, para fins de comparação da eficácia do processo de discriminação, empregaram-se, neste trabalho, diferentes métodos de agrupamentos baseados em técnicas multivariadas. Foi utilizado do o termo “população” para designar o conjunto de populações caracterizadas pelas populações genitoras, híbridos F_1 e gerações de retrocruzamento obtidos utilizando cada um dos genitores como recorrentes.

a. Medida de dissimilaridade

Na análise de agrupamento, com base em valores genotípicos, foi utilizada, como medida de dissimilaridade, a distância genética de Nei (1972). Na análise de agrupamento, com base em valores fenotípicos, foi utilizado uma matriz de dissimilaridade cujos elementos eram representados pelos valores da distância generalizada de Mahalanobis. A distância de Mahalanobis ($D^{2)}$ entre as populações *i* e *i'* é dada por:

$$D^2_{ii'} = (\bar{X}_i - \bar{X}_{i'})' S^{-1} (\bar{X}_i - \bar{X}_{i'}) \quad (1)$$

S é a matriz de variâncias e covariâncias amostral comum a todos os indivíduos;

X_i = vetor de médias observadas para as populações i e i'.

b. Métodos de agrupamentos hierárquicos

A matriz de dissimilaridade gerada com os dados foi utilizada no método de agrupamento hierárquico da ligação média entre grupos (UPGMA) e também o método de ligação simples, também denominado de vizinho mais próximo. Como regra geral, a construção de um dendrograma parte de um grupo inicial que é formado pelas populações mais similares. Em seguida calcula-se a distância entre estas populações e as demais, caracterizadas, neste trabalho por populações de genitores, ou híbrido F1 ou gerações de retrocruzamentos. Nestes novos passos são novamente agrupados populações de maior similaridade. A distância entre os grupos é determinada pela média das distâncias entre pares de populações pertencentes aos diferentes grupos, sendo uma expressão geral para estes métodos:

$$d_{(ij)k} = \frac{n_i}{n_i+n_j} d_{ik} + \frac{n_i}{n_i+n_j} d_{jk} \quad (2)$$

Em que $d_{(ij)k}$ é a distância entre o grupo (*ij*), com tamanho interno n_i e n_j, respectivamente, caracterizando *i*, *j* e *k* como populações ou grupos de populações (Cruz et al. 2008).

Já no método da ligação simples ou do vizinho mais próximo o dendrograma é estabelecido pelas populações com maior similaridade, sendo a distancia entre uma população k e um grupo, formado pelas populações i e j, dada por:

$$d_{(ij)k} = mín\{d_{ik}, d_{jk}\} \quad (3)$$

em que $d_{(ij)k}$ é dada pelo menor elemento do conjunto das distâncias dos pares de populações (i e k) e (j e k). As conexões entre populações e grupos são feitas por ligações simples entre populações, ou seja, a distancia entre os grupos é definida como aquela entre as populações mais parecidas dentre esses grupos. A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = mín\{d_{ik}, d_{jk}, d_{il}, d_{jl}\} \quad (4)$$

ou seja a distância entre dois grupos formados, respectivamente, pelas populações (i e j) e (k e l) é dada pela menor distancia entre os pares (i e k), (i e l), (j e k), (j e l) (Cruz et al., 2008).

c. Método de agrupamento por otimização

Também foi utilizado o método de agrupamento por otimização ou método de Tocher, apresentado em Cruz & Carneiro (2006). O método de Tocher caracteriza por ser uma estratégia de agrupamento simultâneo, em que a separação dos genótipos em grupos é realizada de uma só vez. O método requer que seja identificado, na matriz de dissimilaridade, o par de populações mais similares e, então, formado o grupo inicial. A partir daí, é avaliada a possibilidade de inclusão de novas populações, adotando-se o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo (Cruz et al., 2008).

A entrada de uma nova população em um grupo sempre aumenta o valor médio da distância dentro do grupo. Para a inclusão de novas populações em um grupo, adota-se o critério de comparar o acréscimo no valor da distância dentro do grupo e um nível máximo permitido, obtido da maior entre as menores distâncias envolvendo cada indivíduo θ. A inclusão, ou não, da população k no grupo é, então, feita considerando o acréscimo médio promovido pela inclusão de uma população k em um grupo previamente estabelecido desde que seja menor que θ (Cruz et al., 2008).

d. Método da projeção gráfica

Outro método utilizado foi a projeção gráfica em plano bidimensional, em que as medidas de dissimilaridade são convertidas em escores relativos às duas variáveis (X e Y) que, quando representadas em gráficos de dispersão, irão refletir, no espaço bidimensional, as distancias originalmente obtidas a partir do espaço v-dimensional, em que v é o numero de caracteres utilizados para se obter as distancias (Cruz et al., 2008).

A viabilidade do uso dessa técnica é avaliada pela correlação entre as distancias originais e gráficas que deve ser superior a 0,8 e também pelo estresse e distorção que se recomenda ser inferiores a 20% (Cruz et al., 2011).

3. RESULTADO E DISCUSSÃO

3.1. Análise de dados genotípicos

Os estudos de diversidade genética têm sido realizados por meio de técnicas multivariadas de análise de agrupamento adaptadas, muitas vezes, de outras áreas da ciência. Assim, é comum o emprego de técnicas hierárquicas, próprias de estudos evolutivos ou de taxionomia, em um conjunto de acessos sem nenhuma estrutura de ancestralidade conhecida ou apropriada de ser aplicada ao conjunto de dados. Neste trabalho algumas técnicas de agrupamento foram utilizadas para avaliar, mesmo que de forma empírica, se os princípios de similaridade genética, com valores previamente conhecidos tendo em vista a contribuição gamética de genitores, possa ser devidamente demonstrado por métodos de agrupamento hierárquico, projeção de distâncias e agrupamentos de otimização. Para tanto foram realizadas as analises separadamente para os dados genotípicos e fenotípicos.

Nas análises de diversidade foram consideradas duas etapas. Na primeira o estudo foi feito considerando 10 genitores, representativos de populações em equilíbrio de Hardy-Weinberg em que o objetivo era a escolha do par de populações mais divergentes que seriam utilizados para o estabelecimento do delineamento genético estruturado em retrocruzamento. Na segunda etapa os genitores foram cruzados e a descendência retrocruzada com cada um dos genitores recorrentes.

Na fase inicial deste trabalho utilizou-se, para inferir sobre a diversidade entre 10 populações potenciais para uso como genitores, a projeção no plano bidimensional, dos elementos da matriz de distancia de Nei, calculada a partir das informações genotípicas simuladas para seleção dos genitores. Na figura 2 é apresentado o resultado que evidencia que as populações 10 e 7 foram os mais divergentes e passam a ser chamados de P_1 e P_2 e empregadas para fins de cruzamentos e estabelecimento do delineamento genético fundamentado em populações de retrocruzamento.

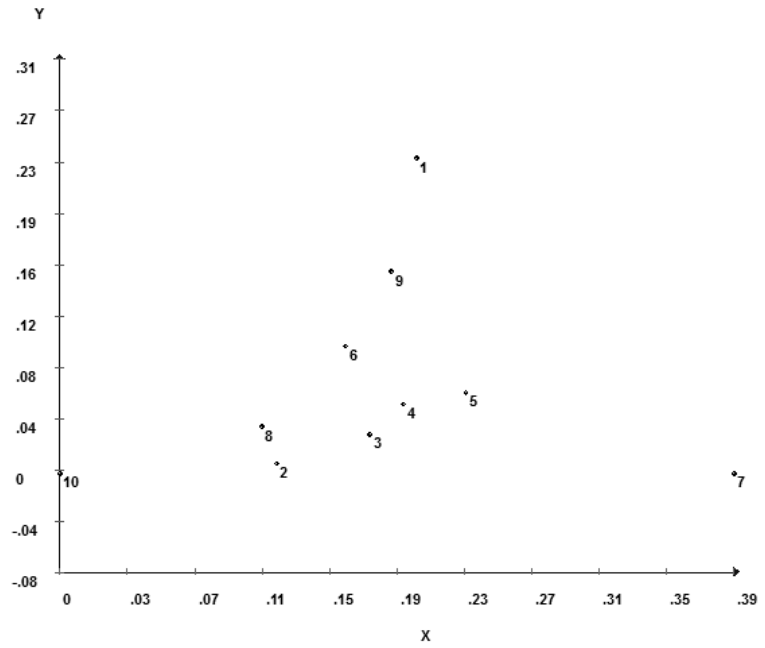


Figura 2: Projeção no plano bidimensional da matriz de dissimilaridade expresso pela Distancia de Nei, calculada a partir das informações genotípicas simuladas para seleção dos genitores.

Segundo Cruz & Carneiro (2006) a diversidade genética entre populações pode ser inferida de forma preditiva através da quantificação das diferenças presentes em características por medidas de dissimilaridade que possa expressar o grau de diversidade genética entre os genitores. A utilização de técnicas multivariadas para estimar a divergência genética, para fins de cruzamento visando explorar heterose e variabilidade em populações segregantes, tem se tornado comum, é empregada em vários trabalhos e em diversas culturas, tais como eucalipto (Scapim et al., 1999), milho (Melo, 2001) e feijão (Bonett et al., 2006; Ceolin et al., 2007; Elias et al., 2007; Ribeiro et al., 2001).

Uma vez identificadas as populações mais divergentes, designadas por P_1 e P_2, elas foram cruzadas gerando a população F_1estabelecida com um número de 100 indivíduos. A população F_1 foi cruzada com os genitores P_1 e P_2 separadamente constituindo a primeira geração de retrocruzamento avançada até a quinta geração em que cada um dos pais foi utilizado como genitor recorrente. Após o estabelecimento das 13 populações foram utilizadas para estudos de diversidade genética tendo-se a expectativa de que os métodos biométricos viesse refletir apropriadamente o padrão de similaridade, em especial com os genitores recorrentes e não-recorrente, com o avanço das gerações de retrocruzamento.

Novamente a dissimilaridade foi calculada utilizando a medida de distancia de Nei e tomado o seu complemento aritmético, que indica a similaridade entre as populações, conforme pode ser visualizado na Tabela 3.

Tabela 3: Índice de similaridade calculado pelo complemento aritmético da distancia de Nei (1972) entre genitores recorrentes P_1 e P_2, F_1 e os seus respectivos retrocruzamentos: RC1a, RC2a, RC3a, RC4a, RC5a, RC1b, RC2b, RC3b, RC4b, RC5b.

	P_2	F_1	RC1a	RC2a	RC3a	RC4a	RC5a	RC1b	RC2b	RC3b	RC4b	RC5b
P_1	0.549	0.902	0.975	0.994	0.998	0.999	0.999	0.749	0.663	0.606	0.581	0.567
P_2		0.894	0.760	0.661	0.604	0.576	0.560	0.977	0.993	0.998	0.998	0.998
F_1			0.975	0.943	0.922	0.912	0.906	0.968	0.938	0.916	0.907	0.902
RC1a				0.992	0.987	0.982	0.977	0.888	0.833	0.796	0.780	0.772
RC2a					0.998	0.995	0.994	0.823	0.753	0.707	0.686	0.676
RC3a						0.998	0.998	0.785	0.707	0.655	0.633	0.620
RC4a							0.999	0.767	0.686	0.631	0.608	0.594
RC5a								0.757	0.672	0.616	0.591	0.578
RC1b									0.994	0.986	0.981	0.980
RC2b										0.998	0.995	0.994
RC3b											0.999	0.998
RC4b												0.998

Conforme esperado o menor valor de similaridade genética foi encontrado entre os genitores P1 e P2 sendo de 0.5499. Apesar das populações não serem relacionadas existe uma quantidade de similaridade genética entre estes genitores.

A similaridade genética entre as gerações de retrocruzamento e seus genitores recorrentes apresentaram resultados coerentes com o que seria esperado pela literatura (Cruz et al. 2011) fundamentada nos pressupostos da contribuição gamética de cada genitor com o avanço das gerações. Em todos os métodos se espera observar posicionamento, em projeção ou em dendrogramas, da população F_1 no centro, delimitado por P_1 e P_2, e a maior aproximação dos retrocruzamentos “a” e “b”com os genitores recorrentes P_1 e P_2, à medida que se avança as gerações de retrocruzamentos.

Percebe-se, na Tabela 3, que no conjunto de retrocruzamento “a”, em que P_1 é o genitor recorrente, a maior similaridade encontrada foi de 0.9994 entre os RC4a e RC5a e a segunda foi de 0.9992 entre P_1 e RC5a. Conforme previsto, o processo de divisão celular, em gerações de retrocruzamento a similaridade genética aumenta com o aumento das gerações de retrocruzamento de forma que o resultado que se espera a partir de uma suposta contribuição gamética também é refletida na dissimilaridade

quantificada biometricamente a partir da segregação de 50 locos codominantes.

Para o conjunto de retrocruzamentos “a” o avanço da similaridade de F_1 a RC5, em relação ao genitor recorrente, foi de 0.9027, 0.9759, 0.9940, 0.9983, 0.9990 e 0,9992, respectivamente.

Ainda analisando a Tabela 3 percebe-se que, para o conjunto de retrocruzamento “b”, em que P_2 é o genitor recorrente a maior similaridade encontrada foi de 0.9990 entre os RC3b e RC4b e a segunda foi de 0.9989 entre P2 e RC4b e RC5b que apresentaram o mesmo valor. Corroborando também com a literatura a similaridade genética aumenta com o aumento das gerações de retrocruzamento 0,9027, 0.9775, 0.9937, 0.9981, 0.9989, 0.9989 considerando as gerações F_1 a RC5 em relação ao genitor P_2.

A projeção das distâncias em plano bidimensional com base na matriz de distâncias genéticas de Nei demonstra a separação dos genitores, que se localizam em pontos extremos do gráfico além da aproximação das populações de retrocruzamento RC1a, RC2a, RC3a, RC4a, RC5a e para o outro lado em RC1b, RC2b, RC3b, RC4b, RC5b relação ao genitor P_2 recorrente. A distorção encontrada foi de 24,57%, o estresse de 31,43% consideradas elevadas tendo em vista a referência de resultado satisfatório relatada por CRUZ et al., (2011). Entretanto, a correlação encontrada foi de 0,9633 demonstrando boa representação gráfica da matriz de distância. A Figura 3 ilustra o processo de recuperação do genoma recorrente pelo programa de retrocruzamentos.

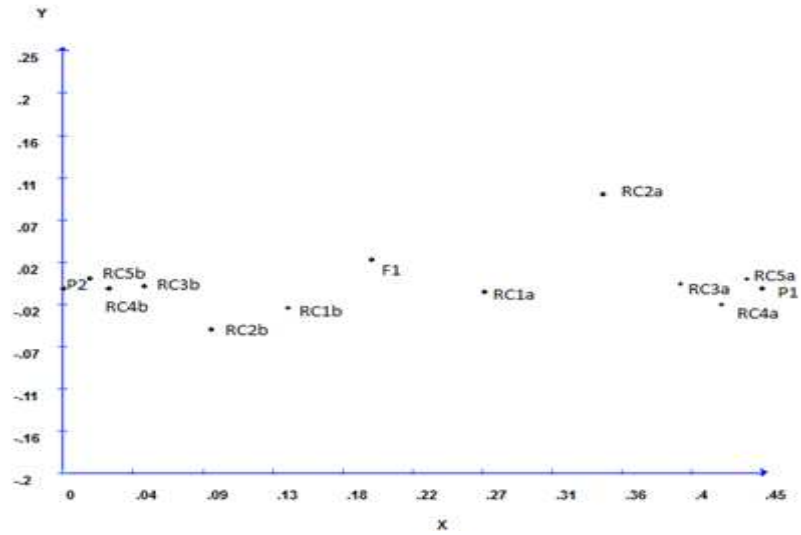


Figura 3: Projeção da dissimilaridade das 13 populações, expressa pela matriz de dissimilaridade de Nei, no espaço bidimensional.

Em RC4, teoricamente, espera-se uma recuperação do genoma recorrente de 96,875%, e em RC5 98,437%. Resultado semelhante do padrão de agrupamento foi encontrado com o emprego do método de Tocher também realizado utilizando-se as distâncias genéticas de Nei (1972). Este método tem como princípio manter a homogeneidade dentro e a heterogeneidade entre os grupos. Na Tabela 4 é visualizada a formação de três grupos distintos separando a população F_1 de seus genitores, e agrupando os genitores recorrentes com os respectivos retrocruzamentos "a" e "b". Percebe-se que as técnicas biométricas fundamentadas em informações de marcadores moleculares, que no presente trabalho refletem as informações da segregação de apenas 50 locos codominantes, em populações constituídas por 100 indivíduos, conseguem traduzir de forma satisfatória o padrão de similaridade que se espera entre as populações estabelecidas no delineamento genético estudado.

Tabela 4: Agrupamento de otimização obtidos pelos dados genotípicos simulados das 13 populações com base na dissimilaridade expressa pela distância de Nei.

Grupo	Populações
1	RC4a RC5a P_1 RC3a RC2a RC1a
2	RC3b RC4b RC5b P_2 RC2b RC1b
3	F_1

Resultados muitos similares foram obtidos usando técnicas de agrupamento hierárquicas conforme podem ser visualizados a partir dos resultados apresentados nas Figuras 4 e 5 referentes aos métodos de agrupamento hierárquico UPGMA e de vizinho mais próximo, respectivamente. Verificou-se, nas duas técnicas de agrupamento, a formação de um grupo contendo o genitor recorrente P_1 e seus retrocruzamentos "a" e outro contendo o genitor recorrente P_2 e os demais retrocruzamentos "b".

Para as populações em estudo, cuja estruturação é previamente conhecida tendo por base o delineamento genético de retrocruzamento, o método do vizinho mais próximo representou melhor as populações já que, no dendrograma que pode ser visualizado na Figura 5 a população F_1 ficou separada numa terceira ramificação ao corte ao nível de 75%, estabelecido pelo método de Mojema, e se agrupou com P_1 e o conjunto "a" de

retrocruzamentos. Já na análise do agrupamento hierárquico de ligação média entre grupos (UPGMA), a população F_1 ficou separada numa terceira ramificação ao corte a um nível inferior a 20% e se agrupou também com P_1 e o conjunto “a” de retrocruzamentos na Figura 4.

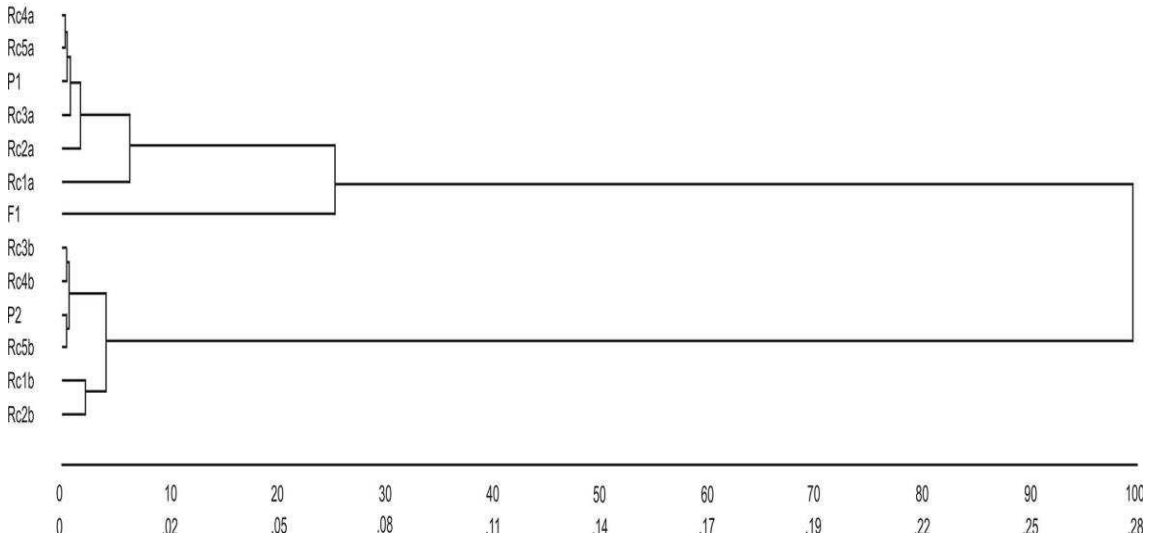


Figura 4: Dendrograma obtido pelo Método de ligação média entre grupos UPGMA baseados na distancia de Nei entre as 13 populações. A primeira linha do eixo da abscissa corresponde a valores percentuais em relação a e no ultimo nível de fusão.

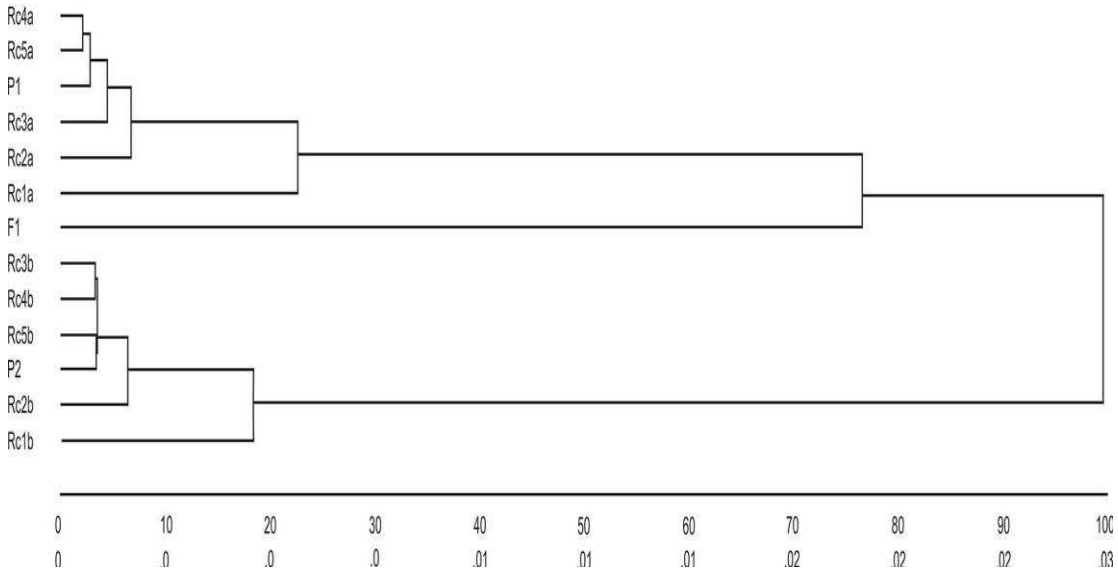


Figura 5: Dendrograma obtido pelo Método de ligação simples- vizinho mais próximo baseado na distancia de Nei entre as 13 populações. A primeira linha do eixo da abcissa corresponde a valores percentuais em relação a e no ultimo nível de fusão.

Pelos resultados encontrados algumas questões de ordem prática e metodológica devem ser suscitadas. Uma primeira abordagem se refere ao

emprego da simulação para estudos biométricos que neste estudo mostrou ser plenamente satisfatória. Esta simulação foi estabelecida com base em princípios estatísticos e biológicos, pois em cada população informações sobre marcadores moleculares foram geradas mantendo-se a estruturação básica de populações em equilíbrio de Hardy-Weinberg. As populações foram submetidas a "cruzamentos" *in silico* e, portanto, as particularidades da formação gamética, segregação de alelos, união aleatória de gametas, viabilidade e erros de amostragem devem ser coerentes com o sistema biológico manifestado entre plantas e animais. Preservar estas características em 13 populações relacionadas demonstra que o processo de simulação é extremamente eficaz em prover dados para valiosos estudos biológicos.

No contexto das técnicas biométricas empregadas é importante realçar que o uso de informações prévias para validar ou viabilizar a obtenção de importantes parâmetros no melhoramento genético é indispensável para o pesquisador e pode ser muito vantajoso em certos estudos comparativos de eficácia de metodologias. Informações prévias são disponíveis em muitos exemplos, de forma que a inferência sobre relacionamento entre populações de genitores e seus respectivos retrocruzamentos constitui um destes exemplos clássico. Neste caso é usual reconhecer o grau de similaridade, de maneira simples e efetiva, fundamentando-se no princípio de que há contribuição equitativa dos genitores para a formação da próxima geração. Nesta mesma linha de pensamento pode-se citar outros exemplos tais como a inferência sobre coeficiente de parentesco, coeficientes de endogamia ou a distribuição dos componentes de variância e dentro de populações estruturadas em famílias. Estas informações são amplamente empregadas admitindo ser o valor paramétrico conhecido e estabelecido por princípios fundamentais da genética quantitativa, mas, na prática, a mensuração destes parâmetros pode ser comprometida e sujeita a imprecisões tendo em vista tamanho populacional, ruídos ambientais e quantidade e qualidade das informações genéticas exploradas. Assim, comparar resultado esperado e observado, sob determinadas condições experimentais, torna-se ótima estratégia para medir eficácia de técnicas alternativas de análise.

Alguns trabalhos têm relatado que diferentes medidas de dissimilaridade e diferentes técnicas de agrupamentos podem proporcionar diferentes partições de um conjunto de populações. Ainda outros trabalhos acrescentam

que diferentes partições de um mesmo conjunto de dados podem ser, sob determinado critério, igualmente satisfatórias (Pereira, 1999). Entretanto, no melhoramento genético o objetivo ao se estudar a diversidade genética geralmente é comum, de forma que os relacionamentos ou distanciamentos genéticos entre partes de genótipos devem refletir prioritariamente a diversidade para fins de orientar cruzamentos, obter máxima heterose na população híbrida e ampla variabilidade na população segregante.

Assim, no presente estudo foi possível reconhecer o potencial das técnicas biométricas, aplicadas em um conjunto de populações estruturadas, com tamanho finito de indivíduos e de informações moleculares, em reconhecer a estrutura hierárquica da similaridade, a proximidade e o distanciamento genotípico. Também foi possível reconhecer a superioridade de resultado obtido por uma técnica em relação a outras concorrentes.

3.2. Análise de dados fenotípicos

Inferir sobre a diferenciação entre populações é assunto de grande relevância tanto sob aspecto do melhoramento genético quanto na adaptação e evolução de grupos populacionais. Fatores como seleção, fluxo gênico, amostragem, dentre outros, podem contribuir para ampliar ou reduzir a diversidade entre populações e demandam estratégias diferenciadas para situações particulares. Deve-se, portanto, inferir com acurácia satisfatória sobre a diversidade entre o material estudado para que as decisões adotadas sejam as mais acertadas.

Neste estudo, à semelhança do tópico anterior, diferentes técnicas multivariadas foram utilizadas para fins de retratar o padrão de similaridade entre as populações estudadas. Entretanto, é considerado, para fins de estudo, valores fenotípicos resultante da ação genotípica acrescidos de efeitos ambientais que normalmente atuam como agentes perturbadores impossibilitando ou dificultando inferir com acurácia sobre o verdadeiro grau de relacionamento entre as populações estudadas. As técnicas biométricas foram aplicadas a dois diferentes Cenários denominados A e B em que o componente ambiental foi progressivamente acrescentado para enriquecimento do presente estudo. O cenário A, foi caracterizado por incluir seis características quantitativas de alta herdabilidade que variam de 0,55 a 0,8, e o cenário B, que

é representado por sete características quantitativas de baixa herdabilidade que variam de 0,2, a 0,5.

Cenário A – Características de Alta Herdabilidade

O efeito ambiental é um dos fatores que devem ser minimizados para que as informações sejam obtidas com maior precisão e as decisões possam ser tomadas de maneira mais acertada pelos melhoristas. Neste trabalho considerou ser apropriado inferir sobre a média fenotípica de cada população que é uma medida mais precisa do potencial genético desta população. Também foi computado o valor da variabilidade dentro da população que, neste caso inclui variações genéticas e não genéticas, mas que poderiam ser úteis como forma de ponderar a importância relativa de cada variável sobre a diversidade genética, sendo tal procedimento utilizado em avaliações de diversidade entre populações naturais. Desta forma, tomou-se a distância generalizada de Mahalanobis como medida de referencia da diversidade entre pares de populações.

Pela Tabela 5 pode ser analisado os elementos da matriz de dissimilaridade de Mahalanobis, representado pelas características de alta herdabilidade, no retrocruzamentos “a” em que P_1 é o genitor recorrente a menor dissimilaridade encontrada foi de 0.0288 entre os RC4a e RC5a e a segunda foi de 0,332 entre P_1 e RC3a .

No retrocruzamento “b”, em que P_2 é o genitor recorrente, a menor dissimilaridade encontrada foi de 0,0238 entre os RC2b e RC3b e a segunda foi de 0.0315 entre P_2 e RC4b e RC5b que apresentaram o mesmo valor.

Conforme esperado, tendo em vista a seleção prévia de genitores, o menor valor de similaridade genética encontrado foi entre os genitores P_1 e P_2. Novamente confrontando a expectativa de diversidade, tendo em vista a contribuição meiótica dos genitores, e aquela obtida com emprego de técnicas biométricas, sobre o conjunto limitado de características, de indivíduos e de efeitos ambientais, pode-se observar alto grau de coerência em que a similaridade genética deve aumentar com o aumento das gerações (Tabela 5).

Tabela 5: Matriz de dissimilaridade de calculada por meio da distância generalizada de Mahalanobis. Foram consideradas as informações fenotípicas das características simuladas de alta herdabilidade para as populações P1 e P2, F1 e os seus respectivos retrocruzamentos: RC1a, RC2a, RC3a, RC4a, RC5a, RC1b, RC2b, RC3b, RC4b, RC5b.

	P_2	F1	RC1a	RC2a	RC3a	RC4a	RC5a	RC1b	RC2b	RC3b	RC4b	RC5b
P_1	6.64	5.933	1.394	0.189	0.033	0.129	0.085	5.897	5.721	6.415	5.908	5.258
P_2		0.739	2.413	4.754	5.899	5.665	5.755	0.323	0.128	0.058	0.045	0.117
F_1			1.747	4.144	5.352	4.813	4.901	0.151	0.292	0.601	0.724	0.842
RC1a				0.640	1.105	0.846	0.905	1.774	1.708	2.193	2.015	1.691
RC2a					0.091	0.137	0.076	4.069	3.966	4.549	4.149	3.620
RC3a						0.094	0.053	5.237	5.064	5.676	5.212	4.602
RC4a							0.029	4.864	4.708	5.413	5.031	4.452
RC5a								4.928	4.816	5.514	5.112	4.524
RC1b									0.092	0.188	0.295	0.426
RC2b										0.092	0.119	0.179
RC3b											0.049	0.163
RC4b												0.046

O padrão de agrupamento está ilustrado na Figura 6 que mostra a representação da raiz quadrada das distâncias de Mahalanobis, em gráfico bidimensional. A técnica biométrica refletiu apropriadamente a medida de dissimilaridade aglomerando as populações mais similares, ou seja, gerações de retrocruzamento e seus genitores recorrentes. É possível observar o RC1a mais distante das gerações RC do tipo "a" e aproximação maior do F_1 com o P_2 e os seus retrocruzamentos. Essa representação apresentou nível de estresse (S) provocado por simplificação da projeção no espaço bidimensional de 2,75% e a distorção de 2,15% valores bastante inferiores aos mínimos de aceitabilidade da técnica que é de 20% (Cruz et al., 2011). A correlação foi de 0,999 o que demonstra a eficiência da técnica em preservar as reais distâncias entre os pares de genótipos avaliados.

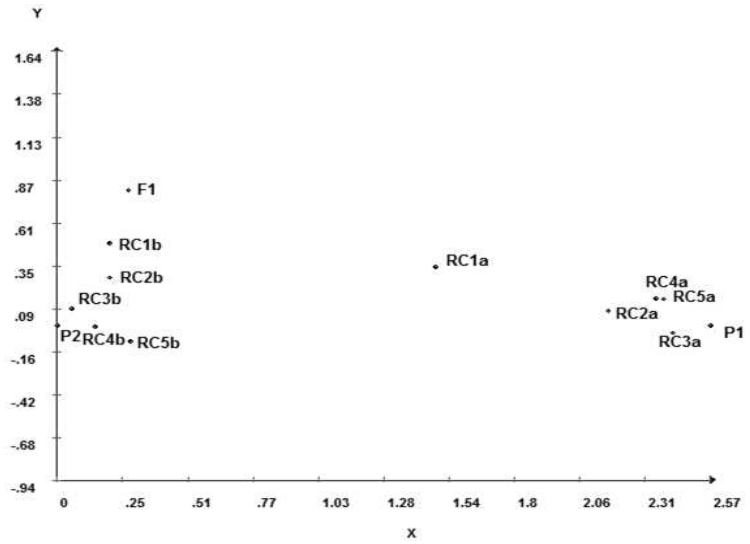


Figura 6: Projeção da dissimilaridade das 13 populações, expressa pela matriz de dissimilaridade de Mahalanobis, no espaço bidimensional **.**

O agrupamento dos genótipos pelo método hierárquico do vizinho mais próximo (Figura 7), UPGMA (Figura 8) e o método de otimização de Tocher (Tabela 6) foram similares na formação de grupos entre populações mais divergentes, mas de certa forma com resultado inapropriado tendo em vista a expectativa meiótica. A concordância entre estas técnicas mostrou que os genótipos pertencentes aos grupos I (P_1 e RC2a, RC3a, RC4a, RC5a), II (P2 e os retrocruzamentos "b" e F_1,), III (RC1a) encontrados por Tocher terem sido os mesmos dos agrupamentos de maior distância pelo vizinho mais próximo que, em um corte ao nível de 25% separou o genótipo RC1a antes de separar F1 de P2 e os retrocruzamentos "b" (16,38%). Segundo Everitt (1977) apud Pereira (1999), o método do vizinho mais próximo possui dificuldade em separar grupos que apresentem pontos intermediários, e ressalta que a escolha do melhor método depende do conhecimento do autor sobre a estrutura dos dados.

Tabela 6: Método de Otimização obtidos pelos dados fenotípicos simulados das 13 populações em Cenário A com base na dissimilaridade expressa pela distância Mahalanobis.

Grupo	Populações
1	RC4a RC5a RC3a P1RC2a
2	P2RC4b RC3b RC5b RC2b RC1b F_1
3	RC1a

Na análise de agrupamento pelo método do vizinho mais próximo (Figura 7), constata-se que os resultados distorceram o real padrão de agrupamento tendo em vista a eficácia da técnica em minimizar os efeitos ambientais indesejáveis. Houve a formação de dois grupos dos conjuntos respectivos de genitores e seus retrocruzamentos "a" e "b". Verifica-se também que a população F_1 ficou separada numa terceira ramificação ao nível de fusão de apenas 10%, enquanto RC1a se separou ao nível de 40%, corroborando com resultados encontrados pelos outros métodos. Entretanto, o mesmo nem sempre separou as populações em concordância com as distancias observadas na matriz de dissimilaridade. O coeficiente de correlação cofenética estimado por meio desse método foi de 0,9032. Esta estimativa foi semelhante à obtida por meio do método UPGMA, que por princípio estatístico é o que apresenta maior magnitude desta correlação. Assim, este valor pode ser

considerado alto (Cruz et al., 2011), não trazendo prejuízos significativos para no processo de representação das distâncias entre as populações. Esses resultados indicam que o método do vizinho mais próximo é afetado por efeitos perturbadores ambientais e proporciona resultados dúbios para representar, nesse grau de dissimilaridade, a separação das populações.

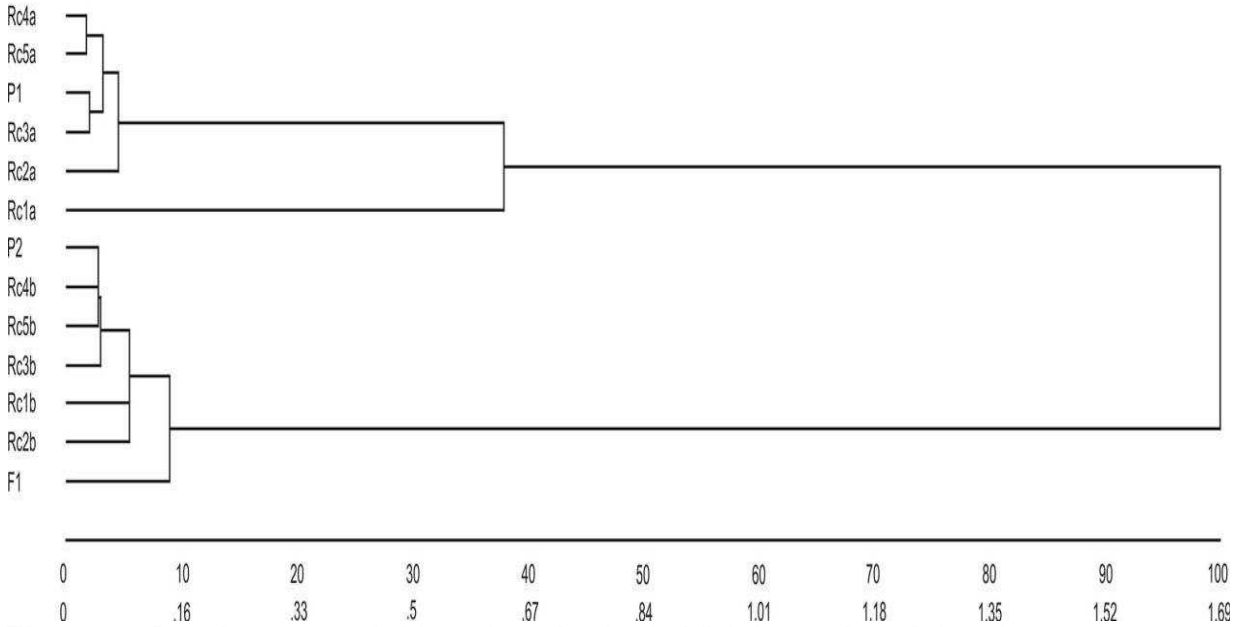
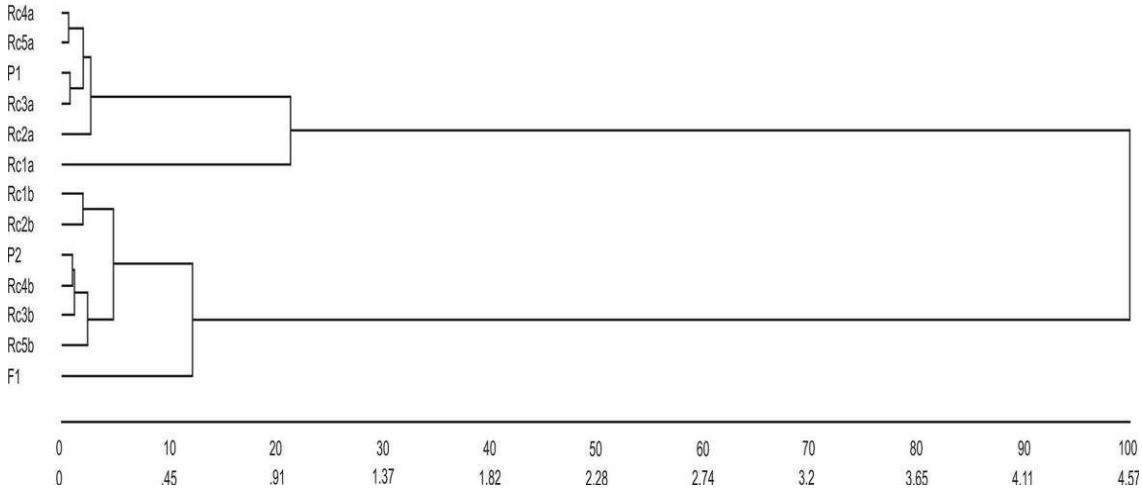


Figura 7: Dendrograma obtido pelo método do Vizinho mais próximo baseados na distancia de Mahalanobis entre as 13 populações simuladas no Cenário A. A primeira linha do eixo da abscissa corresponde a valores percentuais em relação ao ultimo nível de fusão.

O dendrograma gerado pelo método UPGMA apresentou valor cofenético de 0.9068 o que mostra fidelidade na representação do conjunto de dados da matriz de distancias original, pois quanto maior o valor do coeficiente, menor é a distorção provocada ao se agrupar os genótipos (Bussab et al.,1990). **Figura 8:** Dendrograma obtido pelo Método de ligação média entre Grupos



UPGMA baseados na distancia de Mahalanobis entre as 13 populações simuladas no Cenário A. A primeira linha do eixo da abscissa corresponde a valores percentuais em relação ao ultimo nível de fusão.

O emprego do algoritmo UPGMA – Unweighted Pair Group Method with Arithmetic Mean – para cálculo dos níveis de fusão e representação das distâncias por meio de diagrama em árvore tem sido exaustivamente adotado por pesquisadores das diferentes áreas da ciência (Oliveira et al., 2002; Scheffer-Basso, 2002; Silva 2012), tendo sua superioridade com relação a outros métodos de agrupamentos hierárquicos constatada por Dias (1998) tomando por base critérios estatísticos como, geralmente, o coeficiente de correlação cofenético. Este método, juntamente com as projeções bidimensionais, pode ser considerado como boa opção de análise de dados na literatura, porém ficou demonstrado que a eficácia destas técnicas pode ser comprometida pela ação de efeitos ambientais.

Cenários B– Características de Baixa Herdabilidade

O uso do cenário, representado pela análise de características de baixa herdabilidade, permite avaliar a eficácia das técnicas em minimizar efeitos ambientais em condições em que os efeitos ambientais são mais pronunciados que os efeitos genéticos.

Novamente foram calculadas as medidas de dissimilaridade expressas pela distância generalizada de Mahalanobis. Pela Tabela 7 constata-se que no retrocruzamento “a”, em que P_1 é o genitor recorrente, a menor dissimilaridade encontrada foi de 0.0382 entre os RC3a e RC4a e a segunda foi de 0.0942 entre P_1 e RC5a. Pequena diversidade entre gerações avançadas e genitor recorrente é esperada em programas de retrocruzamento.

No retrocruzamento “b”, em que P_2 é o genitor recorrente, a menor similaridade encontrada foi de 0.0238 entre os RC2b e RC3b e a segunda foi de 0.0315 entre P_2 e RC4b. Conforme previsto por princípios de meiose e contribuição gamética equitativa entre genitores, a similaridade genética com o genitor recorrente deve aumentar com o aumento das gerações de retrocruzamento. Entretanto, observa-se que, para as características fenotípicas de baixa herdabilidade, houve pequenas variações se comparados aos dados genotípicos.

O retrocruzamentos RC5b apresentou valor de dissimilaridade 0,06 um pouco maior que o RC4b 0,0315 e o RC3a apresentou a menor dissimilaridade em relação a P_1. De acordo com o esperado a maior distancia genética foi

encontrada entre os genitores P_1 e P_2 sendo de 6,23. Segundo Benin et al. (2002), a escolha de genitores, deve ser baseadas na magnitude de suas dissimilaridades e também no potencial *per se* dos genitores. Os genótipos reunidos em grupos mais distantes dão um indicativo de serem dissimilares, podendo ser considerados como promissores em cruzamentos artificiais.

Tabela 7: Matriz de dissimilaridade de calculada por meio da distância generalizada de Mahalanobis. Foram consideradas as informações fenotípicas das características simuladas de baixa herdabilidade para as populações P_1 e P_2, F_1 e os seus respectivos retrocruzamentos: RC1a, RC2a, RC3a, RC4a, RC5a, RC1b, RC2b, RC3b, RC4b, RC5b.

	P_2	F1	RC1a	RC2a	RC3a	RC4a	RC5a	RC1b	RC2b	RC3b	RC4b	RC5b
P_1	6.23	5.26	1.59	0.34	0.13	0.12	0.09	4.94	6.59	6.38	5.95	5.82
P_2		0.83	1.91	3.79	4.71	5.17	5.84	0.42	0.07	0.07	0.03	0.06
F_1			1.12	3.02	3.90	4.12	4.72	0.55	0.69	0.62	0.91	0.90
RC1a				0.52	0.93	1.07	1.40	1.30	1.99	1.87	1.84	1.78
RC2a					0.07	0.12	0.27	2.74	4.05	3.87	3.59	3.47
RC3a						0.04	0.10	3.54	5.05	4.83	4.47	4.34
RC4a							0.06	3.83	5.46	5.23	4.93	4.78
RC5a								4.47	6.14	5.92	5.60	5.43
RC1b									0.56	0.50	0.36	0.25
RC2b										0.02	0.16	0.20
RC3b											0.15	0.21
RC4b												0.06
RC5b												

Na Figura 9 a dispersão gráfica dos elementos de diversidade da matriz de distâncias genéticas de Mahalanobis demonstra a separação dos genitores, que se localizam em pontos extremos do gráfico e a aproximação das populações de retrocruzamento RC1a, RC2a, RC3a, RC4a, RC5a e o genitor recorrente P_1 e para o outro lado em RC1b, RC2b, RC3b, RC4b, RC5b em relação ao genitor P_2 recorrente. Entretanto, nos dados fenotípicos F_1 se agrupou mais próxima a P_2 e seus retrocruzamentos, não se concentrando a centro da projeção. A distorção foi de 4,89%, o estresse de 6,13% e uma correlação de 99,77%, consideradas como satisfatória por Cruz et al., (2011). Resultados semelhantes foram observados no trabalho de Fonseca et al., (2009) que, ao estudar a diversidade genética em populações de retrocruzamentos de maracujá, verificaram, ao comparar a dispersão gráfica baseada na matriz de distâncias genéticas, de plantas RC5 ou RC4 e seus genitores observaram a maior aproximação das plantas RC5 em relação ao genitor recorrente, além da maior uniformidade de distribuição dessas plantas e o genitor doador no outro ponto extremo.

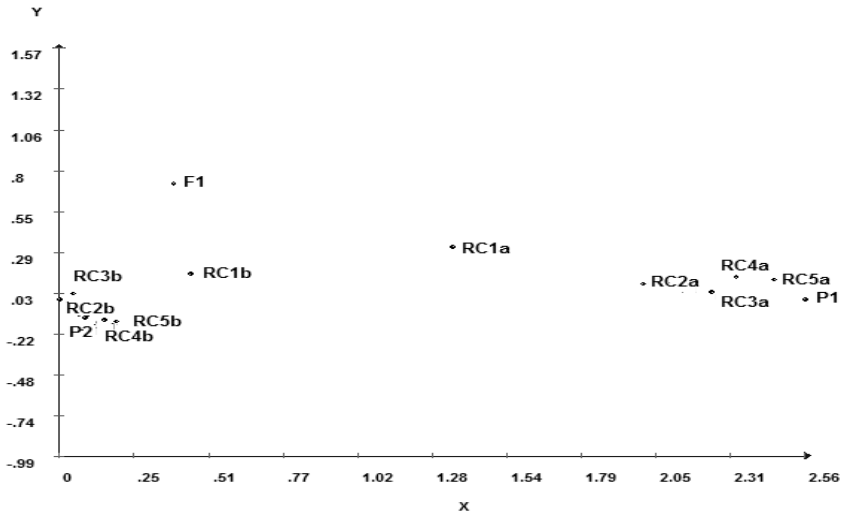


Figura 9: Projeção da dissimilaridade das 13 populações, expressa pela raiz quadrada da matriz de dissimilaridade de Mahalanobis, no espaço bidimensional.

O agrupamento das populações pelo método de Tocher (Tabela 8) foi realizado utilizando-se as distâncias genéticas de Mahalanobis. Este método tem como princípio manter a homogeneidade dentro e heterogeneidade entre os grupos, e possibilitou a reunião das populações em quatro grupos distintos separando a população, RC1a, F_1, e agrupamento os genitores recorrentes com os respectivos retrocruzamentos.

A separação dos três grandes grupos representada pelos genitores recorrentes P_1 e P_2, e os retrocruzamentos "a" e "b", e de F_1 foi mascarada nas análises de agrupamentos de otimização (Tabela 8). Os efeitos ambientais promoveram uma diversidade adicional que não foi removida pelas técnicas biométricas.

Tabela 8: Agrupamento de Otimização obtidos pelos dados fenotípicos simulados das 13 populações em Cenário B com base na dissimilaridade expressa pela distância Mahalanobis.

Grupo	Populações
1	RC4a RC5a P_1 RC3a RC2a
2	RC3b RC4b RC5b P_2 RC2b RC1b
3	RC1a
4	F_1

Verifica-se, na análise de agrupamento hierárquico do vizinho mais próximo (Figura 10), a formação de dois grupos dos conjuntos respectivos de genitores e seus retrocruzamentos "a" e "b". Verifica-se também que a

população F1 ficou separada numa terceira ramificação ao corte ao nível de 50% e se agrupou com P2 e o conjunto “b” de retrocruzamentos, o que era esperado pela dissimilaridade encontrada entre P1 e F1, 5,2626, ser bem superior a de P2 e F1 de 0,8290. O método foi condizente com a dissimilaridade esperada tendo em vista o delineamento genético utilizado.

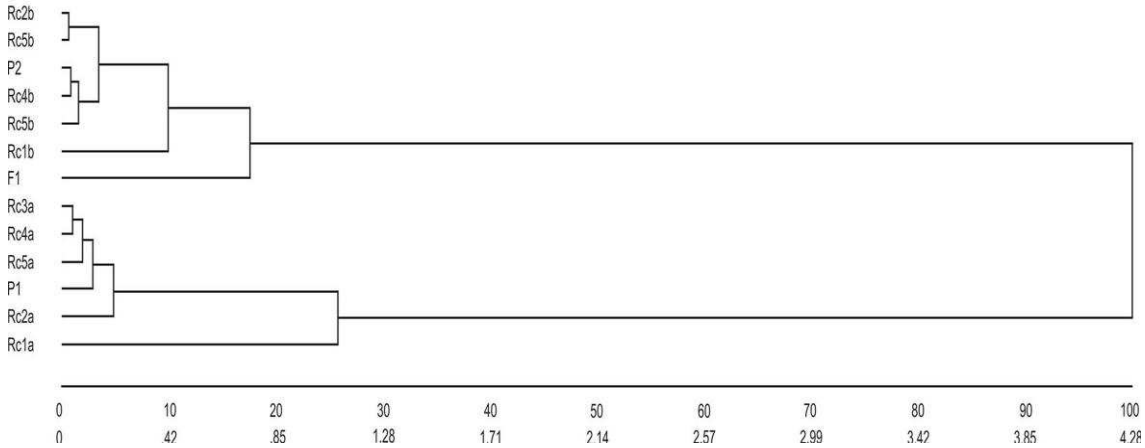


Figura 10: Dendrograma obtido pelo método do Vizinho mais próximo baseados na distancia de Mahalanobis entre as 13 populações simuladas no Cenário B. A primeira linha do eixo da abscissa corresponde a valores percentuais em relação a dissimilaridade no ultimo nível de fusão 1,12.

Na análise do agrupamento hierárquico de Ligação Média entre grupos UPGMA (Figura 11), também houve a formação de dois grupos dos conjuntos respectivos de genitores e seus retrocruzamentos “a” e “b”. Entretanto, a população F_1 ficou separada numa terceira ramificação ao corte ao nível inferior a 20% e se agrupou com P_2 e o conjunto “b” de retrocruzamentos.

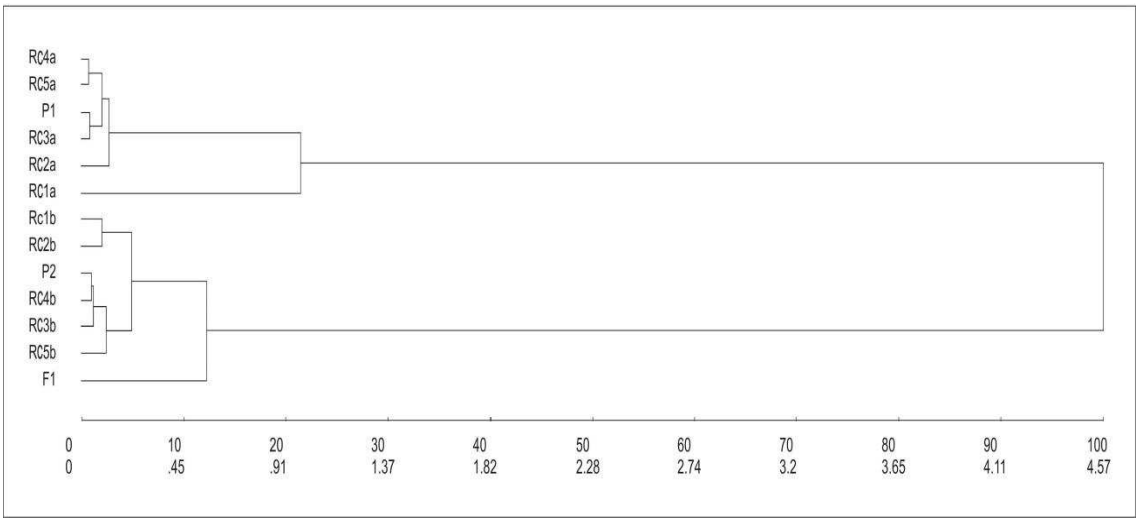


Figura 11: Dendrograma obtido pelo Método de ligação média entre Grupos UPGMA baseados na distancia de Mahalanobis entre as 13 populações simuladas no Cenário B. A primeira linha do eixo da abcissa corresponde a valores percentuais em relação ao ultimo nível de fusão 4,28.

A não concordância do padrão de agrupamento em relação aos métodos de agrupamento hierárquicos foi relatado por Moraes et al. (2005) em estudo de divergência genética entre genitores de soja para uso em programas de retrocruzamentos, sendo que no método de UPGMA observaram a formação de cinco grupos e no de Tocher nove.

No método de UPGMA para os dados fenotípicos do presente estudo em ambos os cenários “A” e “B” pode-se observar que os retrocruzamentos se agruparam de forma que os genitores e suas gerações mais próximas de RCs estivessem sempre bem próximos, e que dois grandes grupos fossem formados. Essa constatação também foi possível, para o método de Tocher, que agrupou os retrocruzamentos do conjunto A, com exceção do RC1a e os retrocruzamentos do conjunto “b” sempre no mesmo grupo, significando que esses genótipos são muito próximos geneticamente. Entretanto, para a população F_1 houve variação quanto a formação de grupos, pois este nem sempre esteve formando um grupo individual, o que indicou que ele apresenta-se mais próximos de P_2 e conjunto B de retrocruzamentos para os dados fenotípicos.

Conforme Rizzo & Braz (2002), o que se pode esperar é que indivíduos constantes em um mesmo grupo sejam similares e, dessa forma, o conjunto de retrocruzamentos “b” bem como os retrocruzamentos do conjunto “a”, exceto RC1a, ficaram sempre em grupos semelhantes indicando boa representação dos método em geral na formação dos grupos.

A relação entre os grupos formados e a variabilidade genética existente na separação de genitores já foi verificada por Cintra et al. (2005), estudando divergência genética entre acessos de *Curcumalonga* L., classificaram 21 acessos em cinco grupos usando a distância generalizada de Mahalanobis, o algoritmo de Tocher e o método do Vizinho Mais Próximo que consideraram uma análise efetiva na constatação de variabilidade. Em estudo de divergência em alface, Oliveira et al. (2004) reuniram 17 cultivares em apenas três grupos pelo método de Tocher. Chioratto et al. (2005) agruparam 993 acessos de feijoeiro em 45 grupos, sendo que 92,3% dos acessos estudados formaram os primeiros oito grupos. Arriel et al. (2005), estudando 39 matrizes de *Cnidoscolus phyllacanthus,* observaram grande divergência genética entre elas, já que estas formaram nove grupos. Dentre 36 acessos de taro, estudados por Pereira et al. (2004) usando métodos de distância generalizada

de Mahalanobis e de otimização de Tocher, houve a formação de seis grupos divergentes, tendo um dos grupos 29 acessos, e dispersão dos demais em grupos diversos, o que demonstrou uma ampla diversidade dos genótipos.

A dissimilaridade entre indivíduos, grupos de indivíduos ou populações tem sido medida por um método específico ou uma combinação de métodos, a partir de diferentes grupos de dados (Mohammadi et al., 2004). Nesse contexto, os estudos de diversidade genética têm se baseado em análises moleculares e morfológicas entre as ferramentas mais utilizadas e informativas para a estimação da diversidade (Máric et al., 2004), contribuindo substancialmente nas diferentes etapas dos programas de melhoramento, por permitir a determinação das singularidades e diferenças em relação à constituição genética e fenotípica de genótipos (Franco et al., 2001).

A expressão fenotípica, por ser influenciada por fatores externos, tais como condições ambientais, têm sido considerada de baixa acurácia para os estudos genéticos a partir de caracteres fenotípicos (Vieira et al., 2007). Já os estudos baseados em métodos moleculares têm se tornado cada vez mais comuns, testemunhando grandes avanços em programas de melhoramento. No entanto, a análise dessas duas categorias de dados separadamente pode resultar em inferências fragmentadas e, muitas vezes, imprecisas, dificultando a compreensão das relações genéticas entre os materiais estudados.

A relação entre o numero de grupos formados e a genealogia das progênies de retrocruzamentos também não foi observada por Ramos (2010) que estudou trinta e duas progênies de retrocruzamento de mamão que foram avaliadas com base em quinze características morfoagronômicas, além da análise de 20 iniciadores ISSR e 19 RAPD. Foi constatado que essas duas análises diferem quanto à coerência entre os grupos formados e a genealogia das progênies. Para o presente estudo, a análise da divergência genética utilizando-se métodos de agrupamentos de UPGMA e o do Vizinho Mais Próximo a formação dos grupos nem sempre apresentou boa concordância com a genealogia das progênies avaliadas com base nos dados fenotípicos quantitativos e mais foi apropriadamente constatada nos dados moleculares.

Segundo Everitt 1977 apud Pereira (1999), muitos dos problemas encontrados na utilização de análises de agrupamento são devido aos critérios que devem ser utilizados pelos pesquisadores na escolha da medida da distancia ou do coeficiente de (di)similaridade, da técnica de agrupamento a ser

utilizado pois os mesmos devem ser escolhidos com base em conhecimentos prévios dos dados, devido a isso, não é possível recomendar um ou outro método, pois a escolha do melhor método depende do conhecimento do autor sobre a estrutura dos dados.

Segundo Sudré et al. (2005), é importante aplicar mais de um método de agrupamento, visando a coerência dos resultados apresentados. E para os métodos utilizados houve boa concordância entre os agrupamentos de UPGMA, Vizinho mais próximo, Tocher e Projeção de distancias, para esses dados fenotípicos, embora nem sempre tenha conseguido representar a genealogia da forma esperada.

Por tudo isso, diante dos resultados expostos, é possível constatar que houve boa concordância dos dados, já que os métodos são bastante sensíveis as particularidades destes. Nota-se que, embora ambos os agrupamentos não sejam 100% coerentes com genealogia, devido a divergência no agrupamento de F_1 e RC1a que não foi esperada, ou seja, não possibilitaram o agrupamento de todas as progênies de acordo com a geração de retrocruzamento, a análise baseada nos dados fenotípicos se aproximou bem do esperado.

3.3. CONCLUSÃO

As técnicas de agrupamento hierárquico de vizinho mais próximo, Tocher, UPGMA e de projeção de distâncias no plano bidimensional foram eficientes em agrupar e descrever o padrão de similaridade genotípica entre populações considerando os diversos níveis de similaridade.

A estrutura de similaridade genética populacional, delineada em populações derivadas de retrocruzamento, tem forte influência ambiental, mas pode ser apropriadamente representada por meio de muitas técnicas multivariadas de UPGMA, Projeção 2d e Tocher aplicadas em dados fenotípicos.

Conjuntos de dados simulados podem ser apropriadamente gerados com potencial de uso diverso, em especial para análises de diversidade em populações delineadas de retrocruzamentos que demanda grande quantidade de informações para fins de treinamento e aprendizagem.

Os dados apropriadamente simulados, por preservarem informações essenciais, podem agregar ou substituir dados históricos algumas vezes não tão facilmente disponíveis.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, R.M.; GARCIA, A.A.F.; CRUZ, E.D.; FIGUEIRA, A. Seleção de descritores botânico-agronômicos para caracterização de germoplasma de cupuaçuzeiro. **Pesquisa Agropecuária Brasileira**, v.38, p.807-818, 2003.

ARAUJO, D.G. de; CARVALHO, S.P.; ALVES, R.M. Divergência genética entre clones de cupuaçuzeiro (*Theobroma grandiflorum* Willd. ex Spreng. Schum.). **Ciência e Agrotecnologia**, v.26, p.13-21, 2002.

BARBOSA, C. D.; VIANA, A. P.; QUINTAL, S. S. R.; PEREIRA, M. G. Artificial neural network analysis of genetic diversity in *Carica papaya* L. **Crop Breeding and Applied Biotechnology**,v. 11, n. 3, p. 224-231, 2011.

BENIN, G.; CARVALHO, F. I. de; ASSMANN, I. C.; CIGOLINI, J; CRUZ, P. J; MARCHIORO, V. S.; LORENCETTI, C.; SILVA, J. A. G. Identificação da dissimilaridade genética entre genótipos de feijoeiro comum (Phaseolus vulgaris L.) do grupo preto. **Revista Brasileira de Agrociência**, Pelotas, v. 8, n. 3, p.179-184, 2002.

BHERING,L.L.**Mapeamento genético em famílas simuladas de irmãos completos.** Tese (Doutorado em Genética e Melhoramento) – Universidade Federal de Viçosa, 166p, 2008.

BONETT, L.P.; GONÇALVES-VIDIGAL, M.C.; SCHUELTER, A.R.; VIDIGAL FILHO, P.S.; GONELA, A.; LACANALLO, G.F. Divergência genética em germoplasma de feijoeiro comum coletado no Estado do Paraná, Brasil. **Semina**: Ciências Agrárias, v.27, p.547-560, 2006.

BORÉM, A.; Miranda, G. V. 2009. **Melhoramento de Plantas.** 5ta ed. Editora UFV, Viçosa, 2009, 523p.

BUSSAB, W.O., MIAZAKI, E.S., ANDRADE, D.F. **Introdução à análise de agrupamentos**. IME-US, 1990.

CEOLIN, A.C.G.; GONÇALVES-VIDIGAL, M.C.; VIDIGAL FILHO, P.S.; KVITSCHAL, M.V.; GONELA, A.; SCAPIM, C.A. Genetic divergence of the common bean (*Phaseolus vulgaris* L.) group Carioca using morpho-agronomic traits by multivariate analysis. **Hereditas**, v.144, p.1-9, 2007.

CHIORATO, A. F.; Carbonell, S. A. M.; Dias, L. A. S.; Resende, M. D. V. (2008) Prediction of genotypic values and estimation of genetic parameters in common bean. **Brazilian Archives of Biology and Technology***,* 51: 465-472.

CINTRA, M.M.D.F. et al. Genetic divergence among Curcuma longa L. accessions. **Crop Breeding and Applied Biotechnology**, v.5, p.410-417, 2005.

CORRÊA, F.J.C. **Avaliação de métodos de seleção tradicionais, assistida por marcadores moleculares e por genes candidatos, com dados**

simulados. Viçosa, MG: UFV, 2001, 54p. Tese (Mestrado em Zootecnia) Universidade Federal de Viçosa, 2001.

CRUZ, C.D. 2013. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**. 35: 271-276.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético ed.5**. Viçosa, UFV**.** 480 p. 2012.

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

Cruz, C. D. (2008) **Programa GENES: Diversidade genética**. Universidade Federal de Viçosa, Viçosa.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos Biométricos Aplicados ao Melhoramento Genético** - Vol 2. 2.ed. Viçosa: UFV, 2006. 585p.

CRUZ, C.D. Programa GENES: Versão Windows, **Aplicativo Computacional em Genética e Estatística, Análise Multivariada e Simulação**. Viçosa: UFV, 2006. 175 p.

DIAS, L.A. dos S.; KAGEYAMA, P.Y.; CASTRO, G.C.T. Divergência genética multivariada na preservação de germoplasma de cacau (Theobromacacao L.). **Agrotrópica,** v.9, p.29-40, 1997.

ELIAS, H. T.; VIDIGAL, M. C. G.; GONELA, A.and VOGT, G. A. Variabilidade genética em germoplasma tradicional de feijão-preto em Santa Catarina. **Pesq. agropec. bras.**2007, vol.42, n.10, pp. 1443-1449.

EVERITT, B.S. The Analysis of Contingency Tables.Wiley.New York, 1977.

FALCONER, D.S.; MACKAY, T.F.C. **Introduction to quantitative genetics.**4.ed. Harlow: LongmanGroup Ltda,1996. 464p.

FALEIRO, F.G.; JUNQUEIRA, N.T.V.; BELLON, G.; KRALH, L.L.; ANJOS, J.R.N.; PEIXOTO, J.R.; BRAGA, M.F.; REZENDE, A.M. Utilização de marcadores moleculares em retrocruzamentos visando a resistência do maracujazeiro-azedo a múltiplas doenças. In: **CONGRESSO BRASILEIRO DE FITOPATOLOGIA**, 36., 2004ª , Gramado. Resumos... p. S325.

FONSECA, K.G.; FALEIRO, F.G.; JUNQUEIRA, N.T.V.; PEIXOTO, J.R.; BELLON, G.; JUNQUEIRA, K.P.; SANTOS, E.C. Análise da recuperação do genoma recorrente em maracujazeiro-azedo com base em marcadores RAPD. **Revista Brasileira de Fruticultura**, Jaboticabal, v. 31, n. 1, p. 145-153, 2009.

FRANCO, F.; CROSSA, J.; RIBAUT, J. M.; BETRAN, J.; WARBURTON, M. L.; KHAIRALLAH, M.A method for combining molecular markersandphenotypicattributes for classifyingplantgenotypes. **Theoretical and Applied Genetic**, 103: 944-952, 2001.

MÁRIC, S.; BOLARIC, S.; MARTINCIC, J.; PEJIC, I.; KOZUMPLIK, V. (2004) Genetic diversity of hexaploid wheat cultivars estimated by RAPD markers, morphological traits and coefficients of parentage. **PlantBreeding**, 123: 366-369.

MELO, W.M.C.; PINHO, R.G.V.; FERREIRA, D.F. Capacidade combinatória e divergência genética em híbridos comerciais de milho. **Ciência e Agrotécnica**, v.25, p.821-830, 2001.

MOHAMMADI, S. A.; PRASANNA, B. M. (2003) Analysis of genetic diversity in crop plants – Salient statistical tools and considerations.**Crop Science**, 43: 1235-1248.

MORAES, R. M. A.; CRUZ, C. D.; BARROS, E. G.; MOREIRA, M. A. Genetic divergence in soybean parents for backcrossing programs. **Crop Breeding and Applied Biotechnology**, Viçosa, v. 5, n. 3, p. 339-346, 2005.

NASCIMENTO FILHO, F.J. do; ATROCH, A.L.; SOUSA, N.R. de; GARCIA, T.B.; CRAVO, M. da S.; COUTINHO, E.F. Divergência genética entre clones de guaranazeiro. **Pesquisa Agropecuária Brasileira**, v.36, p.501-506, 2001.

NEI, M. Genetic distance between populations. **American Naturalist**.Chicago, v. 106, p. 238-292, 1972.

OLIVEIRA, R.P. de; CRISTOFANI, M.; AGUILAR-VILDOSO,C.I.; MACHADO, M.A. Diversidade genética entre híbridos detangerina 'Cravo' e laranja 'Pera'. **Pesquisa Agropecuária Brasileira**, v.37, p.479-484, 2002.

OLIVEIRA, E. Q. D., BEZERRA, N. F., Negreiros, M. Z. D., Barros Júnior, A. P., FREITAS, K., SILVEIRA, L., & LIMA, J. Desempenho agroeconômico do bicultivo de alface em sistema solteiro e consorciado com cenoura." **Horticultura Brasileira**, v. 22, n. 04, p. 712-717, 2004.

OLIVEIRA, A. C. B.; Sakiyama, N. S.; Mistro, J. C.; Giomo, G. S.; Fazuoli, L.C. Similaridade genética entre progênies de retrocruzamento e o genitor recorrente com base em marcadores moleculares. In: **Simpósio de Pesquisa dos Cafés do Brasil** (4. : Londrina, PR : 2005). Anais. Brasília, D.F. : Embrapa - Café, 2005. (1 CD-ROM), 6p.

PEREIRA, J.J.; CRUZ, C.D. Comparação demétodos de agrupamento para o estudo dadiversidade genética de cultivares de arroz. Revista**Ceres**, Viçosa, v.50, n.287, p.41-60, 2003.

PEREIRA, F.H.F.; PUIATTI, M.; MIRANDA, G.V.; SILVA, D.J.H.; FINGER, F.L. Divergência genética entre acessos de taro. **Horticultura Brasileira**, Brasília, v.22, n.1, p. 55-60, jan-mar 2004.

REGAZZI, A.J. Análise multivariada, notas de aula est746, departamento de estatística dauniversidade federal de viçosa. V.2. 2006.

RIBEIRO, N. D.; MELLO, R. M.; DALLA COSTA, R.; SLUSZZ, T. Correlações genéticas de caracteres agromorfológicos e suas implicações na seleção de genótipos de feijão carioca. **Revista Brasileira de Agrociência**, Pelotas, v. 7, p. 93-99, 2001.

RIZZO, A. A. do N.; BRAZ, L. T. Divergência genética entre cinco genótipos de melão rendilhado. **Horticultura Brasileira**, Brasília, DF, v. 20, n. 2, p. 167-170, jun. 2002.

RODRIGUES, L.S.; ANTUNES, I.F.; TEIXEIRA, M.G.; SILVA, J.B. Divergência genética entre cultivares locais e cultivaresmelhoradas de feijão. **Pesquisa Agropecuária Brasileira**, v.37, p.1275-1284, 2002.

SCAPIM, C.A.; PIRES, I.E.; CRUZ, C.D.; AMARAL JUNIOR, A.T.; BRACCINI, A. e L.; OLIVEIRA, V.R. Avaliação da diversidade genética em EucalyptuscamaldulensisDehnh, por meio da análise multivariada. **Revista Ceres,** v.6, p.347-356, 1999.

SCHEFFER-BASSO, S.M. ORSATO, J. MORO, G.V. ALBUQUERQUE, A.C.S. Divergência genética em germoplasma de aveias silvestres com base em caracteres multicategóricos e quantitativos. **Revista Ceres**, Viçosa, MG, v. 59, n. 5, p. 654-667, 2012.

SILVA, F.; REIS, R.Q.; LIMA REIS, C.A.; NUNES, D. Um Modelo de Simulação de Processo de Software baseado em Agentes Cooperativos**. 13º Simpósio Brasileiro de Engenharia de Software** – SBES 1999, Florianópolis: SBC/UFSC, Outubro, 1999.

SILVA, C.M.; GONÇALVES-VIDIGAL, M.C.; VIDIGAL FILHO, P.S.; SCAPIM, C.A.; DAROS, E.; SILVÉRIO, L. Genetic diversity among sugarcane clones (Saccharum spp.). **Acta Scientiarum**. Agronomy, v.27, p.315-319, 2005.

SOUZA, F.F.; QUEIRÓZ, M.A.; DIAS, R.S.C. Divergência genética em linhagens de melancia.**Horticultura Brasileira**, v.23, p.179-183, 2005.

SUDRÉ, C. P.; GONÇALVES, L. S. A.; AMARAL JÚNIOR, A. T.; RIVA-SOUZA, E. M.; BENTO, C. S. Genetic variability in domesticated Capsicum spp. as assessed by morphological and agronomic data in mixed statistical analysis. **Genetic and Molecular Research**, 9: 283-294, 2010.

VASCONCELOS, E. S. de; CRUZ, C.D.; BHERING, L.L.; RESENDE JÚNIOR, M.F.R. Método Alternativo para Análise de Agrupamento. **Pesquisa Agropecuária Brasileira**, Brasília, v.42,n.10, p.1421-1428, out. 2007.

VIEIRA, E. A; CARVALHO, F. I. F.; BERTAN, I.; KOPP, M. M.; ZIMMER, P. D.; BENIN, G.;SILVA, J. A. G.; Hartwig, I.; Malone, G.; Oliveira, A. C. (2007) Associationbetweengeneticdistances in wheat (Triticumaestivum L.) as estimatedbyAFLPandmorphologicalmarkers. **Geneticand Molecular Biology**. 30: 392-399.

CAPITULO 2

ANÁLISES DICRIMINANTE DE ANDERSON, DE FISHER E REDES NEURAIS ARTIFICIAIS EM ESTUDOS CLASSIFICATÓRIOS

RESUMO

A correta classificação de indivíduos é de extrema importância para fins de preservação da variabilidade genética existente bem como para a maximização dos ganhos o em programas de melhoramento genético. As técnicas de estatística multivariada comumente utilizada nessas situações são as funções discriminantes de Fisher e as funções discriminantes de Anderson, que permitem alocar um indivíduo inicialmente desconhecido em uma das g populações ou grupos pré-definidos. Entretanto, para altos níveis de similaridade como é o caso de populações de retrocruzamentos esses métodos têm se mostrado pouco eficientes. Atualmente, muito se fala de um novo paradigma de computação, as redes neurais artificiais, que podem ser utilizadas para resolver diversos problemas da Estatística, como agrupamento de populações similares, previsão de séries temporais e em especial, os problemas de classificação. O objetivo desse trabalho foi realizar um estudo comparativo entre as funções discriminantes de Fisher e de Anderson e as redes neurais artificiais quanto ao número de classificações incorretas de indivíduos sabidamente pertencentes a diferentes populações, com crescentes níveis de similaridade. A dissimilaridade, medida pela distância de Mahalanobis, foi um conceito de fundamental importância na utilização das técnicas de discriminação, pois quantificou o quanto as populações eram divergentes. Quanto maior o valor observado para essa medida, menos similares foram as populações em análise. A obtenção dos dados foi feita através de simulação utilizando o programa computacional Genes (CRUZ, 2006). As redes neurais artificiais mostraram-se uma técnica promissora no que diz respeito a problemas de classificação, uma vez que apresentaram um número de classificações incorretas de indivíduos menor que os dados obtidos pelas funções discriminantes.

ABSTRACT

The correct classification of individuals has a top importance for the genetic variability preservation as well as to maximize gains. The multivariate statistical techniques commonly used in these situations are the Fisher and Anderson discriminant functions, allowing to allocate an initially unknown individual in a g population or predefined groups. However, for higher levels of similarity such as backcross populations these methods has proved to be inefficient. Currently, much hás been Said about a new paradigm of computing, artificial neural networks, which can be used to solve many statistical problems as similar subjects grouping, time-series forecasting and in particular, the classification problems. The aim of this study was to conduct a comparative study between the Fisher and Anderson discriminant functions and artificial neural networks through the number of incorrect classifications of individuals known to belong to different simulated backcross with increasing levels of populations similarity. The dissimilarity, measured by Mahalanobis distance, was a concept of fundamental importance in the use of discrimination techniques, due the quantification of how much populations were divergent. Data collection was done through simulation using the software Genes (Cruz, 2013). The artificial neural network is shown as a promising technique to solve classification problems, since it had a number of incorrect individuals classifications smaller than the data obtained by the discriminant functions.

1. INTRODUÇÃO

Estudos que visam a discriminação de populações têm sido de grande importância para o desenvolvimento de programas de melhoramento genético e para conservação de biodiversidade. Análises da diversidade genética, por meio de características fenotípicas, têm orientado a escolha de genitores apropriados, em programas de melhoramento, levando à otimização dos ganhos seletivos, devido à variabilidade encontrada na descendência de cruzamentos entre grupos divergentes. Além disso, as análises de diversidade genética têm permitido a quantificação da variabilidade existente e facilitado o gerenciamento dos bancos de germoplasma, poupando tempo e recursos.

Uma das técnicas de análise multivariada que permite alocar um novo indivíduo a uma das varias populações distintas, previamente conhecidas, é a análise discriminante. Esta consiste na obtenção de funções que permitam diferenciar um determinado “indivíduo”, com base em medidas de várias características, em uma entre várias populações distintas, buscando minimizar a probabilidade de uma classificação errônea.

A utilização dessa técnica é bastante frequente, uma vez que é simples e possui alta eficiência para uma ampla variedade de estruturas populacionais. Ainda, de acordo com Cruz et al., (2012) , a sua aplicação busca minimizar a probabilidade de uma classificação errônea, isto é, de se classificar o referido “indivíduo” em uma população, quando este na realidade pertence a outra. Fundamentalmente estas técnicas dependem de informações de estatísticas representativas das populações sumarizadas em vetores de médias e em matrizes de variâncias e covariâncias entre um grupo de características.

Em diversas situações o pesquisador dispõe de dados experimentais apropriados para realização das analises biométricas necessárias para avaliação dos experimentos com objetivo de discriminar indivíduos ou populações, porém as análises biométricas nem sempre são capazes de produzir resultados satisfatórios, pois o modelo adotado e as estatísticas requeridas (médias, variâncias e covariâncias) podem ser insuficientes para descrever e caracterizar convenientemente as particularidades de cada população. Neste contexto, a realização de análises por meio de métodos computacionais que sejam capazes de aprendizagem e generalização a partir de toda a informação disponível, sendo tolerantes a ruídos, representa um

grande avanço para os estudos envolvendo procedimentos estatísticos e para o melhoramento genético.

A inteligência artificial tem permitido uma nova abordagem no processo de tomada de decisão em diversas áreas da ciência com grande potencial no melhoramento genético animal e vegetal Ventura et al.,(2013), em especial em análise classificatória Nascimento et al.,(2013). Um novo paradigma pode ser empregado no melhoramento genético para fins de seleção que não envolve modelagem estocástica, mas princípios de aprendizado em abordagem de inteligência computacional. Assim, a informação de cada indivíduo passa assumir considerável importância, pois constituem exemplos, no processo de aprendizagem, do padrão de cada população analisada.

A inteligência artificial tem permitido, com o avanço da tecnologia computacional e do entendimento da neurociência, a criação de modelos de neurônios artificiais muito próximos dos neurônios biológicos que conseguem se conectar formando as Redes Neurais Artificiais. Essas conexões as tornam capazes de análises de diversas situações, aprendizagem e reconhecimento de padrões (Braga et al., 2007).

O desempenho da rede é determinado pelas conexões entre os seus elementos. E por isso, pode-se treinar uma rede neural para executar uma função particular ajustando-se os valores das conexões entre os elementos (Haykin, 2001). Esse processo permite uma adaptação da RNA às particularidades de um problema, que a torna capaz de generalizações, permitindo as mesmas respostas para estímulos similares.

As RNAs caracterizam-se pela sua arquitetura e pelo ajustamento de seus pesos às conexões durante o processo de aprendizado. A arquitetura de uma rede neural é definida pelo número de camadas (camada única ou múltiplas), pelas conexões entre camadas, pelo número de neurônios em cada camada, pelo tipo de conexão entre eles (feedforward ou feedback) e pelo algoritmo de aprendizado (Haykin, 2001).

O avanço crescente na área de informática tem permitido estudos cada vez mais abrangentes nas ciências em geral. Isso ocorre tanto pelo desenvolvimento de tecnologias, que levaram ao sequenciamento do genoma de varias espécies, quanto pela descoberta de mecanismos que permitam o armazenamento e análises dessas sequências, tornando possível estuda-las das mais diversas formas, o que permite um avanço inimaginável. Outro

aspecto é a possibilidade de se gerar, *in silico*, um grande volume de dados, estruturados segundo modelos genéticos pré-definidos por meio de simulação computacional com algoritmos refinados. A utilização de dados simulados tem sido cada vez mais frequente em trabalhos associados ao melhoramento genético, pois pode ser introduzida em qualquer etapa do experimento, permitindo estudos exaustivos e poupando tempo e dinheiro.

A simulação permite que diferentes métodos sejam testados de forma que ocorra uma otimização dos mesmos para dados reais. O uso de dados simulados em redes neurais é muito importante, pois pode suprir a necessidade metodológica em que uma fase, denominada de treinamento, demanda grande quantidade de informações para o aprendizado eficaz das redes. Os dados para fins de treinamento são obtidos, normalmente, por bancos de dados históricos, mas em sua falta, o uso de dados simulados pode ser uma alternativa viável.

Assim, o presente trabalho tem por objetivo comparar a eficiência de técnicas multivariadas e das redes neurais em estudos de classificação de população. E também teve como objetivo demonstrar a potencialidade da técnica de simulação em estudos genéticos. Assim, procura-se evidenciar a importância da simulação em gerar populações a partir de características pré-estabelecidas (herdabilidade, variância e média), fazer a replicação e, ou, a ampliação de conjuntos populacionais, preservando as mesmas características pontuais de média, variância e herdabilidade e de estruturação (matriz de covariância ou de correlações) destas populações. Estando disponível todo conjunto de dados gerados *in silico*, pretende-se utilizar as técnicas baseadas em funções discriminantes de Fisher e Anderson, bem como as Redes Neurais Artificiais, para verificar a eficiência das mesmas na classificação de populações por meio da comparação das taxas de erro aparente encontradas. Foi proposto o uso do delineamento genético fundamentado em populações derivadas de retrocruzamentos, pois estas apresentam diferentes graus de distinguibilidade, ou dificuldade de discriminação, e o padrão de similaridade, entre cada par de população, é parametricamente conhecido tendo em vista os princípios meióticos fundamentados na contribuição gamética equitativa de genitores em cada geração de cruzamento.

2. MATERIAL E MÉTODOS

2.1 Simulação dos dados:

-Simulação dos dados genotípicos de populações estruturadas no delineamento genético de retrocruzamentos

A simulação e análise dos dados foram realizadas no Laboratório de Bioinformática da Universidade Federal de Viçosa, localizado no Instituto de Biotecnologia aplicada a Agropecuária (BIOAGRO) utilizando aplicativo computacional Genes (CRUZ, 2013).

Dados genotípicos foram, originalmente, simulados para dez populações em equilíbrio de Hardy-Weinberg, com 100 indivíduos cada. Foram geradas informações relativas a 50 locos manifestando dois alelos codominantes. Este conjunto prévio de dados foi utilizado para o cálculo de uma medida de dissimilaridade genotípica de Nei (1972) e também para estudos da diversidade genética das populações pela projeção gráfica das distâncias em plano bidimensional. O par de populações mais divergente foi tomado para gerar de um sistema hierárquico de retrocruzamentos apresentado na figura 1.

As relações de parentescos e a estruturação hierárquica foram estabelecidas considerando populações genitoras geneticamente divergentes, híbrido F_1 e cinco gerações de retrocruzamento em relação a cada um dos genitores, permitindo estabelecer parâmetros para quantificação da eficácia das metodologias testadas fundamentado no grau de distingüibilidade entre cada par de populações.

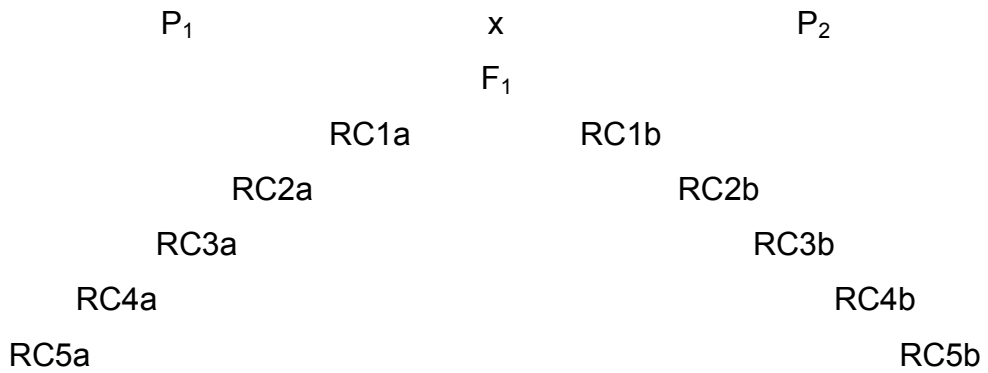


Figura 1: Esquema estruturado dos cruzamentos entre os genitores P_1 e P_2 e seus respectivos retrocruzamentos designados como recorrentes “a” e “b”.

-Simulação de valores fenotípicos

Neste estudo foram utilizadas informações moleculares e fenotípicas de 13 populações π_j, j=1,2,...,p=13, constituídas por indivíduos (n_j=100 para todo j) mensurados em relação a v=13 características quantitativas, contínuas, com distribuição normal, tendo média e variância previamente conhecidas. Para cada população simulada, considerou-se igualdade das matrizes de variância e covariâncias, uma vez que sem essa pressuposição perde-se a linearidade das funções discriminantes. Todas as populações foram simuladas por meio do uso do aplicativo computacional Genes (CRUZ, 2013).

Foram simulados 13 variáveis com valores de média e herdabilidade previamente estabelecidos. O delineamento experimental adotado na simulação foi o delineamento inteiramente casualizado em que cada população apresentava 100 genótipos, assumindo-se herdabilidade de 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75 e 80%, valores de média semelhante ao valor da herdabilidade para fins didáticos. Estas características foram estabelecidas pela ação de alelos de 20 locos, tomados ao acaso entre os 50 previamente genotipados, com efeito aditivo diferencial, com pesos da importância do loco, sobre a variabilidade genotípica total do caráter, estabelecidos a partir de uma distribuição binomial e grau médio de dominância nulo.

Para cada variável devem ser estabelecidos os valores da média e da herdabilidade, e utilizado o seguinte modelo estatístico:

$$Y_{ij} = \mu + G_i + \varepsilon_{ij}$$

Em que:

Y_{ij}: observação simulada de uma dada característica;

μ: média geral da característica, cujo valor é especificado pelo pesquisador;

G_i: efeito associado ao i-ésimo indivíduo da j-ésima população;

ε_{ij}: erro aleatório, sendo $\varepsilon_{ij} \sim N(0, \sigma^2)$

Como foram estabelecidos valores prévios da herdabilidade e da média e o controle genético do caráter (número de locos e ação de cada loco) é conhecido, as estimativas da variância genética e ambiental podem ser calculadas. Por princípios de genética quantitativa, as variâncias genéticas esperadas nas gerações de retrocruzamento podem ser previstas pelas frequências genotípicas e efeitos aditivos e, se for o caso dos efeitos de dominância que, neste trabalho, foram considerados nulos. O modelo aditivo foi

empregado de forma que o valor genotípico de cada indivíduo foi gerado e, acrescido de um efeito ambiental, simulado admitindo distribuição normal, com média zero e variância σ^2, resultando em um valor fenotípico final submetido às análises. Segundo Cruz et al., (2012), esse modelo além de mais simples, tem sido rotineiramente utilizado no melhoramento por fornecer valiosas informações para o êxito de programas de melhoramento.

Para fins didáticos as características utilizadas foram divididas em dois cenários: Cenário A- Alta herdabilidade (v = 55,..., 80) e Cenário B – Baixa herdabilidade (v= 20,..., 50), conforme pode ser visualizado nas tabelas 1 e 2.

Tabela 1: Médias paramétricas das características simuladas das 13 populações constituindo o Cenário A de alta herdabilidade.

Populações	Características					
	h²=55	h²=60	h²=65	h²=70	h²=75	h²=80
P1	50,00	55,00	60,00	65,00	70,00	75,00
P2	25,00	27,50	30,00	32,50	35,00	37,50
F1	37,50	41,25	45,00	48,75	52,50	56,25
RC1a	43,75	48,13	52,50	56,88	61,25	65,63
RC2a	46,88	51,56	56,25	60,94	65,63	70,31
RC3a	48,44	53,28	58,13	62,97	67,81	72,66
RC4a	49,22	54,14	59,06	63,98	68,91	73,83
RC5a	49,61	54,57	59,53	64,49	69,45	74,41
RC1b	31,25	34,38	37,50	40,63	43,75	46,88
RC2b	28,13	30,94	33,75	36,56	39,38	42,19
RC3b	26,56	29,22	31,88	34,53	37,19	39,84
RC4b	25,78	28,36	30,94	33,52	36,09	38,67
RC5b	25,39	27,93	30,47	33,01	35,55	38,09

Tabela 2: Médias paramétricas das características simuladas das 13 populações constituindo o Cenário B de baixa herdabilidade.

Populações	Características						
	h²=20	h²=25	h²=30	h²=35	h²=40	h²=45	h²=50
P1	20,00	25,00	30,00	35,00	40,00	45,00	20,00
P2	10,00	12,50	15,00	17,50	20,00	22,50	10,00
F1	15,00	18,75	22,50	26,25	30,00	33,75	15,00
RC1a	17,50	21,88	26,25	30,63	35,00	39,38	17,50
RC2a	18,75	23,44	28,13	32,81	37,50	42,19	18,75
RC3a	19,38	24,22	29,06	33,91	38,75	43,59	19,38
RC4a	19,69	24,61	29,53	34,45	39,38	44,30	19,69
RC5a	19,84	24,80	29,77	34,73	39,69	44,65	19,84
RC1b	12,50	15,63	18,75	21,88	25,00	28,13	12,50
RC2b	11,25	14,06	16,88	19,69	22,50	25,31	11,25
RC3b	10,63	13,28	15,94	18,59	21,25	23,91	10,63
RC4b	10,31	12,89	15,47	18,05	20,63	23,20	10,31
RC5b	10,16	12,70	15,23	17,77	20,31	22,85	10,16

Simulação dos cenários de distinguibilidade

O conjunto de populações e características foi aplicado a seis diferentes cenários com diferentes graus de distinguibilidade tendo em vista o padrão de similaridade das populações envolvidas e repetidos para cada *cenário de características*. No cenário de características foi considerado diferentes magnitudes de ruídos, expresso por efeitos ambientais aleatórios, envolvendo os grupos A (características de alta herdabilidade) e B (características de baixa herdabilidade) de forma a eficiência das análises discriminantes e das redes neurais, como definido na Tabela 3.

Tabela 3: Constituição dos cenários de distinguibilidade utilizados pelas funções discriminantes e redes neurais artificiais para características de alta e baixa herdabilidade.

Cenário	Delineamento Genético	Observações
1	P1, P2, F1	300
2	P1, P2, F1, RC1a, RC1b	500
3	P1,P2,F1, RC1a, RC1b, RC2a, RC2b	700
4	P1,P2,F1, RC1a, RC1b, RC2a, RC2b, RC3a, RC3b	900
5	P1,P2,F1, RC1a, RC1b, RC2a, RC2b, RC3a, RC3b, RC4a, RC4b	1100
6	P1,P2,F1, RC1a, RC1b, RC2a, RC2b, RC3a, RC3b, RC4a, RC4b, RC5a, RC5b	1300

2.2. Funções Discriminantes

Os métodos de análise discriminante aplicam-se a populações que possuem uma partição definida a priori, descritas por diversas variáveis explicativas. O objetivo é discriminar as classes da partição, através das características definidas pelas variáveis explicativas. Pretende-se, então, construir uma regra de decisão que permita, no futuro, alocar novos indivíduos, minimizando os erros de alocação.

Análise Discriminante de Fisher

A Função Discriminante Linear de Fisher, uma combinação linear das características observadas que apresenta melhor poder de discriminação entre os grupos, constitui a base de todo o estudo na análise discriminante. Esta função tem a propriedade de minimizar a probabilidade de má classificação, quando as populações são normalmente distribuídas, com média e variância conhecidas. Contudo, tal situação pode não ocorrer na prática, necessitando-

se, portanto de estimativas e métodos de estimação dessas probabilidades ótimas (Cruz et al.,2012).

Considerando duas populações (π_i e $\pi_{i'}$) com vetor de médias v-variado μ_i e $\mu_{i'}$ e matriz de variâncias e covariâncias comuns $\sum$, de ordem v, define-se a função discriminante linear de Fisher pela expressão 1.

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = \left(\mu_i - \mu_{i'}\right)' \Sigma^{-1}\tilde{x} \quad (1)$$

Assim, a função discriminante $D_{ii'}(\tilde{x})$ é uma combinação linear do conjunto de caracteres que possibilita alocar um determinado indivíduo, com vetor de observações $\tilde{x}$, em uma população π_i, ou $\pi_{i'}$, com máxima probabilidade de acerto. Define-se também o ponto médio entre duas populações π_i e $\pi_{i'}$ pelo valor m, expresso pela equação 2 ou 3.

$$m_{ii'} = \frac{1}{2}\left(\mu_i - \mu_{i'}\right)' \Sigma^{-1}\left(\mu_i + \mu_{i'}\right) = \alpha' u = \frac{1}{2}\left(\alpha' \mu_1 + \alpha' \mu_2\right) \quad (2)$$

ou

$$m_{ii'} = \frac{1}{2}\left[D(\mu_1) + D(\mu_2)\right] \quad (3)$$

Com a função discriminante estimada, adota-se a regra de classificação conforme as expressões 4 e 5.

- Aloca-se $\tilde{x}$ em π_i se:

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = \left(\mu_1 - \mu_2\right)' \Sigma^{-1}\tilde{x} \geq m_{ii'} \quad (4)$$

- Aloca-se $\tilde{x}$ em $\pi_{i'}$ se:

$$D_{ii'}(\tilde{x}) = \alpha'\tilde{x} = \left(\mu_1 - \mu_2\right)' \Sigma^{-1}\tilde{x} < m_{ii'} \quad (5)$$

A ideia básica da Análise Discriminante de Fisher foi transformar observações multivariadas X em observações univariadas Y derivadas das populações π_1 e π_2 em que estas apresentassem o maior grau de separação possível. Fisher sugere tomar combinações lineares de X para criar as combinações Y's, pois tais combinações podem ser facilmente manipuladas.

Análise discriminante de Anderson

De acordo com Anderson (1958), quando se dispõe de várias populações e se deseja alocar um novo indivíduo a cada uma delas, um

procedimento importante é que os indivíduos estejam divididos em populações distintas, e que seja estabelecida as probabilidades “a priori” para as várias populações, pois há casos nos quais a probabilidade de um determinado indivíduo pertencer a uma dada população pode ser muito distinta da dele pertencer a outra, de forma que a experiência do pesquisador torna-se de extrema importância.

Com estas informações, são geradas funções, que são combinações lineares das características avaliadas, e que tem por finalidade obter a melhor discriminação entre os indivíduos, alocando-os em suas devidas populações.

Além disso, as funções permitem também a classificação de novos genótipos, de comportamento desconhecido, nas populações já conhecidas. A eficácia das variáveis utilizadas em promover a discriminação também é avaliada, permitindo conhecer a adequação da função estimada.

Para o estabelecimento da função discriminante de Anderson, considera-se que, para uma população π_j (j = 1, 2, ..., g), o vetor da variável aleatória $\tilde{x}$ tem distribuição $N_v(\mu_j, \Sigma)$, com a seguinte função densidade de probabilidade (equação 6).

$$f_j(\tilde{x}) = |2\pi\Sigma|^{-1/2} \; e^{-\frac{1}{2}\left[\left(\tilde{x}-\mu_j\right)'\Sigma^{-1}\left(\tilde{x}-\mu_j\right)\right]} \qquad (6)$$

Também é admitido que a probabilidade de uma observação pertencer a uma determinada população é p_j ($\sum_{j=1}^{g} p_j = 1$), conhecida *a priori*. Assim, pode-se estabelecer a função discriminante, dada pela probabilidade de $\tilde{X}$ pertencer a π_j, por meio do logaritmo da função densidade de probabilidade de $\tilde{x}$ e da probabilidade a *priori*, de forma que se tenha:

$$D_j(\tilde{x}) = -\frac{1}{2}\left[\ln(2\pi) + \ln|\Sigma_j|\right] - \frac{1}{2}\left[\left(\tilde{x}-\mu_j\right)'\Sigma_j^{-1}\left(\tilde{x}-\mu_j\right)\right] + \ln(p_j) \qquad (7)$$

ou, de forma simplificada:

$$D_j(\tilde{x}) = \ln(p_j) + \left(\tilde{x} - \frac{1}{2}\mu_j\right)'\Sigma^{-1}\mu_j \qquad (8)$$

Com base nas médias de cada população e na matriz de variância e covariância entre as médias das populações, obtiveram-se as respectivas funções discriminantes. Cada função é uma combinação linear das v

características avaliadas, existindo tantas funções quanto for o número de populações avaliadas. A partir das funções discriminantes, estima-se, para cada genótipo, o valor discriminante, permitindo, classificar o i-ésimo indivíduo, com vetor de média $\tilde{x}_i$, na população π_j se e somente se $D_j(\tilde{x}_i)$ for o maior entre os elementos do conjunto $\{D_1(\tilde{x}_i), D_2(\tilde{x}_i),..., D_g(\tilde{x}_i)\}$.

Taxa de erro Aparente

Conhecidas as funções discriminantes de Fisher e Anderson, utilizou-se a taxa de erro aparente para avaliar a sua eficácia quanto a classificação de novos indivíduos em populações previamente conhecidas. A taxa de erro aparente é dada em função da probabilidade de má classificação, que para cada população, é dada pela equação 6:

$$p_j = \frac{m_j}{n_j}, (j = 1,2,3) \quad (6)$$

onde m_j é o número de observações retiradas de uma população, que foram, por meio da técnica avaliada, classificadas em outra população e n_j é o número de indivíduos da população j. Assim, tem-se:

$$\mathrm{TEA} = \frac{1}{N}\sum_{j=1}^{g} m_j \quad (7)$$

em que N é o número total de observações avaliadas e g é o número de populações consideradas.

Ampliação simulada de dados de experimentos para fins de treinamento da rede

Para fins de treinamento das redes neurais um novo conjunto de dados foi simulado, mas que representavam a ampliação dos dados originais por preservar particularidades dos dados das populações originais tomados para fins de estudo de análise discriminante, como descrito a seguir.

Os valores simulados para as observações da população foram tomados como uma variável aleatória $Y \sim N(\phi, \Sigma)$. Os dados foram transformados em uma variável aleatória $Z \sim N(\phi, I)$ por meio da transformação linear $Z = F'Y$, sendo F obtida por meio do processo de decomposição espectral de Σ, tal que $\Sigma^{-1} = FF'$. O processo de ampliação consiste na simulação de novos valores de Y, considerando $Y \sim N(\phi, (F')^{-1}Z)$. Foi considerado um arquivo de dados ampliados de 100 para 200 indivíduos em cada população, consistindo um total de 2600

genótipos para treinamento da Rede Neural. O processo de ampliação de dados também foi realizado por meio do aplicativo computacional GENES (Cruz, 2013).

Em procedimentos de simulação de um conjunto de dados, ou replicação de uma estrutura de dados conhecidos, ou mesmo a ampliação de um conjunto a partir da estrutura de outro, algumas pressuposições devem ser atendidas. A primeira delas é que o conjunto de dados deve ter uma distribuição conhecida e, a princípio, média igual a zero e variância igual a V. Para satisfazer a essa exigência, utilizou-se o Teorema de Box Muller, que garante que as variáveis x e y são normalmente distribuídas com média zero e variância V, como nas equações 8 e 9, em que:

$$x = \sqrt{-2log_e(RND)V}\cos(2\pi RND) \quad (8)$$

e

$$y = \sqrt{-2log_e(RND)V}\sin(2\pi RND) \quad (9)$$

Sendo *RND* um número aleatório.

Para garantir que a covariância do conjunto X de dados é nula, a metodologia recomendada é o uso da técnica de componentes principais e se baseia na simplificação do conjunto de dados para um conjunto reduzido de componentes, os quais apresentam as propriedades de reter o máximo da variação originalmente disponível e ser independentes entre si (Cruz, 2006).

Considere a variável aleatória $\mathrm{Y}\sim\mathrm{N}(\phi,\Sigma)$ que desejamos transformar em uma variável aleatória $\mathrm{Z}\sim\mathrm{N}(\phi,\mathrm{I})$. Por meio do processo de decomposição espectral, tem-se que $\Sigma^{-1} = \mathrm{FF}'$. Então, $\left(\Sigma^{-1}\right)^{-1} = \left(\mathrm{FF}'\right)^{-1} = \left(\mathrm{F}'\right)^{-1}\mathrm{F}^{-1} = \Sigma$

Se $\mathrm{Z} = \mathrm{F}'\mathrm{Y}$, então $\mathrm{E}(\mathrm{Z}) = \mathrm{E}\left(\mathrm{F}'\mathrm{Y}\right) = \mathrm{F}'\mathrm{E}(\mathrm{Y}) = \mathrm{F}'\phi = \phi$ e $\mathrm{V}(\mathrm{Z}) = \mathrm{F}'\mathrm{V}(\mathrm{Y})\mathrm{F} = \mathrm{F}'\Sigma\,\mathrm{F} = \mathrm{F}'\left[\left(\mathrm{F}'\right)^{-1}\mathrm{F}^{-1}\right]\mathrm{F} = \mathrm{I}$

Consideremos que $\mathrm{Z}\sim\mathrm{N}(\phi,\mathrm{I})$. Se $\mathrm{Z} = \mathrm{F}'\mathrm{Y}$, então $\mathrm{Y} = \left(\mathrm{F}'\right)^{-1}\mathrm{Z}$

Portanto, $\mathrm{V}(\mathrm{Y}) = \mathrm{V}\left(\left(\mathrm{F}'\right)^{-1}\mathrm{Z}\right) = \left(\mathrm{F}'\right)^{-1}\mathrm{V}(\mathrm{Z})[(\mathrm{F})^{-1}]' = \left(\mathrm{F}'\right)^{-1}\mathrm{I}\,\mathrm{F}^{-1} = \Sigma$

No conjunto de dados ampliados, foram obtidos arquivos com 2600 genótipos para treinamento da Rede Neural Artificial.

2.3. Caracterização da Rede Neural

No presente trabalho a arquitetura de Rede utilizada foi a Multilayer Perceptron (MLP), estabelecida adotando-se a seguinte configuração: uma camada de entrada, três camadas intermediária e uma camada de saída. A Rede MLP foi processada no software Matlab por meio de script apresentado no *módulo integração* no aplicativo computacional GENES.

As redes neurais artificiais apresentadas neste trabalho apresentavam duas configurações diferentes. A primeira delas foi utilizada no *Cenário de varíavel* A e apresentava seis entradas (correspondentes às diferentes características avaliadas). Já a segunda apresentava uma entrada a mais, pois no cenário B havia sete características a serem avaliadas.

As funções de ativação utilizadas foram a linear (purelin), para a camada de saída e, para as camadas ocultas, foram estabelecidas todas as combinações possíveis das funções de ativação tangente hiperbólica (tansig) e logarítmica (Logsig). O algoritmo escolhido para treinamento foi Trainbr– Backpropagation. O Número de ciclos de treinamento foi fixado em 2000 épocas. Teve-se o cuidado de limitar o número de iterações, para que esse não se tornasse excessivo, o que poderia levar à perda do poder de generalização.

O número de neurônios nas camadas intermediárias variou de 6 a 15 neurônios na primeira camada, de 10 a 40 na segunda camada e de 10 a 40 na terceira camada. A camada de saída foi composta por um neurônio e a saída foi representada por um vetor cujos elementos eram o número da população, onde esse valor era conhecido no treinamento e desconhecido na validação. A melhor arquitetura da rede foi estabelecida por aquela com acurácia média superior, considerando as 64800 possibilidades, calculada pela multiplicação do número de neurônios em cada camada e as funções de ativação possíveis (15X40X40X3X3X3).

Assim, foi escolhida a rede mais eficiente, para cada um dos cenários adotando como critério a menor taxa de erro aparente. As Figuras 2 e 3 ilustram as configurações de rede adotadas para os Cenários A e B.

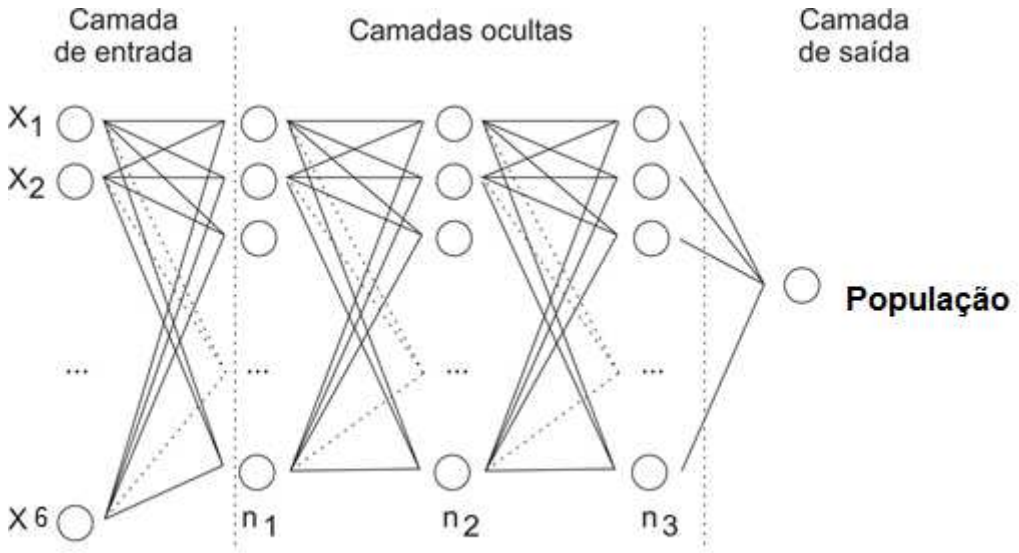


Figura 2. Arquitetura da RNA. Entradas de (x_1) a (x_6) na camada entrada estão relacionadas com as características simuladas e são consideradas como entrada. As camadas ocultas foram compostas por n_i (ni variando de um a 15 ou 40 nós), com funções de ativação purelin, tansig ou logsig. Todas as combinações foram exploradas. Na camada de saída, as RNA's retornaram a população que o individuo pertencia.

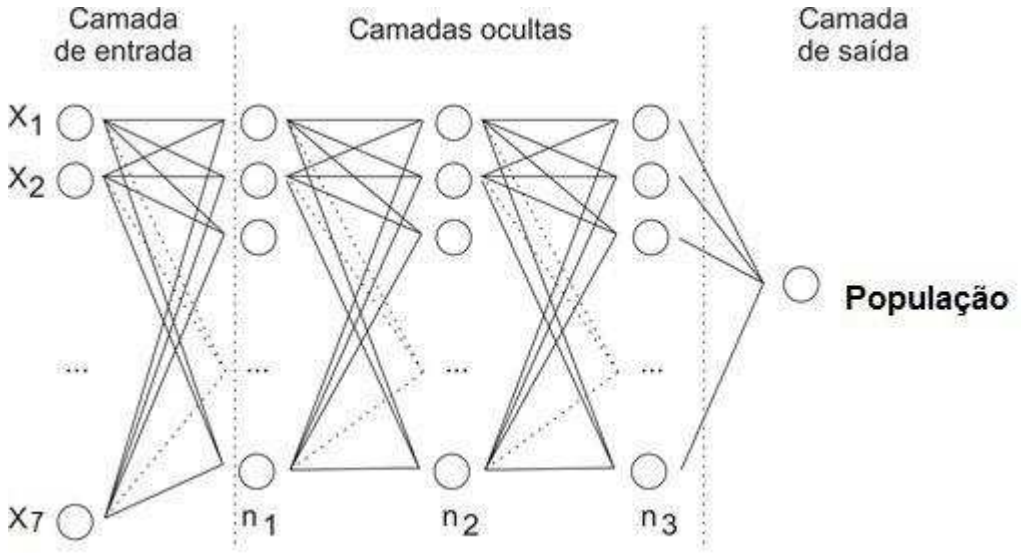


Figura 3. Arquitetura da RNA. Entradas de (x_1) a (x_7) na camada entrada estão relacionadas com características simuladas e são consideradas como entrada. As camadas ocultas foram compostas por ni (ni variando de um a 15 ou 40 nós), com funções de ativação purelin, tansig ou logsig. Todas as combinações foram exploradas. Na camada de saída, as RNA's retornaram a população que o individuo pertencia.

3. RESULTADO E DISCUSSÃO

Análises Discriminantes de Fisher e Anderson

As análises discriminantes de Fisher e Anderson foram igualmente ineficazes na discriminação das populações em todos os cenários de distinguibilidade e de variáveis, como pode ser observados nas Tabelas 4 e 6, visto que não foram capazes de diferenciar as populações de retrocruzamentos com características quantitativas de alta ou baixa herdabilidade. É valido lembrar que, a cada cenário de distinguibilidade, o nível de dificuldade na classificação das populações aumentava já que espera-se, considerando a contribuição gamética equitativa, que a recuperação do genitor recorrente na geração x (representado por RC_x) ocorra na proporção de $(2^{x+1} - 1) / 2^{x+1}$, em que x é número de retrocruzamentos com o genitor recorrente (Cruz, 2005). Assim, após cinco gerações de retrocruzamento a proporção esperada de similaridade entre a RC_5 e o genitor recorrente, é de, aproximadamente, 98,44% (Borém, 2009). Esse grau de similaridade torna-se bastante difícil a discriminação das diferentes populações de retrocruzamento entre si e com o respectivo genitor recorrente que, na verdade, é o propósito da estratégia de melhoramento por retrocruzamento. Portanto, as taxas de erro aparente, medidas separadamente em cada cenário, foram superiores a 50% para quase todos os cenários analisados, e chegaram a mais de 80% no Cenário de distingüibilidade 6, em que as cinco gerações de retrocruzamento estavam presentes.

Inferir sobre a diferenciação entre populações é assunto de grande relevância tanto sob aspectos de melhoramento genético quanto na adaptação e evolução de grupos populacionais. No melhoramento genético cita-se o estudo realizado por Pereira et al. (1999) cujo objetivo era a diferenciação de cultivares de arroz considerados de padrão moderno e tradicional. Fatores como seleção, fluxo gênico, amostragem, dentre outros, podem contribuir para ampliar ou reduzir a diversidade entre populações e demandam estratégias diferenciadas para situações particulares. Deve-se, portanto, inferir com acurácia satisfatória sobre a diversidade entre o material estudado para que as decisões adotadas sejam as mais acertadas.

Também é ressaltado que os estudos de diversidade e de discriminação entre populações ou genótipos são de grande valia em programas de

melhoramento cujo objetivo é estabelecer grupos heteróticos e identificar combinações híbridas de maior vigor. O resultado obtido neste trabalho mostra que, em situações reais, a utilização das técnicas biométricas para análise discriminante de populações com grau de distinguibilidade semelhante às analisadas seria inviável a partir do momento que se observa baixo grau da capacidade de diferenciação entre elas e alta taxa de erro aparente. Assim, a busca por uma metodologia que evidencie, com a máxima eficácia possível, o grau de diferenciação entre populações é de grande importância.

Como pode ser visualizado na Tabela 4, para o cenário de variáveis "A", em que são utilizadas variáveis quantitativas de alta herdabilidade na discriminação dos indivíduos das populações P_1 e P_2 e F_1, que constituem o Cenário de distinguibilidade 1, os resultados obtidos pelas funções discriminantes já foi pouco satisfatória, uma vez que a taxa de erro entre elas foi superior a 20%.

É possível observar, pelos resultados considerando o cenário de distinguibilidade 6 apresentados na Tabela 5, que as populações P_1 e P_2 são bem diferenciadas, pois apenas um dos indivíduos da população P_1 foi classificado de forma equivocada na população P_2 e, de forma inversa, o mesmo resultado foi encontrado. A população F1, que apresentou maior taxa de alocação correta de 43%, teve 19% dos seus indivíduos alocados incorretamente em RC1b, o que poderia indicar sua maior similaridade com o genitor P2.

Tabela 4. Taxa de erro aparente (TAE) calculada nas funções discriminante de Fisher (FIS) e de Anderson (AND) estabelecidas pela combinação linear entre as características do *Cenário de variáveis* A para fins de discriminação entre as populações formados nos seis *Cenários de distinguibilidade.*

Cenário de variável - A				
Cenários de distinguibilidade	Tamanho Populacional	Similaridade(*)	TAE (%) Fisher	TAE (%) Anderson
1	300	50,00%	22,67	22,67
2	500	75,00%	54,60	54,60
3	700	87,50%	67,57	67,57
4	900	93,75%	74,00	74,00
5	1100	96,875%	78,00	78,00
6	1300	98,43%	80,01	80,01

(*)similaridade genética máxima esperada entre as populações de retrocruzamento.

Tabela 5. Resumo da classificação incorreta da Função Discriminante de Anderson no cenário A de alta herdabilidade e cenário 6 de distinguibilidade, para todos os conjuntos de populações analisados.

Resumo - % de Classificação correta e incorreta de cada grupo Cenário A													
POP	P1	P2	F1	RC1a	RC2a	RC3a	RC4a	RC5a	RC1b	RC2b	RC3b	RC4b	RC5b
P1	24	1	6	8	12	9	22	8	2	0	1	0	7
P2	1	24	17	5	1	1	1	0	9	5	11	8	17
F1	0	6	43	8	2	1	0	2	19	5	3	2	9
RC1a	9	0	21	20	8	7	13	0	4	1	4	4	9
RC2a	16	2	7	10	12	13	17	7	4	1	4	2	5
RC3a	27	2	7	13	11	14	20	2	1	0	0	0	3
RC4a	16	2	6	13	12	14	24	8	0	0	1	1	3
RC5a	16	1	6	18	12	13	18	12	1	1	0	0	2
RC1b	1	14	26	7	0	0	2	2	16	5	12	8	7
RC2b	1	13	28	8	6	0	1	0	4	8	12	8	11
RC3b	0	12	26	9	5	1	0	0	6	3	18	4	16
RC4b	0	18	16	4	3	3	0	0	9	2	14	13	18
RC5b	0	10	14	9	4	1	1	0	6	3	15	13	24

*A tabela deve ser lida no sentido vertical.

Na Tabela 6 é possível observar que, para o cenário “B”, em que são utilizadas variáveis quantitativas de baixa herdabilidade na discriminação dos indivíduos das populações P_1 e P_2 e F_1, que constituem o Cenário de distinguibilidade 1, pelas funções discriminantes também foi pouco satisfatória, uma vez que a taxa de erro entre elas foi de a 27%.

Tabela 6. Taxa de erro aparente (TAE) calculada nas funções discriminante de Fisher e de Anderson estabelecidas pela combinação linear entre as características do Cenário B para as populações nos seis Cenários.

Cenário de variável - B				
Cenários de distinguibilidade	Tamanho Populacional	Similaridade(*)	TAE (%) Fisher	TAE (%) Anderson
1	300	50%	27,00	27,00
2	500	75%	52,80	52,80
3	700	87,5%	65,57	65,57
4	900	93,75%	74,88	74,88
5	1100	96,875%	77,63	77,63
6	1300	98,43%	81,07	81,70

(*)similaridade genética máxima esperada entre as populações de retrocruzamento.

De maneira geral, observa-se, pela Tabela 7 em que são apresentas as classificações incorretas relativas ao cenário de distinguibilidade 6, que as populações P_1 e P_2 são bem diferenciadas, pois nenhum dos indivíduos da população P_1 foi classificado de forma equivocada na população P_2 e, de forma inversa, o mesmo resultado foi encontrado. Já a população F_1 foi a que

apresentou maior taxa de alocação correta (42%). A maior taxa de alocação incorreta dessa população foi de 15% dos indivíduos que foram classificados em RC1a o que poderia indicar, contraditoriamente, sua maior similaridade com o genitor P_1.

Uma comparação entre os resultados das Tabelas 5 e 7, com ênfase nos valores obtidos para as classificações corretas (elementos da diagonal das matrizes representadas nestas tabelas) permitem concluir que o efeito ambiental mais pronunciado contribui para a redução deste tipo de classificação.

Tabela 7. Resumo da classificação incorreta da Função Discriminante de Anderson no cenário de baixa herdabilidade e cenário 6 de distinguibilidade, para todos os conjuntos de populações analisados.

Resumo - % de Classificação correta e incorreta de cada grupo Cenário B

POP	P1	P2	F1	RC1a	RC2a	RC3a	RC4a	RC5a	RC1b	RC2b	RC3b	RC4b	RC5b
P1	36	0	4	12	3	6	12	22	4	0	0	0	1
P2	0	8	15	4	4	0	2	2	10	15	13	14	13
F1	0	2	42	15	1	0	0	2	12	6	7	8	5
RC1a	10	0	16	19	4	2	5	11	15	6	4	4	4
RC2a	18	3	8	16	3	7	10	18	5	3	1	3	5
RC3a	21	0	4	14	5	6	10	25	5	2	2	3	3
RC4a	22	0	6	10	3	6	16	25	3	0	1	5	3
RC5a	23	0	5	8	0	9	16	31	2	2	2	1	1
RC1b	0	4	20	10	4	4	2	1	21	7	8	9	10
RC2b	2	6	18	4	1	1	1	0	17	19	15	9	7
RC3b	1	4	18	7	2	0	0	0	15	16	14	12	11
RC4b	1	10	21	6	2	2	1	1	13	13	4	18	8
RC5b	2	6	8	9	3	2	0	0	22	12	5	18	13

*A tabela deve ser lida no sentido vertical.

As funções discriminantes foram obtidas com base nas médias de cada população e na matriz de variâncias e covariâncias, obtidas dentro de cada população, admitidas ser homogêneas. As funções foram estimadas considerando probabilidade *a priori* igual para cada uma das populações. De maneira geral, os resultados apresentados neste trabalho apontam que os procedimentos multivariados fundamentados nas funções discriminantes produziram resultados não tão ineficientes para a discriminação de populações delineadas em sistema genético de retrocruzamento tanto para Cenários de características de alta herdabilidade quanto de baixa herdabilidade em todos os cenários de distinguibilidade.

Aplicação das Redes Neurais Artificiais

A qualidade e o tamanho do conjunto de dados para treinamento são de extrema importância para a eficiência da rede neural (Kavzoglu, 2009). Devido a isso é necessário que os dados utilizados no treinamento para a classificação das populações sejam representativos de todos os genótipos e, ainda, numerosos o suficiente, pois há uma relação direta entre o tamanho do conjunto de dados para treinamento e a confiabilidade das estimativas dos dados de validação da rede. O tamanho da amostra esta relacionada principalmente com as competências utilizadas pelas redes neurais para o treinamento. Uma amostra com poucos indivíduos não é suficiente para uma rede neural reconhecer todas as classes possíveis, já uma amostra maior pode tornar a rede mais especifica e melhorar a confiabilidade dos resultados, porém exige maior tempo computacional para a execução das tarefas de treinamento da rede (Kavzoglu, 2001). Entretanto, um exagero na analise dos dados pode levar a rede a decorar exemplos se tornando incapaz de generalização (Braga et al., 2007).

Um dos meios de obtenção do conjunto de treinamento do tamanho adequado é pelo processo de ampliação dos dados. Com o objetivo de conseguir uma população de treinamento que mantivesse a estrutura genética do experimento inicial, foi utilizado o processo de ampliação das populações simuladas no programa GENES.

Nas Tabelas 8 e 9 são apresentados os resultados obtidos pelo uso das RNA. Em cada um dos seis cenários de distinguibilidade considerou-se a possibilidade do treinamento da rede ser feito a partir de um conjunto ampliado de dados, gerados independentemente, tomando por base os vetores de médias e as matrizes de variâncias e covariâncias estimadas em cada população. Na fase de treinamento, a taxa de erro aparente, ou seja, de classificação equivocada, foi nula, para quase todos os cenários de distinguibilidade, se mostrando perfeitamente capaz de diferenciar totalmente as populações até a 3 geração de retrocruzamento. No cenário de distinguibilidade 5, que apresentava todas as populações de retrocruzamentos até a quarta geração - RC4, a taxa de erro aparente foi de 1,36% para as populações com características de alta herdabilidade - Cenário A- e de 0,72% para as populações com características de baixa herdabilidade- o Cenário B.

Tabela 8. Taxa de erro aparente (TAE%) apresentada pela Rede Neural aplicada aos seis cenários de distinguibilidade, com características de alta herdabilidade.

Fase de treinamento da RNA - Cenário de variáveis A			
Cenários de distinguibilidade	Observações	Similaridade(*)	TAE (%) RNA
1	600	50%	0%
2	1000	75%	0%
3	1400	87,5%	0%
4	1800	93,75%	0%
5	2200	96,875%	1,36%
6	2600	98,43%	19,00%

(*) similaridade genética máxima esperada entre as populações de retrocruzamento

A análise dos resultados das Tabelas 8 e 9 mostra que, mesmo as RNAS treinadas com essas configurações, são capazes de diferenciar também as populações com cinco gerações de retrocruzamentos, cenário de distinguibilidade 6, tanto para características de alta ou baixa herdabilidade, com uma taxa de erro de 19% e 19,23%, para os cenários de variáveis A e B, respectivamente.

Tabela 9. Taxa de erro aparente (TAE%) apresentada pela Rede Neural aplicada aos seis cenários com características de baixa herdabilidade.

Fase de treinamento da RNA - Cenário de variáveis B			
Cenários de distinguibilidade	Observações	Similaridade(*)	TAE (%)RNA
1	600	50%	0%
2	1000	75%	0%
3	1400	87,5%	0%
4	1800	93,75%	0%
5	2200	96,875%	0,72%
6	2600	98,43%	19,23

(*) similaridade genética máxima esperada entre as populações de retrocruzamento.

As RNAs, ao serem empregadas no conjunto de dados original (arquivo de validação), proporcionaram resultados bem mais satisfatórios do que os obtidos pelas funções discriminantes de Fisher ou de Anderson, com taxa de erro aparente inferior a 15,0% no pior cenário (Tabela 11). Entretanto, é válido ressaltar que essa melhor contribuição se deve ao fato da capacidade das RNAS de extraírem características dos dados, não se baseando apenas nos parâmetros estáticos de média e variância, como as funções discriminantes.

Neste trabalho, a utilização de características de baixa ou alta herdabilidade pouco influenciou na eficácia das redes neurais artificiais aplicadas aos diferentes cenários. Novamente, para as populações até a terceira geração de retrocruzamentos (RC3), a taxa de erro aparente foi nula, se mostrando 100% capaz de diferenciar as populações. Para o cenário em que as quatro gerações de retrocruzamentos estavam representadas a taxa de erro foi praticamente nula sendo de 0,27% para as características quantitativas de alta herdabilidade e de 0,09% para as de baixa. Para o cenário de distinguibilidade 6 que possuía até a quinta geração de retrocruzamentos, cujo nível de similaridade ultrapassava 98% houve uma taxa de erro aparente 12,53 para o Cenário A e de 14,61 para o B. Esses resultados podem ser observados nas Tabelas 10 e 11.

Tabela 10. Taxa de erro aparente (%) apresentada pela Rede Neural aplicada nos seis cenários com características de Alta herdabilidade.

Fase de validação cenário A			
Cenários de Distinguibilidade	Observações	Similaridade	TAE (%)RNA
1	300	50%	0%
2	500	75%	0%
3	700	87,5%	0%
4	900	93,75%	0%
5	1100	96,875%	0,27%
6	1300	98,43%	12,53%

(*) similaridade genética máxima esperada entre as populações de retrocruzamento.

Tabela 11. Taxa de erro aparente (%) apresentada pela Rede Neural aplicada nos seis cenários com características de Baixa herdabilidade.

Fase de validação cenário B			
Cenários de distinguibilidade	Observações	Similaridade	TAE (%)RNA
1	300	50%	0%
2	500	75%	0%
3	700	87,5%	0%
4	900	93,75%	0%
5	1100	96,875%	0,09%
6	1300	98,43%	14,61

(*) similaridade genética máxima esperada entre as populações de retrocruzamento.

Os resultados foram obtidos com a rede estabelecida com três camadas ocultas e algoritmo de treinamento *Trainbr– Backpropagation*. A descrição da configuração das camadas ocultas encontra-se descritas na Tabela 12, indicando que a variação do número de neurônios de 6 a 15 para a primeira camada, 15 a 40 na segunda, 15 a 40 na terceira camada intermediaria e as combinações das funções de ativação ativação tangente hiperbólica (tansig) e logarítmica (Logsig), foram adequados na solução do problema apresentado. Foi observado que, para as camadas um e dois, houve predomínio da função ativação tangente hiperbólica (tansig) e para a terceira camada intermediaria ouve predominância da logarítmica (Logsig

Tabela 12: Descrição da RNA em relação ao número de neurônios e função de ativação nas camadas ocultas (O1, O2 e O3) e taxa de erro aparente (TEA) nos processos de treinamento (TEAt) e validação (TEAv) nos conjuntos de dados dos 6 cenários de distinguibilidade em alta (Cenário de variáveis A) e baixa herdabilidade(Cenário de variáveis B).

Cenário de	Nós			Função de ativação			Cenário A		Cenário B	
distinguibilidade	O1	O2	O3	O1	O2	O3	TEAt	TEAv	TEAt	TEAv
1	6	10	18	tansig	tansig	logsig	0	0	0	0
2	6	30	30	tansig	tansig	tansig	0	0	0	0
3	15	30	35	tansig	logsig	logsig	0	0	0	0
4	15	35	40	tansig	tansig	logsig	0	0	0	0
5	15	40	40	logsig	tansig	logsig	1,36	0,27	0,72	0,09
6	15	40	40	tansig	tansig	logsig	19	12,53	19,23	14,61

Segundo Hetch & Nielsen (1989) o número de neurônios na camada intermediária utilizadas em problemas de classificação e padrões de filtragem linear é dado por (2i +1), em que i é o numero de variáveis de entrada. Entretanto, os mesmo autores informam que isso não é suficiente para problemas muito complexos como é o caso do presente estudo. Embora haja relatos na literatura de que na segunda camada intermediária haja menos neurônios que na terceira, esse resultado não foi condizente com o observado para esses dados.

A fase crucial para garantir a eficiência de uma RNA é determinar a quantidade e tamanho das camadas ocultas, esta etapa, entretanto, pode ser determinada empiricamente (Braga et al., 2007). A especificação do número e do tamanho das camadas ocultas das RNAs é um ponto crítico para garantir a

capacidade de aprendizagem das características dos conjuntos de dados de treinamento e reconhecer novos dados que são inseridos durante o processo de validação e teste. O número de nós nas camadas intermediárias define a complexidade do modelo de rede neural para descrever as relações e a estrutura inerente aos dados de treinamento (Kavzoglu & Mather, 2003). Neste trabalho o número de camadas ocultas foram estabelecidas arbitrariamente como sendo igual a 3, mas o número de neurônios por camada foi determinado empiricamente investigando, dentre um conjunto de possibilidades, o número que conduzisse a resultados mais satisfatório em termos de taxa de erro aparente.

Segundo Ardö et al., (1997) não existe relação entre a acurácia das RNAs e o numero de camadas intermediárias. No presente trabalho, assim como no trabalho de Pereira (2009) que utilizou RNAs para classificação de populações de milho, foram utilizadas três camadas intermediárias. Devido a complexidade dos dados proveniente da similaridade das populações de retrocruzamento e também pelo número total de indivíduos do estudo as redes neurais artificiais foi gasto um tempo relativamente longo para o processamento dos dados. Este fato ocorre devido ao número de interações entre o número de camadas ocultas, o número de neurônios em cada camada e o número de funções de ativação, demonstrando que seria inviável a utilização de mais camadas e provavelmente a utilização de menos camadas intermediarias apresentaria um desempenho insatisfatório.

Portanto, estudos que evidenciem a eficiência de arquiteturas utilizadas nos programas de melhoramento, seja na predição do valor genético ou na classificação de populações são de grande valia já que facilitariam a escolha da estrutura utilizada nas redes neurais artificiais, pois embora haja grande particularidade nos dados utilizados, ainda pouco se sabe sobre como escolher as melhores estratégias biométricas que são determinantes do tempo e do sucesso na execução do trabalho, ressaltando a importância do presente estudo.

As configurações de rede utilizadas também foram apropriadas para estabelecimento de RNA com erro nulo para quase todas as populações na etapa de treinamento (Tabelas 8 e 9). A mesma eficácia foi observada nas populações de validação, que apresentaram para maioria dos cenários erros nulos e um erro inferior a 19,2% para o Cenário de distinguibilidade 6.

De maneira geral verifica-se que, no processo de treinamento, o erro em classificação ocorreu apenas nas populações dos Cenários de distinguibilidade 5 e 6, em que o nível de similaridade, envolvendo um par de populações, ultrapassava 96% como se espera entre uma população genitora recorrente a quarta e a quinta geração de retrocruzamento. Na Tabela 10 e 11 observa-se que o erro de classificações em treinamento ocorreu de forma menos acentuada do que aquele verificado por meio do uso das funções discriminante.

Resultados satisfatórios também foram encontrados por Barbosa et al., (2011) que trabalharam com 37 acessos de mamão (Carica papaya L.) e oito características quantitativas com o objetivo de avaliar uma estratégia nova para analises da diversidade genética. Eles comparam o desempenho das redes neurais artificiais e da análise discriminante de Anderson na classificação dos acessos. A RNA classificou os acessos em 4 grupos. A análise discriminante de Anderson classificou 91,90% dos acessos nos grupos formados pela RNA. De acordo com a análise discriminante de Anderson, as redes neurais classificaram corretamente 94,44% no grupo 1, 100% no grupo 2, 88,89% no grupo 3 e 87,5% no grupo 4. Os autores concluíram que as redes neurais artificiais classificam de forma eficiente os acessos para estudo de diversidade genética.

O resultado encontrado no presente estudo corroborou com a literatura uma vez que, segundo Braga et al. (2007), as RNAs possuem superioridade de desempenho em relação aos modelos convencionais, sugerindo que as populações possam ser bem discriminadas por esta abordagem. O processo de ampliação por simulação mostrou ser viável em disponibilizar um conjunto de dados de maior proporcionalidade mantendo as propriedades das informações contidas nos dados reais. Neste trabalho considerou ser suficiente preservar, para os dados ampliados, a função de distribuição, considerada normal para cada uma das treze variáveis, e suas estatísticas básicas tais como médias, variância e covariâncias.

CONCLUSÃO

A simulação utilizada foi eficaz em preservar a estrutura genética das populações e descrever a sua dinâmica ao longo de sucessivas gerações de retrocruzamento.

As redes neurais artificiais foram eficientes em classificar populações derivadas de retrocruzamentos até mesmo no conjunto de dados de baixa distinguibilidade em que havia par de populações com mais de 96% de similaridade, como ocorre entre o genitor recorrente e a quinta geração de retrocruzamento.

A estrutura da rede com três camadas intermediárias com de 6 a 15 neurônios na primeira camada, 15 a 40 na segunda camada e 15 a 40 neurônios na terceira camada foram eficientes na classificação das populações de retrocruzamentos estudadas.

Em análises classificatórias a abordagem por meio de redes neurais artificiais se mostra bem superior em relação às técnicas multivariadas de análise discriminantes tais como as propostas por Fisher ou Anderson.

4.6. REFERÊNCIAS BIBLIOGRÁFICAS

ANDERSON T.W. **An Introduction to Multivariate Statistical Analysis.** New York: John Wiley & Sons, 1958, 345 p.

ARDÖ, J.; PILESJÖ, P.; SKIDMORE, **A. Neural networks, multitemporal Landsat Thematic Mapper data and topographic data to classify forest damages in the Czech Republic**. Canadian Journal of Remote Sensing,v. 23, n. 3, p. 217-229, 1997.

Barbosa, C. D., Viana, Al. P. Quintal, S. S. R. Pereira, M. G. "Artificial neural network analysis of genetic diversity in Carica papaya L." **Crop Breeding and Applied Biotechnology**: 224-231, 2011.

BRAGA, A.P.; CARVALHO, A. P. L. F.; LUDERMIR, T. B. **Redes Neurais Artificiais - Teoria e aplicações** – 2. ed. Rio de Janeiro: LTV, 2011. 226p.

BORÉM, A.; Miranda, G. V. 2009. **Melhoramento de Plantas.** 5ta ed. Editora UFV, Viçosa, 2009, 523p.

CRUZ, C. D. GENES - a software package for analysis in experimental statistics and quantitative genetics. **Acta Scientiarum**. v.35, n.3, p.271-276, 2013.

CRUZ, C. D.; REGAZZI, A. J.; CARNEIRO, P. C. S. **Modelos biométricos aplicados ao melhoramento genético**- ed.5. Viçosa, UFV. 480 p. 2012.

CRUZ, C. D. **Programas Genes: análise multivariada e simulação**. Viçosa, ed. UFV, 2006. 175p.

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos Biométricos Aplicados ao Melhoramento Genético** - Vol 2. 2.ed. Viçosa: UFV, 2006. 585p.

FISHER, R.A. The use of multiple measurements in taxonomic problems. Annals of **Eugenics**, v.7, p.179-188, 1936.

HAYKIN, S.S. **Redes Neurais: princípios e práticas** / Symon Haykin; trad Paulo Martins Engel. 2 ed. – Porto Alegre : BOOKMAN, 2001. 900p.

KAVZOGLU, T. **An investigation of the design and use of feed forward artificial neural networks in the classification of remotely sensed images**. University of Nottingham, 2001.

KAVZOGLU, T.; MATHER, P. The use of backpropagating artificial neural networks in land cover classification. **International Journal of Remote Sensing**, v. 24, n. 23, p. 4907-4938, 2003.

KAVZOGLU, T. Increasing the accuracy of neural network classification using refined training data. **Environmental Modelling & Software**,v. 24, n. 7, p. 850-858, 2009.

Hecht-Nielsen. R. **Theory of the back-propagation neunn network.** In Proceedings of the 1989 International Joint (Confirence on Neural Networks [pp 1:593-6061 New York: [EEF Press.

MATLAB version 7.10.0. Natick, Massachusetts: The Math Works Inc., 2010.

NEI, M. Genetic distance between populations. **American Naturalist**.Chicago, v. 106, p. 238-292, 1972.

PEREIRA, J.J. **Análises de agrupamento e discriminante no melhoramento genético - aplicação na cultura do arroz (*Oryza sativa L.*).** (Tese de Doutorado), Universidade Federal de Viçosa: UFV, 1999. 191 p.

PEREIRA, T.M **Discriminações de populações com diferentes graus de similaridade por redes neurais artificiais.** Viçosa: UFV, 2009. 73 p.

VENTURA, R.V.; SILVA, M.A.; MEDEIROS, T.H.; DIONELLO, N.L.; MADALENA, F.E.; FRIDRICH, A.B.; VALENTE, B.D.; SANTOS, G.G.; FREITAS, L.S.; WENCESLAU, R.R.; FELIPE, V.P.S.; CORRÊA, G.S.S. Uso de redes neurais artificiais na predição de valores genéticos para peso aos 205 dias em bovinos da raça Tabapuã. **Arq.Bras. Med. Vet. Zootec.**, v.64, n.2, p.411-418, 2012.

5.0 Conclusões Gerais

A simulação utilizada foi eficaz em preservar a estrutura genética das populações e descrever a sua dinâmica ao longo de sucessivas gerações de retrocruzamentos.

A estrutura de similaridade genética populacional, delineada em populações derivadas de retrocruzamento, foi apropriadamente representada por meio de técnicas multivariadas aplicadas em dados fenotípicos.

A simulação dos dados fenotípicos e genotípicos realizada se mostrou eficiente para estudos de discriminação de populações com alto grau de similaridade utilizando redes neurais artificiais.

Técnica de agrupamento hieráquica e de projeção de distâncias no plano são eficientes em agrupar e descrever o padrão de similaridade entre populações considerando os diversos níveis de similaridade. A estrutura de similaridade genética populacional, delineada em populações derivadas de retrocruzamento, foi apropriadamente representada por meio de técnicas multivariadas aplicadas em dados fenotípicos.

A obtenção de dados experimentais preservando propriedades estatísticas, tais como média e variância é eficiente e pode ser realizada usando princípios estocásticos de distribuição tais como enunciado no teorema de Box-Muller.

Os conjuntos de dados preservados ou ampliados podem ser apropriadamente gerados e utilizados em estudos de redes neurais que demandam grande quantidade de informações para fins de treinamento e aprendizagem. Os dados apropriadamente simulados, por preservarem informações essenciais, podem agregar ou substituir dados experimentais nem sempre viáveis.

A ampliação do conjunto de dados para fins de análise classificatória deve ser feita considerando a função de distribuição das variáveis, no vetor de médias e na matriz de variâncias e covariância estimada em cada população mensurada.

A utilização de RNA mostrou ser mais satisfatória em apontar a diferenciação de todas as populações do que as técnicas multivariadas e as análises descriminante de Anderson e Fisher.