UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
EMBRAPA MANDIOCA E FRUTICULTURA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS
CURSO DE MESTRADO

ESTRUTURA POPULACIONAL EM LOCI MULTIVARIADOS E REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DE RENDIMENTO EM TABACO (*Nicotiana tabacum* L.)

LUCIANA LIMA DOS REIS

CRUZ DAS ALMAS - BAHIA
FEVEREIRO DE 2021

ESTRUTURA POPULACIONAL EM LOCI MULTIVARIADOS E REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DE RENDIMENTO EM TABACO (*Nicotiana tabacum* L.)

LUCIANA LIMA DOS REIS
Engenheira Florestal
Universidade Federal do Recôncavo da Bahia (UFRB), 2018.

Dissertação submetida ao Colegiado de Curso do Programa de Pós-Graduação em Recursos Genéticos Vegetais da Universidade Federal do Recôncavo da Bahia e Embrapa Mandioca e Fruticultura, como requisito parcial para obtenção do Grau de Mestre em Recursos Genéticos Vegetais.

Orientador: Prof. Dr. Ricardo Franco Cunha Moreira
Coorientador: Prof. Dr. Jair Wyzykowski

CRUZ DAS ALMAS - BAHIA
FEVEREIRO DE 2021

FICHA CATALOGRÁFICA

R375e

Reis, Luciana Lima dos.

Estrutura populacional em loci multivariados e redes neurais artificiais na predição de rendimento em tabaco (*Nicotiana tabacum* L.) / Luciana Lima dos Reis._ Cruz das Almas, Bahia, 2021.

60f.; il.

Orientador: Ricardo Franco Cunha Moreira.

Coorientador: Jair Wyzykowski.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas, Mestrado em Recursos Genéticos Vegetais.

1.Diversidade genética – Fumo. 2.Inteligência computacional – Redes neurais (Computação). 3.Fenótipo – Análise. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Agrárias, Ambientais e Biológicas. II.Título.

CDD: 576.58

Ficha elaborada pela Bibliloteca Central de Cruz das Almas - UFRB.

Responsável pela Elaboração - Antonio Marcos Sarmento das Chagas (Bibliotecário - CRB5 / 1615).

(os *dados para catalogação foram enviados pelo usuário via formulário eletrônico*).

UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS AGRÁRIAS, AMBIENTAIS E BIOLÓGICAS
EMBRAPA MANDIOCA E FRUTICULTURA
PROGRAMA DE PÓS-GRADUAÇÃO EM RECURSOS GENÉTICOS VEGETAIS
CURSO DE MESTRADO

ESTRUTURA POPULACIONAL EM LOCI MULTIVARIADOS E REDES NEURAIS ARTIFICIAIS NA PREDIÇÃO DE RENDIMENTO EM TABACO (*Nicotiana tabacum* L.)

Comissão examinadora da defesa de dissertação de
Luciana Lima dos Reis

Aprovado em: 12 de Fevereiro de 2020

Prof. Dr. Ricardo Franco Cunha Moreira
Universidade Federal do Recôncavo da Bahia, Bahia.
(Orientador)

Prof. Dr. Liniker Fernandes da Silva
Universidade Federal do Recôncavo da Bahia, Bahia
(Examinador Interno)

Prof. Dr. Angelo Gallotti Prazeres
IF Baiano, Bahia
(Examinador Externo)

DEDICATÓRIA

Dedico este trabalho a Deus por não ter me deixado desistir dos meus sonhos e me fortalecer sempre.

A minha saudosa vó Zuleide e minha mãe Anaildes, que acreditaram em mim desde os meus primeiros passos. Nada disso teria sentido se vocês não existissem na minha vida.

AGRADECIMENTOS

Agradeço primeiramente a Deus pela oportunidade de estar concluindo mais uma jornada árdua e de grande aprendizado. Guiando meus passos a cada momento vivido, presenciando a magnitude da sua benção na minha vida.

A minha queridíssima vó Zuleide e minha mãe Anaildes pelo amor, amizade, encorajamento e compreensão dedicados a mim, com seus ensinamentos a cada dia da minha vida. Me mostrado o caminho a ser seguido para a realização de todos os meus sonhos.

A meu marido Lucas pelo amor, compreensão e a familia linda que formamos juntos, tornando meus dias mais fáceis de serem caminhados ao seu lado.

A minha filha Maria Luisa por ter sido um raio de luz na minha vida, trazendo amor, paz e muita sabedoria, me mostrando assim a importância do amor mais sincero.

Ao meu Orientador, Ricardo Franco por toda disposição, incentivo, sorriso e conversa que me trouxe sabedoria e norteou a busca por conhecimento. E pela amizade que desenvolveu-se em cada conversa.

Ao Prof. Dr. Jair Wyzykowski por toda disposição, incentivo e suporte para a realização deste trabalho.

Aos meus colegas do Programa de Pós-Graduação em Recursos Genéticos Vegetais pelo carinho e apoio.

A Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), pela concessão da bolsa de mestrado.

A todos que contribuíram diretamente e indiretamente, deixo aqui o meu muitíssimo
obrigado!

Estrutura populacional em locos multivariados e Redes Neurais Artificiais na predição de rendimento em tabaco

Autor: Luciana Lima dos Reis
Orientador: Prof. Dr. Ricardo Franco Cunha Moreira
Co-orientador: Prof. Dr. Jair Wyzykowski

Resumo: A fonte primaria de informações sobre indivíduos e suas populações, são seu genoma e suas características fenotípicas. Elas são capazes de descrever detalhadamente os acontecimentos biológicos sofridos pelos indivíduos no espaço-tempo e as modificações ocasionadas nas alterações dos genomas. Com vista nisto, a pesquisa tem buscado predizer os possíveis eventos e postular soluções para as mais diversas situações genéticas, com ajuda de softwares que possam simular o avanço no tempo e suas características diversas. Atualmente a biotecnologia tem investido, de forma significativa, na compreensão e aprimoramento de dados simulados, por serem uma ferramenta viável no processo de acompanhamento evolutivo, gerando informações robustas e maior número de cenários complexos e mais realistas. Desta forma, o presente trabalho foi realizado em dois capítulos, onde: I) Simulação de dados para predição genética de populações naturais multi-locus avançadas no tempo; II) Predição de fenótipos de tabaco através das redes neurais artificiais, objetivando demonstrar a importância do uso de dados simulados no planejamento de pesquisas e utilização de equações matemáticas na predição de dados fenotípicos, por meio de resultados e informações de alta confiabilidade, que possibilitam através destes a produção de informações e conhecimento que subsidiará pesquisadores na estratégia de conservação de populações e indivíduos de forma rápida e altamente eficiente.

Palavras-chaves: Diversidade genética, inteligência computacional, previsão fenotípica.

Population structure in multivariate loci and Artificial Neural Networks in predicting tobacco yield

Author: Luciana Lima dos Reis
Advisor: Prof. Dr. Ricardo Franco Cunha Moreira
Co-Advisor: Prof. Dr. Jair Wyzykowski

Abstract: The primary source of information about individuals and their populations is their genome and phenotypic characteristics. They are able to describe in detail the biological events suffered by individuals in space-time and the changes caused by changes in genomes. With this in mind, the research has sought to predict possible events and postulate solutions for the most diverse genetic situations, with the help of software that can simulate the advance in time and its diverse characteristics. Currently, biotechnology has significantly invested in the understanding and improvement of simulated data, as they are a viable tool in the evolutionary monitoring process, generating robust information and a greater number of complex and more realistic scenarios. Thus, the present work was carried out in two chapters, where: I) Simulation of data for genetic prediction of natural multi-locus populations advanced in time; II) Prediction of tobacco phenotypes through artificial neural networks, aiming to demonstrate the importance of using simulated data in research planning and the use of mathematical equations in the prediction of phenotypic data, through highly reliable results and information, which make it possible through of these, the production of information and knowledge that will support researchers in the strategy of conserving populations and individuals in a fast and highly efficient manner.

Keywords: Genetic diversity, computational intelligence, phenotypic prediction.

SUMÁRIO

Introdução

O desenvolvimento de modelos matemáticos complexos que possam auxiliar o conhecimento sobre os mais diversos assuntos, tem sido pauta de estudo em muitas áreas do conhecimento biológico (PENG et al., 2007; CURRAT et al.,2019). Viabilizando a projeção de situações biológicas distintas, que convencionalmente levariam anos para serem observadas e detectadas por conta dos altos custos de manutenção da mesma.

A viabilidade de simular dados em sua ampla magnitude, proporciona uma saída viável e de baixo custo, que traz a evolução genética e suas magnitudes como fator principal na elaboração de modelos. Pois, segundo Yuan et al (2012), a complexidade dos dados computacionais é importante para satisfazer os interesses das variações demográficas e populacionais exigidas pela pesquisa de avanço temporal.

Para subsidiar as pesquisas, alguns modelos têm utilizado polimorfismos de nucleotídeo único (SNPs) como base e *output* de dados, por serem marcadores pontuais, aumentando a acurácia e a complexidade, estreitando consideravelmente a simulação e qualidade dos dados obtidos (STRAND, 2002; FRANÇOIS & CAYE, 2017; MEYER & BIRNEY, 2018). Por serem marcadores pontuais e um dos principais causadores de variações genômicas e fenotípicas, estes podem ser produzidos em abundancia nos genomas simulados (MEYER & BIRNEY, 2018; GAYNOR, 2020).

Alguns softwares como o R (R Development Core Team, 2019) e GENES (CRUZ, 2006) podem ser utilizados para geração de indivíduos e populações com avanço temporal. Onde a constituição de populações são realizadas através da combinação de fatores de mutação, seleção natural, cruzamentos e outros fatores, capazes de atuar sobre as populações.

Os avanços compreendem a seleção evolutiva e os fatores que atuam nos ambientes, onde a distância demográfica pode ser considerada fator atuante na diversidade entre e dentre as populações (CURRAT et al., 2019). Sendo detectáveis através de modelos biométricos, que utilizados de forma adequada, descreve diversos eventos sofridos pelas populações.

Neste sentido, as redes neurais artificiais (RNAs) tem se destacado no âmbito da predição de dados, por serem uma abordagem alternativa para predição

(BINOTI et al., 2013; BINOTI et al., 2014; BHERING et al., 2015). As RNAs são modelos de aprendizagem baseados em neurônios humanos interligados, com recomendações pré-definidas pelo usuário (EBERHART & RUSSELL, 1966; BINOTI, 2010; BINOTI et al., 2012). Estas não necessitam do conhecimento das relações das variáveis, colaborando para a utilização em diversas áreas do conhecimento (NASCIMENTO, 2013; BINOTI et al.,2014).

Desta forma, a utilização de modelos biométricos e softwares de simulação de dados tornam-se eficiente na mensuração de variáveis de natureza quantitativa e qualitativas e na compreensão dos eventos genéticos. Pois são capazes de captar variações de acordo a origem auxiliando na compreensão da relação de indivíduos e populações com as variáveis envolvidas, por meio de fatores genéticos ou ambientais.

Revisão de literatura

1. Tabaco (*Nicotiana tabacum* L.)

A espécie *Nicotiana tabacum* L. comumente conhecida como tabaco ou fumo, pertencente à família Solanaceae, apresenta como centro de origem a América tropical (SOARES et al., 2008; FLORA DO BRASIL, 2020). São plantas, que apresentam flores grandes e rosadas, caule ereto, fruto em forma de cápsula ovoide e podem atingir até 2 metros de altura na fase adulta (BOIEIRO, 2008).

Conforme Oliveira (2006), as principais utilizações do fumo estão ligadas a produção de charuto, cigarrilha e cigarros. Com vista nisto, o fumo é uma das culturas não alimentícias mais cultivadas do mundo (LIMA, 2006). Possuindo este uma ampla produção entre e os continentes da Ásia (65,7%), Américas (18,7%), África (12,7%) e Europa (2,8%) (FAOSTAT, 2021).

No ano de 2019, o Brasil produziu 769,801 toneladas de tabaco, ocupando assim o terceiro lugar no ranking entre os maiores produtores mundiais desta cultura (FAOSTAT, 2021). De acordo com Mesquita & Oliveira (2003), na Bahia o fumo foi introduzido no ano de 1757, no município de Cachoeira com introdução de indústrias de charuto. Sendo considerado, o segundo maior estado produtor com participação de 41% na produção do Nordeste (MESQUITA & OLIVEIRA, 2003; OLIVEIRA, 2006). Tendo como base de cultivo, as variedades Sumatra,

Virgínia e Brasil-Bahia, com produção voltada para fumo em folha (OLIVEIRA, 2006).

A utilização de genótipos de tabaco diversificado, tem relação direta com as condições edafoclimáticas, que visam o maior rendimento dos mesmos, com o cultivo em regiões semelhantes às de origem (LIMA, 2006). Com isso, o conhecimento dos germorplasmas existentes, tornam-se importantes para verificação e seleção das características morfoagronomicas de interesse para o plantio. Pois o conhecimento sobre os genótipos, nortea o cultivo das variedades para as regiões e finalidades de uso, de acordo com as características genotípicas e fenotípicas de interesse dos pesquisadores (MESQUITA & OLIVEIRA, 2003).

Desta forma, se faz importante a caracterização das variedades para observação de características de interesse e determinação da variabilidade existente entre os genótipos de tabaco (GUIMARÃES et al., 2007).

2. Marcadores moleculares codominantes

O conhecimento das características presentes no genoma de espécies vegetais apresenta grande importância para a compreensão do cenário genético de populações e suas gerações, sendo capaz de fornecer informações por meio dos marcadores moleculares, que auxiliarão em estudos para a conservação da diversidade e melhoramento genético (SOUZA, 2015).

Os marcadores moleculares permitem a descrição dos genomas e a obtenção de outras informações de interesse agronômico para a cultura a ser avaliada (CAMPOS et al., 2013). Diante desses, torna-se possível o mapeamento e manipulação de locos para verificar a expressão de características, predição de híbridos simples, estudo de filogenias, entre outros (EUCLYDES & GUIMARÃES, 1997).

Com o avanço da biotecnologia é possível obter uma ampla gama de marcadores. No entanto, a eficácia destes marcadores está diretamente relacionada ao conhecimento das suas aplicações, bem como suas vantagens e desvantagens (SEMAGN et al., 2006).

Para Turchetto-Zolet et al. (2017), conforme sua herança alélica, os marcadores podem ser subdivididos em dominantes e codominantes. Os dominantes identificam apenas indivíduos homozigotos em suas análises, através da detecção de presença/ausência do alelo de interesse (HOFFMANN & BARROSO, 2006). RAPD (*Randon Amplified Polymorphic DNA*) (WILLIAMS et al., 1990; WELSH & MCCLELLAND, 1990) e AFLP (*Amplified Fragment Length Polymorphism*) (VOS et al., 1995) são exemplos de marcadores dominantes.

Os marcadores codominantes são capazes identificar indivíduos heterozigotos e homozigotos (TURCHETTO-ZOLET et al., 2017). Dentre estes, tem-se: SSR (*Simple Sequence Repeats*) (TAUTZ, 1989), RFLP (*Restriction Fragment Length Polymorphism*) (BOTSTEIN et al., 1980), SNP (*Single Nucleotide Polymorphism*) (CHING et al., 2002).

Para Caetano (2009), os SSR, surgiram como alternativa de genotipagem para os estudos em animais, proporcionando a construção dos primeiros mapas genéticos. Os RFLP foram os primeiros marcadores a serem utilizados na década de 80, nos estudos de animais (CHARDON et al., 1985; BECKMANN et al., 1986; GEORGES et al.,1987; CAETANO, 2009).

Segundo os autores François & Caye (2017) e Caetano (2009), os marcadores SNP são os mais utilizados pelos cientistas atualmente, por possuírem benefícios múltiplos quando comparado a tecnologia dos demais. Por serem na maioria das vezes de caráter bi-alélico, estes são encontrados em grande quantidade no genoma de indivíduos, podendo subsidiar estudos populacionais com simulação de dados, por apresentarem polimorfismos ligados a características fenotípicas (MEYER & BIRNEY, 2018; GAYNOR, 2020).

3. Simulação de dados

O desenvolvimento de mecanismos que possam subsidiar o conhecimento sobre genes e genomas de plantas, são frequentemente utilizados para observar o comportamento de populações a longo prazo. Com base nas variações ocorridas no espaço-tempo, a simulação de dados tem facilitado a busca por compreensão sobre os possíveis eventos evolutivos como deriva genética e mutações, que podem ocorrer em uma população natural (ASSIS, 2005; FERREIRA, 2007; DAETWYLER et al., 2013).

A complexidade da simulação de dados está ligada a necessidade de imitar a realidade dos eventos genéticos, com a ampliação das hipóteses e rapidez nas respostas aos vários aspectos dos dados em estudo, permitindo a utilização de diferentes modelos de simulação que agreguem diversas fontes de variabilidade em níveis hierárquicos (ASSIS, 2005; SUN, 2012; DAETWYLER et al., 2013).

Para Sun (2012), no melhoramento genético, a utilização de dados computacionais está diretamente relacionada ao processo de planejamento, em função de sua maior viabilidade e menor custo. Sendo comumente empregados para: predição de valores genéticos, simulação de genes e genomas e dinâmica genômica em diferentes períodos de tempo (MUIR, 1997; ASSIS, 2005). Já a conservação genética, utiliza a simulação de dados como maneira para compreender o padrão evolutivo das populações, prevendo os possíveis riscos de extinção, força de seleção e endogamia (HOBAN et al., 2012; HOBAN, 2014).

Na literatura, podem ser encontrados vários modelos matemáticos para subsidiar a simulação de dados (KINGMAN, 2000; YUAN et al. 2017). No entanto, não existe um único modelo que possa descrever todos os eventos genéticos de interesse (HOBAN et al.,2012). Desta forma, a simulação dos parâmetros é ajustável para as variações dos eventos, facilitando sua adequação ao interesse dos pesquisadores e tornando-se ferramenta de grande importância para a ciência.

Neste sentido, existem softwares que possibilita a aplicação de grande variedade de modelos matemáticos para obtenção de dados simulados. De acordo com Hoban et al. (2012), na literatura podem ser encontrados uma multiplicidade de simuladores que conseguem subsidiar a aquisição de dados com presença de polimorfismo. Dentre os softwares existentes, pode-se citar: programa R (R DEVELOPMENT CORE TEAM 2019), GENES (CRUZ, 2006), GENEPOP (RAYMOND & ROUSSET,1995) e ARLEQUIN (EXCOFFIER et al.,2005).

Dentre os softwares citados anteriormente, pode-se destacar o R, como sendo um dos mais versáteis para incorporar os modelos, sendo possível verificar a sua versatilidade por meio dos pacotes que possuem implementação e extensão de outros programas, facilitando o trabalho dos pesquisadores, na utilização de um único software com várias alternativas de simulação dos dados.

3.1 Genética populacional avançada no tempo

A aquisição de dados populacionais distribuídos no espaço-tempo, tem tornado a simulação populacional uma ferramenta altamente eficiente, por apresentar versatilidade nos modelos de simulação (ANDRELLO & MANEL, 2015). A utilização desta ferramenta é considerada uma estratégia viável e de baixo custo, para estudos de conservação e melhoramento genético (FAUX et al., 2016).

Conforme Strand (2002), o avanço da tecnologia dos programas computacionais proporciona viabilidade para simulação de cenários possíveis para as populações naturais, com a flexibilidade na implementação dos parâmetros nas análises e abordagem diferenciada das possibilidades. A tecnologia do software R, tem disponibilizado uma ampla gama de ferramentas para a simulação da dinâmica dos processos evolutivos no estudo das populações de organismos vivos (FALCONER & MACKAY,1996; HOBAN et al., 2012).

No programa R, diversos pacotes foram desenvolvidos para atender a demanda de informações preditivas de populações. Como pode ser observado em trabalho de Wimmer et al. (2012), no qual foi proposto o pacote “synbreed” como uma alternativa para predição genômica de populações, com simulação de dados para genótipos e fenótipos, utilizando o método de validação cruzada.

De acordo com Faux et al. (2016), as populações podem ser geradas pela simulação de aspectos como seleção genômica e edição de genoma, pelo usuário do pacote “AlphaSimR”. Pois a integração de biotecnologias ao pacote favorece a contribuição real para o desenvolvimento de pesquisas voltadas ao melhoramento genético com ênfase em populações de gerações avançadas.

Os métodos para variação demográfica e avanço no tempo, com base na estrutura da população genética, foram abordados por Strand (2002) e Andrello & Manel (2015) em desenvolvimento dos pacotes “Metasim” e “MetaPopGen”, respectivamente. No “Metasim”, a simulação dos dados pode variar desde a análise do estádio de vida (mudas, juvenis, adultos vegetativos, reprodutivos e dormentes) do indivíduo até os cenários evolutivos realistas, baseados em modelos próprios (STRAND, 2002). Segundo Andrello & Manel (2015), para o “MetaPopGen”, os parâmetros são manipulados para obtenção de populações

com base em número de genótipos, sendo este indicado para projetos de cenários complexos e populações numerosas.

Desta forma, torna-se importante salientar que os dados e a implementação das populações dependerão da demanda de cada usuário. Tais ferramentas relatadas, irão subsidiar qualquer estudo de dados para populações avançadas no tempo, se o pesquisador tiver o conhecimento sobre os modelos e métodos empregados para executar as simulações.

3.2 Populações de gerações avançadas (Fn)

Em programas de melhoramento e conservação genética, a condução do trabalho tem como base a busca de respostas sobre as populações analisadas, a fim de verificar o comportamento de seus alelos após o cruzamento dos indivíduos, auxiliando na escolha das linhagens mais promissoras (SANTOS et al., 2001).

Desta forma, com o intuito de aumentar a eficiência dos programas de melhoramento, utilizam-se as populações avançadas (Fn), que serão avaliadas mediante o desempenho dos alelos de interesse, de acordo com os objetivos do programa (SANTOS et al.,2001; FERREIRA, 2007).

Segundo Ramalho et al. (2001), as populações avançadas podem ser obtidas por métodos distintos. No entanto, a utilização dos métodos depende do conhecimento prévio das suas metodologias, para obter a eficácia na predição das populações segregantes (VALÉRIO et al., 2009).

Entre os diversos métodos, podem ser encontrados trabalhos na literatura envolvendo cruzamentos dialélicos para a cultura do trigo (*Triticum aestivum* L.) (PIMENTEL et al.,2013a; PIMENTEL et al.,2013b) e melancia (SOUZA et al., 2013), trabalhos com o método Jinks & Pooni (1976) para o arroz (*Oryza sativa* L.) (SANTOS et al.,2001) e feijão (*Phaseolus vulgaris* L.) (ROCHA et al.,2015).

3. Análises para detecção de variabilidade genética

A detecção da variabilidade genética num determinado grupo de indivíduos, depende da análise adequada dos dados, para extração de todas as informações

com eficiência, pois a caracterização de variabilidade é crucial para o programa de melhoramento e conservação (CAVALCANTE & LIRA, 2010).

Em vista disto, revisar os métodos biométricos descritos na literatura, torna-se parte importante do trabalho que norteará a avaliação dos dados. Assim, pode-se identificar genitores em potencial para cruzamentos futuros, incrementar informações a respeito de culturas, unidades taxonômicas e compreender o padrão biogeográfico das populações (SILVA & RUSSO, 2000; FERREIRA, 2007; PESSANHA et al., 2011).

Segundo Cruz et al. (2020), no início dos estudos sobre a diversidade biológica, os cientistas se norteavam por dados de natureza fenotípica. No entanto, com o avanço das tecnologias, a variabilidade genética pode ser detectada em nível de DNA.

De acordo com Cavalcante & Lira (2010), a utilização de dados genotípicos é considerada enriquecedora e complementar nas informações extraídas por dados fenotípicos. Através deles, os pesquisadores conseguem eliminar duplicatas em programas de melhoramento, verificar o fluxo gênico e avaliar a diversidade existente entre e dentro de populações (ROBINSON, 1998; CRUZ et al., 2020).

O modelo de análise multivariada, pode ser empregado para caracterizar a variabilidade das populações, pela da utilização de caracteres qualitativos e quantitativos de forma simultânea (MOURA et al.,1999). As técnicas de análise multivariada mais empregadas são a de variáveis canônicas, métodos aglomerativos e análise dos componentes principais (CRUZ & CARNEIRO, 2003, CAVALCANTE & LIRA, 2010).

Para o estudo das variáveis canônicas e análise dos componentes principais, são utilizados gráficos de dispersão, os quais interpretados através de plano cartesiano. Já o método aglomerativo, baseia-se nas medidas de dissimilaridade obtidas usualmente pela distância euclidiana e a distância generalizada de Mahalonobis (D^2) (CRUZ, 2005; CAVALCANTE & LIRA, 2010). Em ambas as metodologias, o objetivo é constatar a formação de grupos com indivíduos semelhantes e a divergência entre os mesmos, podendo ser observado o uso das técnicas e sua eficiência em várias publicações (PERONI et al.,1999; COELHO et al.,2007; CORREIA & GONÇALVES, 2012; SEBIM et al., 2016; PEREIRA et al., 2019).

Outra forma para obter a caracterização da variabilidade, é por meio de parâmetros genéticos populacionais, como a heterozigosidade esperada e observada, número de alelos e índice de Shannon-Wiener; os quais detectam o polimorfismo existente (FERREIRA, 2007). Para Dias (1998), as informações sobre a diversidade, auxiliam no entendimento da estrutura da população, verificando assim as ações evolutivas sofridas pela mesma e ajudando a predizer a magnitude das variações sujeitas na população (CRUZ et al., 2020).

4. Redes neurais artificias (RNAs)

As redes neurais artificias (RNAs), são heurísticas que buscam a inserção da relação entre variáveis, com base no modelo de neurônios semelhantes ao do cérebro humano, em que estes possuem relações entre si, para obtenção da resposta final do trabalho proposto (FERNEDA, 2006; BINOTI, 2010; BINOTI et al., 2013). Obtendo características sobre os dados que não são apresentados explicitamente (KOVÁCS, 1996).

As RNAs fazem a interação dos neurônios de acordo com seus pesos, podendo ser ajustado e diferenciado através do mesmo, sendo o conjunto de neurônios responsável pela aprendizagem, capazes de deduzir os resultados de acordo com o exemplo ao qual foi implementado (BRAGA et al., 2000; FERNEDA, 2006; FIORIN et al. 2011). Existem diversas arquiteturas para as RNA's sendo destacada a *Adaline* (*Adaptive Linear Neuron*) apresentado por Widrow e Hoff no ano de 1960 e *Percepton,* proposto por Frank Rosenblat em 1958 (FIORIN et al. 2011).

Atualmente, as *Perceptrons de Multiplas Camadas* (*Multilayer Perceptron* - MLP) são as mais requisitadas para estudos, por ser de fácil utilização (FIORIN et al. 2011). Esta apresenta-se dividida em camadas de entrada, intermediárias e saída.

A camada de entrada é constituída por uma ou mais variáveis que estarão relacionadas (sinais de entrada), a variável alvo da análise (LACERDA, 2019). As camadas intermediárias ou ocultas são responsáveis pelas conexões, processamento e aprendizagem dos neurônios, extraindo as correlações entre as variáveis e suas características. Já na de saída, ocorre a obtenção dos dados processados (FIORIN et al.,2011; LACERDA, 2019).

Com vista no desempenho e capacidade de resposta, atualmente as RNAs tem sido aplicada para verificação de diversidade entre indivíduos (BRASILEIRO et al., 2015), para filtragem de snps (SILVA, 2013), na correlação de valor genético e ambiente (BATTEY et al., 2019), predição de volume de madeira (BHERING et al., 2015) e identificação de genótipos com características adaptativas (TEODORO et al., 2015).

Referências

ANDRELLO, M.; MANEL, S. MetaPopGen: anrpackage to simulate population genetics in large size metapopulations. **Molecular Ecology Resources**, v. 15, n. 5, p. 1153-1162, 27 jan. 2015.

ASSIS, Giselle Mariano Lessa de. **Efeito do número de genes na avaliação genética utilizando dados simulados**. 2005. 102 f. Tese (Doutorado). Programa de pós-graduação em Genética e Melhoramento- Universidade Federal de Viçosa, Viçosa. 2005.

BATTEY, C.J; RALPH, P. L.; KERN, A. D. Predicting Geographic Location from Genetic Variation with Deep Neural Networks. **bioRxiv,** Dezembro, 2019.

BRAGA, A. P.; CARVALHO, A. C. P. L. F.; LUDEMIR, T. B. **Redes neurais artificiais**: **teoria e aplicações**. Rio de Janeiro: LTC, 2000.

BRASILEIRO, B.P.; MARINHO, C.D.; COSTA, P.M. de A.; CRUZ, C.D.; PETERNELLI, L.A.; BARBOSA, M.H.P. Selection in sugarcane families with artificial neural networks. **Crop Breeding and Applied Biotechnology**, v.15, p.72-78, 2015.

BECKMANN, J.S.; KASHI, Y.; HALLERMAN, E. M. Restriction fragment length polymorphism among Israeli Holstein-Friesian dairy bulls. **Animal Genetics** , v.17, n.1, p.25-38, 1986.

BHERING, L.L.; CRUZ, C.D.; PEIXOTO, L. de A.; ROSADO, A.M.; LAVIOLA, B.G.; NASCIMENTO, M. Application of neural networks to predict volume in eucalyptus. **Crop Breeding and Applied Biotechnology**, v.15, p.125-131, 2015. DOI: 10.1590/1984-70332015v15n3a23.

BINOTI, M. L. M. S. **Redes neurais artificiais para prognose da produção de povoamentos não desbastados de eucalipto**. 2010. 54f. Dissertação (Mestrado) Programa de pós-graduação em Ciência Florestal – Universidade Federal de Viçosa, Viçosa, MG, 2010.

BINOTI, D. H. B.; BINOTI, M. L. M. S.; LEITE, H. G.; SILVA, A. Redução dos custos em inventário de povoamentos equiâneos utilizando redes neurais artificiais. **Rev. Bras. Ciênc. Agrár.** Recife, v.8, n.1, p.125-129, 2013.

BINOTI, M. L. M. S.; BINOTI D. H. B; LEITE, H. Aplicação de redes neurais artificiais para estimação da altura de povoamentos equiâneos de Eucalipto. **Revista Árvore**, Viçosa-MG, v.37, n.4, p.639-645, 2013.

BINOTI, M.L.M.S.; BINOTI D. H. B; LEITE, H. G.; GARCIA, S. L. R; FERREIRA, M. Z.; RODE, R.; SILVA, A. A. L. Redes neurais artificiais para estimação do volume de árvores, **Revista Árvore**, Viçosa-MG, v.38, n.2, p.283-288, 2014.

BOIEIRO, M. **Tabaco**. Portugal, 2008. Disponível em: <http://www.institutohipocrates.pt/index.php/medicinasnaoconvencionais/fitoterapia/192-tabaco.html>. Acesso em: 31/12/2021.

BOTSTEIN, D.; WHITE, R.; SKOLNICK, M.; DAVIS, R.W. Construction of a genetic map in man using restriction fragment length polymorphisms. **Am. J. Hum. Genet.**, v.32, 1980, p.314–331.

CAETANO, A. R. Marcadores SNP: conceitos básicos, aplicações no manejo e no melhoramento animal e perspectivas para o futuro. **R. Bras. Zootec**., v.38, p.64-71, 2009.

CAMPOS, G.; HICKEY, J. M.; PONG-WONG, R.; DAETWYLER, H. D.; CALUS, M. P. L. Whole-Genome Regression and Prediction Methods Applied to Plant and Animal Breeding. **Genetics**, v. 193, p.327–345, 2013.

CAVALCANTE, M.; LIRA, M. A. Variabilidade genética em *Pennisetum purpureum* Schumacher. **Revista Caatinga**, Mossoró, v. 23, n. 2, p. 153-163, 2010.

CHARDON, P.; VAIMAN, M.; KIRSZENBAUM, M.; et al. Restriction fragment length polymorphism of the major histocompatibility complex of the pig**.** **Immunogenetics** ,v.21, n.2, p.161-71, 1985.

CHING, A.; CALDWELL, K.S.; JUNG, M.; DOLAN, M. ;SMITH, O.S. SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines. **BMC Genetics**, 2002, 3:19-32.

COELHO, C. M. M.; COIMBRA, J. L. M.; SOUZA, C. A.; BOGO, A.; GUIDOLINI, A. F. Diversidade genética em acessos de feijão (*Phaseolus vulgaris* L.). **Ciência Rural**, v.37, n.5, 2007.

CORREA, A. M.; GONÇALVES, M. C. Divergência genética em genótipos de feijão comum cultivados em Mato Grosso do Sul. **Rev. Ceres**, Viçosa, v. 59, n.2, p. 206-212, 2012.

CRUZ, C. D. **Princípios de genética quantitativa***.* Viçosa, MG: UFV, 2005. 394 p.

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

CRUZ, C.D.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. Viçosa: UFV, 2003. 579p.

CRUZ, C.D.; REGAZZI, A.J.; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético**. 4.ed. Viçosa: Ed. da UFV, 2012. 514p.

CRUZ, C.D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética.** Viçosa : UFV, 2020, 614p.

CURRAT, M.; ARENAS, M.; QUILODRA`N, C. S.; EXCOFFIER, L.; RAY, N. SPLATCHE3: simulation of serial genetic data under spatially explicit evolutionary scenarios including long-distance dispersal**. Bioinformatics**, v.35, n.21, p. 4480–4483, 2019.

DAETWYLER, H. D; CALUS, M. P. L.; PONG-WONG, R.;.CAMPOS, G.; HICKEY, J. M. Genomic Prediction in Animals and Plants: Simulation of Data, Validation, Reporting, and Benchmarking. **Genetics**, v. 193, p.347–365, 2013.

DIAS, L. A. S. Análises multidimensionais. In: Alfenas, A. C. (Ed.) **Eletroforese de isoenzimas e proteínas afins**: fundamentos e aplicações em plantas e microorganismos. Viçosa: UFV, 1998, cap.9, p.405-475.

EBERHART, S. A.; RUSSELL, W. A. Stability parameters for comparing varieties. **Crop Science**, v.6, p.36-40, 1966.

EUCLYDES, R. F.; GUIMARÃES, S. E. F. Associação dos métodos tradicionais de seleção à seleção assistida por marcadores moleculares.I Fórum Nacional de Equideocultura In: REVISTA BRASILEIRA DE REPRODUÇÃO ANIMAL, 21, 1997, BELO HORIZONTE, MINAS GERAIS. **Fórum...** Belo Horizonte, MG, 1997, p. 89-96.

EXCOFFIER, L.; LAVAL,G.; SCHNEIDER, S. Arlequin ver. 3.0: an integrated software package for population genetics data analysis. **Evolutionary Bioinformatics Online.** v.1, p.47-50, 2005.

FALCONER, D.; MACKAY, T. **Introduction to Quantitative Genetics.** Longman Technical. 1996.

FAUX, A-M.; GORJANC,G.; GAYNOR, R. C.; BATTAGIN, M.; EDWARDS, S. M.; WILSON, D. L.; HEARNE, S. J.; GONEN, S.; HICKEY,J. M. AlphaSim: Software for Breeding Program Simulation. **The Plant Genome**, v. 9, n. 3 ,2016.

FERNEDA, E. Redes neurais e sua aplicação em sistemas de recuperação de informação. **Ci. Inf.**, Brasília, v. 35, n. 1, p. 25-30, jan./abr. 2006.

FERREIRA, F. M. **Diversidade em populações simuladas com base em locos multiálelicos**. Tese de Doutorado em Genética e Melhoramento. Universidade Federal de Viçosa, Viçosa, 2007.

FIORIN D, V.; MARTINS, F. R.; SCHUCH, N. J.; PEREIRA, E. B. Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. **Revista Brasileira de Ensino de Física**, v. 33, n. 1, 1309 ,2011.

FLORA DO BRASIL 2020 EM CONSTRUÇÃO. **Jardim Botânico do Rio de Janeiro.** Disponível em: < http://floradobrasil.jbrj.gov.br/ >. Acesso em: 31/01/2021

FRANÇOIS, O.; CAYE, K. Naturalgwas: An R package for evaluating genomewide association methods with empirical data. **Molecular Ecology Resources**,v. 18, n.4, p.789–797, 2017.

Food and Agriculture Organization of the United Nations (FAOSTAT) . Disponível em: < http://www.fao.org/faostat/es/#data/QC/visualize>. Acesso em: 31/01/2021.

GAYNOR, C. (2020). **AlphaSimR: Breeding Program Simulations**. R package version 0.12.1. Disponível em: < https://CRAN.R-project.org/package=AlphaSimR >. Acesso em: 31/01/2021.

GEORGES, M.; LEQUARRÉ, A.S.; HANSET, R.; VASSART, G. Genetic variation of the bovine thyroglobulin gene studied at the DNA level. **Animal Genetics**, v.18, n.1, p.41-50, 1987.

GUIMARÃES, W. N. R. Caracterização morfológica e molecular de acessos de feijão-fava (*Phaseolus lunatus* L.). **R. Bras. Eng. Agríc. Ambiental**, v.11, n.1, p.37–45, 2007.

HAYKIN, S. **Neural networks and learning machines**. 3rd ed. New York: Prentice Hall, 2009. 936p.

HOBAN, S. An overview of the utility of population simulation software in molecular ecology. **Molecular Ecology 23** , p.2383-2401. 2014.

HOBAN, S.; BERTORELLE, G.; GAGGIOTTI, O E.Computer simulations: tools for population and evolutionary genetics. **Nature Reviews Genetics,** v 13, p. 110-122. 2012.

HOFFMANN, L. V.; BARROSO, P. A. V**. Marcadores Moleculares como Ferramentas para Estudos de Genética de Plantas.** Campina Grande, 2006. 35p. (Embrapa Algodão. Documentos, 147).

JINKS, J. L.; POONI, H. S. Predicting the properties of recombinant inbreed lines derived by single seed descent. **Heredity**, Oxford, v. 36, n. 2, p. 243-266, 1976.

KINGMAN, J.F.. Origins of the coalescent: 1974–1982. **Genetics 156**, p.1461–1463, 2000.

KOVACS, Z. L. **Redes neurais artificiais**: fundamentos e aplicações. 2.ed. São Paulo: Colledium Cognitio, 1996. 174p.

LACERDA, W. S. **Guia de aulas práticas de redes neurais artificiais**. Lavras: UFLA, 2019. 70 p. : il.

LIMA, R. S. **Descritores morfoagronômicos e divergência genética entre genótipos de tabaco da variedade Bahia no Recôncavo Baiano**. Dissertação (Mestrado). Universidade Federal do Recôncavo da Bahia, Cruz das Almas, 2016, 73p.

MESQUITA, A. S.; OLIVEIRA, J. M. C. A cultura do fumo na Bahia da excelência à decadência. **Bahia Agríc.,** v.6, n.1, nov .2003.

MEYER, H.V; BIRNEY, E.. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. **Bioinformatics**, v.34, n.17, p. 2951-2956, 2018.

MOURA, W. M. et al. Divergência genética em linhagens de pimentão em relação a eficiência nutricional de fósforo. **Pesquisa Agropecuária Brasileira**, Brasília, v. 34, n. 2, p. 217-224, 1999.

MUIR, W.M. Genetic selection strategies: computer modeling. **Poultry Science**,76: 1066-1070. 1997.

NASCIMENTO, M.; PETERNELLI, L. A.; CRUZ, C. D.; NASCIMENTO, A. C. C.; FERREIRA, R. de P.; BHERING, L. L.; SALGADO, C. C. Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes. **Crop Breeding and Applied Biotechnology.** v.13, p.152-156, 2013.

OLIVEIRA, J. M. C. A cultura do fumo na Bahia: refletindo sobre a Convenção-Quadro. **Bahia Agríc**., v.7, n.2, abr. 2006.

PATERSON, A.H.; TANKSLEY, S.D.; SORRELLS, M. E. DNA markers in plant improvement. **Advances in Agronom**y, San Diego, v.46, p.39-90, 1991.

PEREIRA, L. D.; SILVA, D. F. P.; SOUZA, L. K. F.; PEREIRA, E. T. L.; ASSUNÇÃO, H. F.; COSTA, M. M. Genetic diversity of bushy cashew (*Anacardium humile* A. St.-Hil.) based on characteristics of fruits. **Rev. Bras. Frutic.**, Jaboticabal, v. 41, n. 5, 2019.

PERONI, N; MARTINS, P. S.; ANDO, A. Diversidade inter- e intra-específica e uso de análise multivariada para morfologia da mandioca (Manihot esculenta Crantz): um estudo de caso. **Sci. agric**., Piracicaba, v.56, n.3, 1999.

PENG, B.; AMOS, C. I.; KIMMEl, M. Forward-Time Simulations of Human Populations with Complex Diseases. **PLoS Genetics**. v.3, n.3, 2007.

PESSANHA, P. G. O.; VIANA, A. P.; AMARAL JÚNIOR, A. T.; SOUZA, R. M.; TEIXEIRA, M. C.; PEREIRA, M. G. Avaliação da diversidade genética em acessos de *Psidum* spp. via marcadores RAPD. **Rev. Bras. Frutic.**, Jaboticabal, v. 33, n. 1, p. 129-136, 2011.

PIMENTEL, A.J.B.; SOUZA, M. A.; CARNEIRO, P. C. S.; ROCHA, J. R. A.S. C.; MACHADO, J. C.; RIBEIRO, G. Análise dialélica parcial em gerações avançadas para seleção de populações segregantes de trigo. **Pesq. agropec. bras.**, Brasília, v.48, n.12, p.1555-1561, dez. 2013a.

PIMENTEL, A.J.B.; RIBEIRO, G.; SOUZA, M. A.; MOURA, L. M.; ASSIS, J. C.; MACHADO, J. C. Comparação de métodos de seleção de genitores e populações segregantes aplicados ao melhoramento de trigo. **Bragantia**, Campinas, v. 72, n. 2, p.113-121, 2013b.

R Core Team (2019). R: **A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. Disponivel em: https://www.R-project.org/.

RAYMOND, M.; ROUSSET, F. GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. **Journal of Heredity**, v.86, p.248-249, 1995.

RAMALHO, M. A. P.; ABREU, A. F. B.; SANTOS, J. B. Melhoramento de espécies autógamas. In: Nass, L. et al. (Eds.) **Recursos genéticos e melhoramento - plantas**. Rondonópolis: Fundação MT, 2001. cap. 9, p. 201-230.

ROBINSON, I. P. Aloenzimas na Genética de Populações de Plantas. In: Alfenas, A. C. (Ed.) **Eletroforese de isoenzimas e proteínas afins**: fundamentos e aplicações em plantas e microrganismos. Viçosa: UFV, 1998, cap. 7, p. 329-380.

ROCHA,G. S.; CARNEIRO, J. E. S.; CARNEIRO, P. C. S.; POERSCH, N. L.; LIMA, M. S.; SILVA, L. C. Estratégias de predição e efeitos de ambientes na avaliação de populações segregantes de feijão. **Rev. Ceres**, Viçosa, v. 62, n.5, p. 438-445, 2015.

ROSENBLATT, F. The perceptron: A probabilistic model for information storage and organization in the brain. **Psychological Review**,v. 65, n.6, p.386–408, 1958.

SANTOS P.G.; SOARES, A. A.; RAMALHO, M. A. P. Predição do potencial genético de populações segregantes de arroz de terras altas. **Pesq. agropec. bras.**, Brasília, v. 36, n. 4, p. 659-670, abr. 2001.

SEBIM, D. E.; OLIVEIRA, P. H.; BRUSAMARELLO, A. P.; BARETTA, D. R. ; BRUM, B. Diversidade genética entre populações de feijão crioulo através da análise multivariada de caracteres morfoagronômicos. **Espacios,** v.37, n.16, p.19, 2016.

SEMAGN, K. et al. An overview of molecular markers methods for plants. **African Journal of Biotechnology**. v.5, n.25, p.2540-68, 2006.

SILVA, E. P.; RUSSO, .C. A. M. Techniques and statistical data analysis in molecular population genetics. **Hydrobiologia**, n. 420, p. 119-135, 2000.

SILVA, B. Z. **Filtragem robusta de SNPs utilizando redes neurais em DNA genômico completo**. Dissertação (Mestrado). Programa de Modelagem Computacional . Universidade Federal de Juiz de Fora, Juiz de Fora, 2013, 100p.

SOARES, E.L.C.; VIGNOLI-SILVA, M.; VENDRUSCOLO, G. S.; THODE, V. A.; SILVA, J. G.; MENTZ, L. A. Família Solanaceae no Parque Estadual de Itapuã, Viamão, Rio Grande do Sul, Brasil. **Revista Brasileira de Biociências**, Porto Alegre, v. 6, n. 3, p. 177-188, jul./set., 2008.

SOUZA, D. C. L. Técnicas moleculares para caracterização e conservação de plantas medicinais e aromáticas: uma revisão. **Revista brasileira de plantas medicinais**, v. 17, n. 3, p. 495-503, 2015.

SOUZA, F.F.; DIAS, R. C. S.; QUEIRÓZ, M. A. Capacidade de combinação de linhagens avançadas e cultivares comerciais de melancia. **Horticultura Brasileira.** v. 31, n. 4, 2013.

STRAND, A. E. Metasim 1.0: an individual-based environment for simulating population genetics of complex population dynamics. **Molecular Ecology Notes** , v.2, p.373–376, 2002.

SUN, X. **Models and methods for computer simulations as a resource in plant breeding.** Dissertation of Doctor of Philosophy in Crop Sciences. University of Illinois at Urbana-Champaign, 2012

TAUTZ, D. Hypervariabilitiy of simple sequences as a general source for polymorphic DNA markers. **Nucleic Acids Res.**, v. 17, p. 6463-6471, 1989.

TEODORO, P. E.; BARROSO, L. M. A.; NASCIMENTO, M.; TORRES, F. E.;SAGRILO, E.; SANTOS, A.; RIBEIRO, L. P. Redes neurais artificiais para identificar genótipos de feijão-caupi semiprostrado com alta adaptabilidade e estabilidade fenotípicas. **Pesq. agropec. bras.**, Brasília, v.50, n.11, p.1054-1060, 2015.

TURCHETTO-ZOLET, A. C.; TURCHETTO, C.; ZANELLA, C. M.; PASSAIA, G. **Marcadores Moleculares na Era genômica: Metodologias e Aplicações**. Ribeirão Preto: Sociedade Brasileira de Genética, 2017. 181 p.

VALÉRIO, I.P.; CARVALHO, F.I.F.; OLIVEIRA, A.C.; SOUZA, V.Q.; BENIN, G.; SCHMIDT, D.A.M.; RIBEIRO, G.; NORNBERG, R.; LUCH, H. Combining ability of wheat genotypes in two models of diallel analyses. **Crop Breeding and Applied Biotechnology**, v.9, p.100-107, 2009.

VOS, P.; HOGERS, R.; BLEEKER, M.; REIJANS, M.; VAN DE LEE, T.; HORNES, M.; FRIJTERS, A.; POT, J.; PELEMAN, J.; KUIPER, M. & ZABEAU, M. AFLP: a new technique for DNA fingerprinting. **Nucleic Acids Research**, Oxford, v.23, n.21, p.4407-4414, 1995.

WELSH, J.; McCLELLAND, M. Fingerprinting genomes using PCR with arbitrary primers. **Nucleic Acids Research**, Oxford, v.18, n.24, p.7213-7218, 1990.

WILLIAMS, J.G.K.; KUBELIK, A.R.; LIVAK, K.J.; RAFALSKI, J.A. & TINGEY, S.V. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. **Nucleic Acids Research**, Oxford, v.18, n.22, p.6531-6535, 1990.

WIMMER, V.; ALBRECHT, T.; AUINGER, H.J.; SCHOEN, C.C. synbreed: a framework for the analysis of genomic prediction data using R. **Bioinformatics**, v.28, n.15 , p.2086-2087, 2012.

WIDROW, B.; HOFF, M.E. in: *Proceedings of IRE WESCON Convention Record*, **Institute of Radio En- gineers**, Los Angeles, v. 4, p. 96-104, 1960.

YUAN, X.; MILLER, D. J.; ZHANG, J.; HERRINGTON, D.; WANG, Y. An Overview of Population Genetic Data Simulation. **Journal of Computational Biology**, v.19, n.1, p. 42–54, 2012.

Capítulo I

Diversidade genética em populações simuladas multi-loci avançadas no tempo

Diversidade genética em populações simuladas multi-loci avançadas no tempo

Autor: Luciana Lima dos Reis
Orientador: Prof. Dr. Ricardo Franco Cunha Moreira
Co-orientador: Prof. Dr. Jair Wyzykowski

Resumo: Entender os cenários evolutivos a partir das populações que habitam determinados locais geográficos é importante para conhecer a estrutura genética populacional, uma vez que a utilização de ferramentas tecnológicas avançadas tornou-se saída viável para acompanhar o comportamento das populações em seus habitats naturais e os possíveis acontecimentos genéticos que produzem diversidade entre e dentre as populações. Objetivou-se no presente trabalho a verificação dos modelos biométricos para a simulação de dados, o desempenho dos mesmos na produção de diversidade populacional e avaliação da diversidade genética nas populações simuladas. Para isto, foram simuladas quatro populações de plantas dióicas, com quatro demes distintas em quatro fases de vida e distribuição no tempo de 40 anos. Os dados resultaram em quatro populações naturais, sendo pop 1= 12, pop 2= 40, pop 3= 25 e pop 4= 29 indivíduos com seis locus polimórficos, com 55 alelos cada. Os valores para o número total de alelos na população (A) variou entre 53 e 55 e a proporção de alelos na população (P_a) apresentou média de 50% por população. As frequências alélicas não ultrapassaram valores de 0,4 em todas as amostra, com média de proporção de loci polimórficos (P) igual a 0,71. Para número médio de alelo por loco (N_m), número médio de alelos efetivos por loco (n_e) e índice de Shannon-Wiener (H), o locus L3 (N_m= 1,42, n_e= 7,25, H= 2,6) e locus L4 (N_m=1,83 ,n_e= 8,78, H=2,27) assumiram os valores acima da média observadas. Na heterozigosidade esperada (He) e observada (Ho), a média foi He=0,85 e Ho= 0,86. O índice de fixação dentro de populações (F_{is}), índice de fixação total (F_{it}) e divergência entre as populações (F_{st}) apresentam valores médios de F_{is}=-0.003, F_{it}= -0,004 e F_{st}=-0,001. De acordo com a AMOVA a diversidade existente dentre as populações é de 100%, sugerindo que os dados simulados apresentaram alta diversidade genética dentre os indivíduos da população, podendo este ser utilizado para predição genética para conservação.

Palavras-chaves: Estrutura genética, marcadores moleculares, recursos genéticos naturais.

Introdução

A biodiversidade existente nas florestas, consiste numa fonte de estudo para as mais diversas áreas do conhecimento vegetal. Por apresentarem grandes extensões, estas possuem uma complexa estrutura e dinâmica de desenvolvimento genético. No entanto, apesar das florestas apresentarem características tão importantes para a manutenção das diversas fontes de vida, estas ainda são alvos de degradação através das ações antrópicas, que resultam na fragmentação de florestas. Neste contexto, o estudo do fragmento florestal é de grande importância para obter diversas informações sobre as populações naturais existentes, compreensão dos processos evolutivos e das novas condições de adaptação decorrentes dos impactos sofridos pelo ambiente ao longo do tempo.

As populações naturais podem estar sujeitas à mudanças (ESTIGARRIBIA et al., 2017) capazes de ocasionar uma desestruturação a nível de separação espacial entre seus integrantes, podendo aumentar sua autofecundação e as taxas de cruzamentos entre os indivíduos aparentados, afetando assim o fluxo gênico e ocasionando a deriva genética (DAL BEM et al., 2015; ZANELLA, 2011). Desta forma, é crucial o desenvolvimento de estudos para avaliação dos efeitos genotípicos e fenotípicos nestas populações, a fim de compreender as mudanças ocorridas no tempo e espaço (MARTINS, 1987; ESTOPA et al., 2006).

Para entender os eventos evolutivos sofridos pelas populações, os pesquisadores têm recorrido a simulação de dados, com a finalidade de verificar possíveis cenários do futuro e seus efeitos nas mesmas. A simulação de dados sobre populações para predição genética tem ganhado espaço em diversos trabalhos científicos; por gerar informações robustas e de alta confiabilidade e ser uma técnica de baixo custo para avaliação temporal dos indivíduos que as constituem (OLIVEIRA et al., 2005; AGUIAR, 2006; PEIXOTO, 2013; GUIMARÃES, 2016).

Com o avanço tecnológico nos softwares, a simulação tem apresentado vários modelos, como: aqueles baseados em números de genótipos, números de indivíduos e genotipagem de alto rendimento (STRAND, 2002; ANDRELLO & MANEL, 2015; WIMMER et al., 2012). Sendo capazes de simular de forma

eficiente uma grande variação de eventos ocorridos no tempo e espaço (WIMMER et al., 2012).

Para verificação da acurácia dos dados, a análise da variabilidade populacional torna-se importante entre e dentre populações. Através desta, é viável determinar a direção e magnitude da alteração gerada nas frequências alélicas dos indivíduos que compõe as populações e através destas é possível direcionar estratégias para conservação dos recursos genéticos (MARTINS, 1987).

Desta forma, o presente trabalho tem como objetivo a simulação de quatro populações multi-locus com a atuação de eventos evolutivos durante 40 anos e analisar a estrutura genética populacional por meio da previsão temporal.

Material e métodos

1. Simulação dos dados

O trabalho foi desenvolvido no software R versão 3.6.3, com a simulação dos dados e análise dos parâmetros populacionais, a fim de verificar a diversidade genética atuante nas populações e seus processos evolutivos.

As simulações foram baseadas em modelos estocásticos, que considera a relação dos indivíduos com a demografia, ambiente e relação entre os mesmos. Para obtenção dos dados, foi utilizado o pacote rmetasim com a simulação de um marcador microssatélites para quatro populações endêmicas de plantas, com ciclo de vida dioico e distribuição inicial em dois estágios de vida (de acordo com as figuras 1a, 1b e 1c). Estes passaram por 30 gerações, onde cada ano avançado, as populações sofreram efeitos da extinção local, reprodução, sobrevivência, crescimento e redução da capacidade de suporte de 400 indivíduos nas quatro populações estimadas. Na reprodução, houve dispersão e movimento de polén entre as subpopulações.

Foi definido taxa de variação em torno de 0,01, para mutação. Onde cada locus sofreu mutação única, sendo esta mantida ao longo das simulações. Os locus foram gerados com base no modelo de mutação gradual para geração de dados em sequencias de microssatélites com comprimentos variáveis.

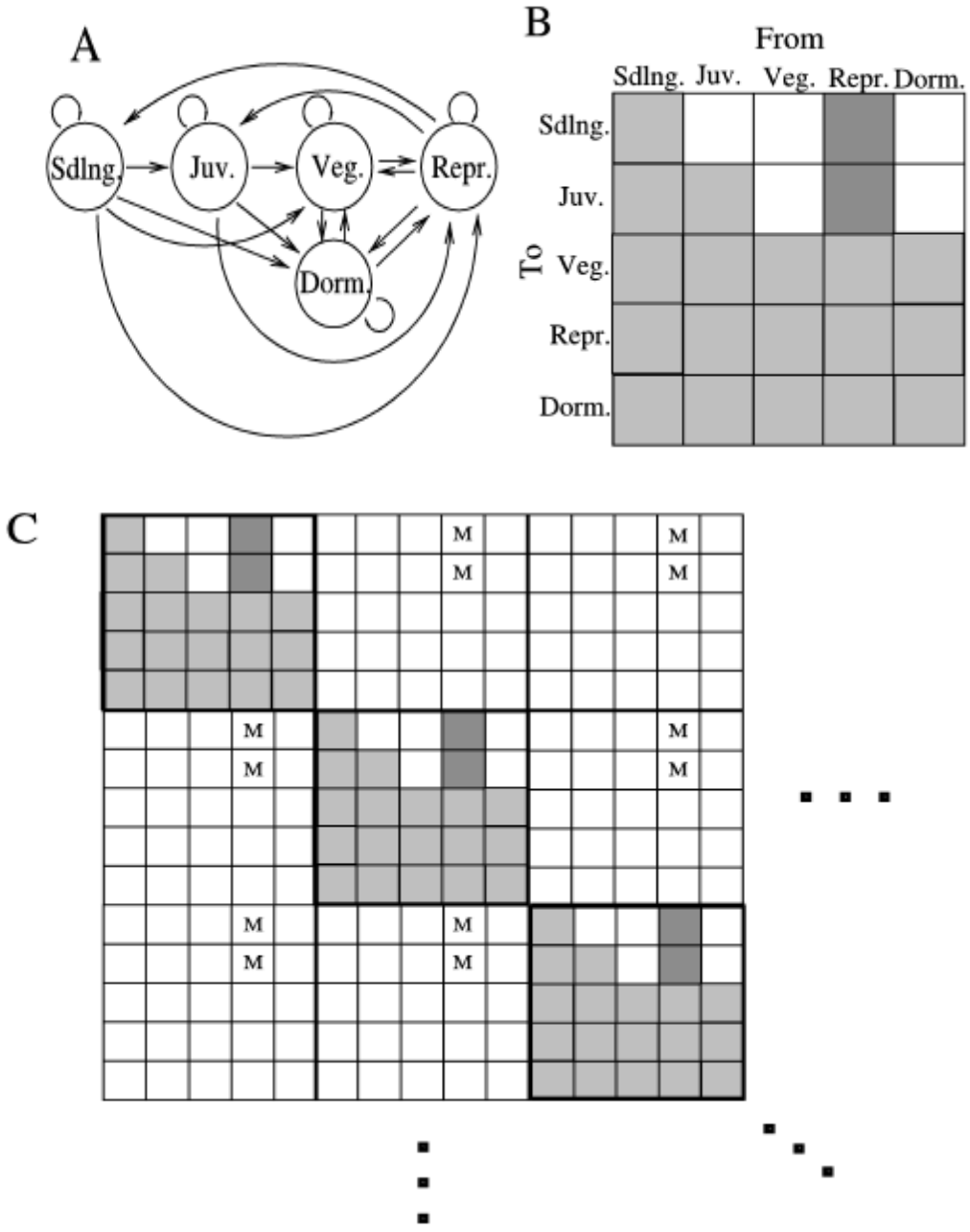


Figura 1: Esquema de simulação do programa rmetasim para uma população, como exemplo evolutivo dos estágios de vida, para os habitats e paisagem. a) Ciclo de vida para os diferentes estágios de desenvolvimento de indivíduos, variando entre mudas (*Sdlng*), juvenis (*Juv.*) , adultos vegetativos (*Veg.*), adultos reprodutivos (*Repr.*) e indivíduos dormentes (*Dorm.*); b) Matriz de trasição entre os estágios de vida dos indivíduos; c)Matriz de relação entre crescimento x sobrevivência x reprodução , de acordo com modelo populacional descrito por Caswell (1989). (Fonte: Strand, 2002).

Obtendo-se ao final, quatro populações com número total de 106 indivíduos, subdivididos entre as mesmas, com 6 locus e 55 alelos presentes na população e tempo de simulação estimado em 40 anos. Estes parâmetros foram definidos de acordo como modelo desenvolvido por Strand (2002) no pacote Rmetasim.

2. Análise da estrutura da população

As análises das populações foram executadas através de parâmetros genéticos para identificação da diversidade intrapopulacional e interpopulacional, obtidas através dos pacotes Rmetasim, Poppr, Adegenet, factoextra, NbClust, cluster, stats, hierfstat e StrataG.

a) Heterozigosidade esperada (He)

A heterozigose pode ser quantificada, com base em Nei (1978):

$$He = 1 - \sum P_i^2$$

Onde:

P_i- frequência estimada do *i*-ésimo alelo.

b) Heterozigosidade observada (Ho)

A quantidade de heterozigose de um determinado loco, pode ser determinada de acordo com Brown & Weir (1983):

$$Ho = 1 - \sum P_{ii}$$

Onde:

P_{ii} - frequência observada de genótipos homozigotos do alelo *i*.

c) Índice Shannon-Wiener

O índice Shannon-Wiener (1949) calcula a diversidade genotípica ou fenotípica da população i no loco c. Sendo definida como:

$$H' = -\sum_{i=1}^{c} Pi \ln(Pi)$$

Onde:

c - número de classes genotípicas para um dado loco;

Pi - frequência do i-ésimo genótipo.

d) Número médio de alelos por loco

É a quantificação de alelos diferentes no loco j, que varia em um intervalo de 1< k <a_j.

e) Número total de alelos (A)

É definida pela soma dos alelos L dos locus analisados dentro da população i (FERREIRA, 2007; CRUZ et al.,2020).

$$A = \sum_{j=1}^{L} a_j$$

f) Número médio de alelos por loco (N_m)

É calculada através da seguinte expressão (CRUZ et al., 2020):

$$N_m = \frac{A}{L}$$

Onde:

A - número total de alelos

L - número de locos analisados.

g) Número efetivo de alelos por loco (n_e)

De acordo com Nei (1987), o número de efetivo de alelos pode ser dado pela expressão**:**

$$n_e = \frac{1}{1 - \widehat{H}_e}$$

h) Proporção de alelos na população (P_a)

Para Cruz et al.(2020), este pode ser calculado através da expressão :

$$P_a = \frac{número\ de\ alelos\ da\ população}{número\ total\ de\ alelos\ da\ espécie}$$

i) Proporção de locos polimórficos *(P)*

Pode ser calculada através da expressão:

$$P = \frac{número\ de\ locos\ polimórficos}{número\ total\ de\ locos\ analisados}$$

Segundo Cole (2003), a proporção de loci polimórficos pode ser classificada de acordo com três critérios:

i) Loco exibindo polimorfismo em pelo menos um indivíduo da amostra;

ii) Loco em que o alelo mais comum tem frequência menor que 99%;

iii) Loco em que o alelo mais comum tem frequência menor que 95%.

j) Índice de fixação (Coeficiente de endogamia) *(f)*

O índice de fixação foi formulado conforme Wright (1965), como uma medida de diversidade para as populações, podendo ser expressa através da equação:

$$\hat{F}_{Is} = 1 - \frac{\hat{H}_o}{\hat{H}_e}$$

Sendo:

F_{IS}- mede o coeficiente de ancestria, como uma medida da correlação de gametas entre as subpopulações.

k) Distância de Nei

Nei et al. (1983), para verificação da distância genética, propôs a expressão:

$$D_{N83,ii'} = \frac{1}{L}\sum_{j=1}^{L}(1 - \sum_{k=1}^{a_j}\sqrt{\hat{p}_{ijk}\hat{p}_{i'jk}}$$

l) Estatística H de Nei

Nei (1973,1977), para obter os estimadores de variabilidade dos retrocruzamentos, os parâmetros calculados foram:

$$\hat{F}_{IS} = \frac{\hat{h}_s - \hat{h}_o}{\hat{h}_s}$$

$$\hat{F}_{IT} = \frac{\hat{h}_T - \hat{h}_o}{\hat{h}_T}$$

$$\hat{F}_{ST} = \hat{G}_{ST} = \frac{\hat{h}_T - \hat{h}_S}{\hat{h}_T}$$

Sendo:

$\hat{F}_{IS}$ - índice de fixação dentro de populações

$\hat{F}_{IT}$ - índice de fixação total

$\hat{F}_{ST}$- divergência entre as populações

I – Individuos

S – populações

T – Total de populações

m) AMOVA

Segundo Excoffier et al. (1992), a análise de variância molecular estima a variância dos diferentes níveis hierárquicos, verificando como está distribuída a variabilidade das populações (VASCONCELOS, 2006). Esta é dada pelos de forma resumida na tabela 1.

Tabela1: Analise de variância molecular (AMOVA), de acordo com Excoffier et al. (1992)

Fonte de variação	GL	SQ	E(QM)
Entre Populações	g-1	SQD_{EP}	$n\sigma_a^2 + 2\sigma_b^2 + \sigma_c^2$
Entre Indivíduos/Populações	N-g	$SQD_{EI/DP}$	$2\sigma_b^2 + \sigma_c^2$
Dentro de Indivíduos	N	SQD_{DI}	σ_c^2
Total	2N-1	SQD_T	σ_T^2

Fonte: Ferreira, 2007

n) Visualização gráfica dos resultados

Dendograma

Os dendrogramas foram obtidos por meio das médias aritméticas (não ponderadas) das medidas de distância, dadas pelo método UPGMA (*Unweighted Pair Group Method with Arithmetic Mean*), com definição dos números ótimo de grupos, de acordo com a similaridade das populações (SNEATH & SOKAL, 1973).

Resultados e Discussão

De acordo com as informações contidas na Tabela 2, as simulações resultaram em quatro populações, sendo as pop 1= 12 indivíduos, pop 2= 40 indivíduos, pop 3= 25 indivíduos e pop 4= 29 indivíduos. Sendo as variações de número de indivíduos explicadas pela carga de suporte computacional e processo de extinção simulado nas populações.

A geração estocástica de números de indivíduos pelo suporte computacional e extinção ocorrente a cada geração pela estocasticidade ambiental (STRAND, 2002). Para os parâmetros propostos, iniciou-se as populações com 1000 indivíduos, os quais foram reduzidos para 400 nos consecutivos processos evolutivos, em função do processo de extinção devido a estocasticidade ambiental, de acordo com o suporte computacional do usuário. Resultando em perda de indivíduos por morte e redução nas populações.

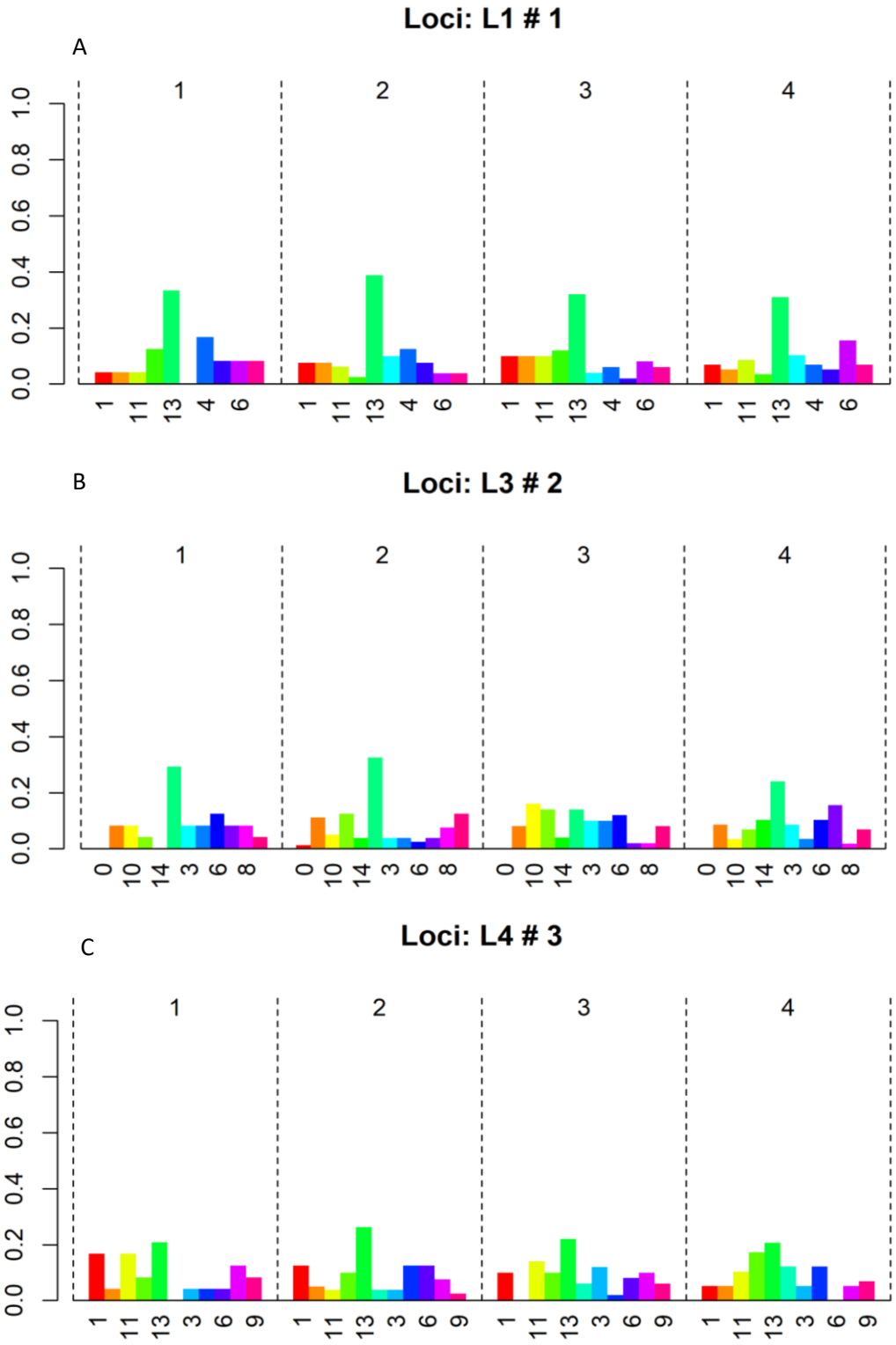
Tabela 2: Número de indivíduos por população (n), número total de alelos na população (A), proporção de alelos na população (P_a) e índice de Shannon-Wiener (H), para 4 populações simuladas.

População	$\boldsymbol{n}$	$\boldsymbol{A}$	$\boldsymbol{P_a}$	$\boldsymbol{H}$
Pop 1	12	50	0,24	2.48
Pop 2	40	55	0,26	3.69
Pop 3	25	53	0,25	3.22
Pop4	29	53	0,25	3.37
Total	106	55	1	-

Os dados simulados resultaram com número total para alelos de 55, distribuídos entre as populações e baixa amplitude na proporção de alelos na população com variação de 0,24 á 026, podendo-se deduzir que as mesmas tiveram condições de geração dos alelos idênticas e que estas não sofreram influência no número de indivíduos. O índice de Shannon-Wiener (H) encontrado para a populações variaram entre 2,48 e 3,69, para a população 1 e 2, respectivamente. Indicando alta diversidade entre as amostras das populações avaliadas.

As frequências alélicas (Figura 2) variaram entre (0 e 0,4) em todos os loci observados, havendo geração de alelos para baixa frequência nas populações verificando desta forma, que as populações não sofrerem efeito de deriva genética. Segundo Faria (2017) o aumento da deriva genética, está ligado com o aumento da taxa de autofecundação. Desta forma, pode-se constar que apesar das populações terem evoluído em muitas gerações e sofrido a influência aleatória do ambiente, os fatores evolutivos gerados

estocasticamente não causaram efeito nas populações. As alterações de frequência alélicas podem ocorrem de acordo com a variação demográfica, gerando deriva genética resultando em isolamento das espécies (GUSSON et al., 2005).



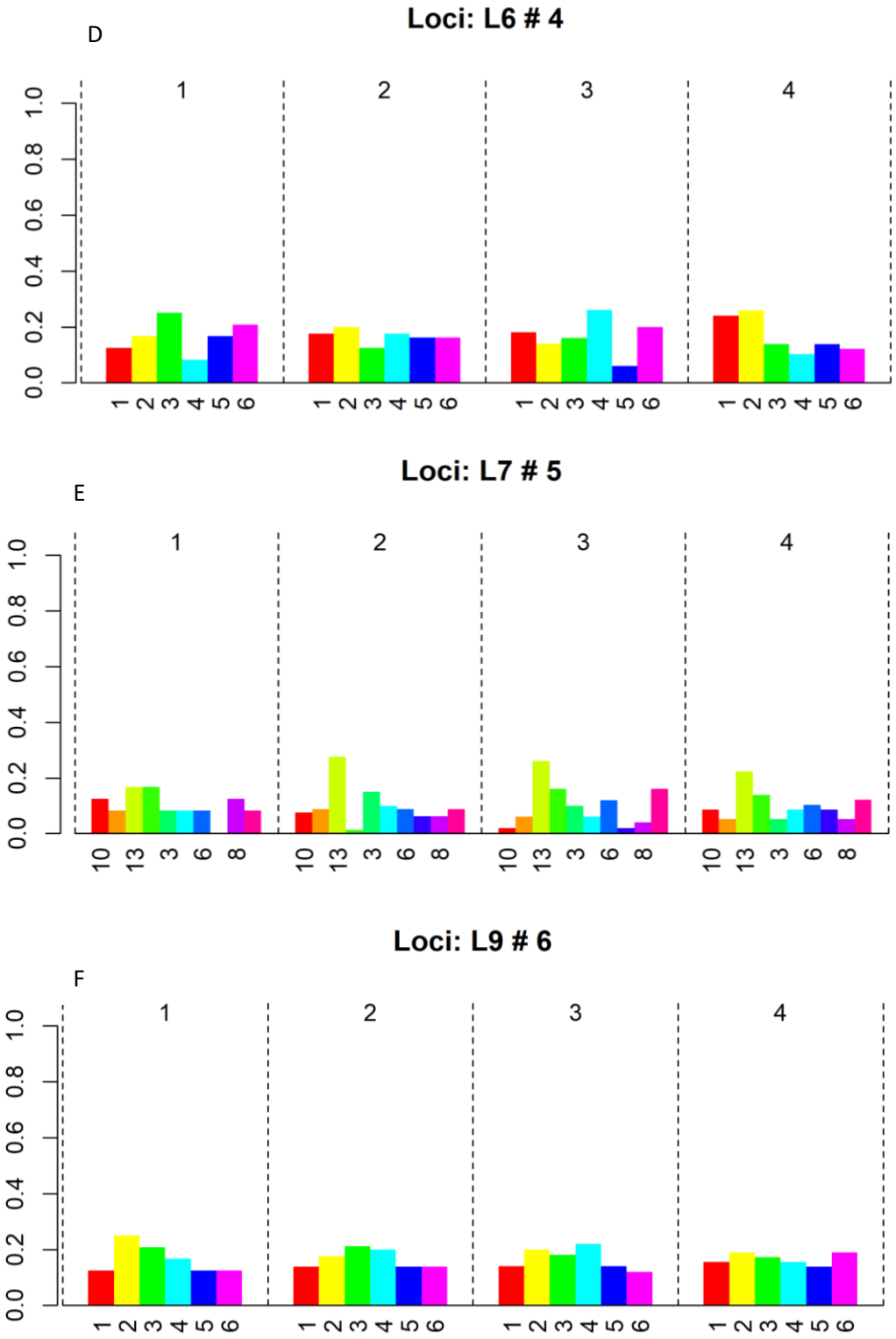


Figura 2: Frequências alélicas por loci para quatro populações simuladas. Em que : Figura 2A: Loco 1; Figura 2B: Loco 3; Figura 2C: Loco 4; Figura 2D: Loco 6; Figura 2E: Loco 7; Figura 32F: Loco 9.

Na Tabela 3, para todos os seis loci analisados, foram observados em número de alelos (Na) a variação de 6 (Loci 6 e 9) a 10 (Loci 1 e 7) alelos por loci com uma média de 9, para o total de 55 alelos analisados. O número médio de alelo

por loco (N_m) apresentou uma variação de 1 a 2 referindo-se aos loci 6,9 e 3, respectivamente.

Em relação a proporção de loci polimórficos (P), foi encontrada para o loci L6 e L9 100% de proporção, e a menor no loci L3 e L4, com 54%. Segundo Cole (2003), a proporção de alelos consideram a existência de polimorfismo. Para Nei e Kumar (2000), fatores evolutivos como mutações e interações genômicas, são causadores de polimorfismo e diversidade de loci. O número médio de alelos efetivos por loco (n_e) variou entre 6,0 para o loco L9 e 8,78 para o loco L4, com uma média de 7,25 para todos os loci avaliados.

Tabela 3: Número de alelos (Na), número médio de alelo por loco (N_m), número médio de alelos efetivos por loco (n_e), proporção de loci polimórficos (P %), Índice de Shannon-Wiener (H), heterozigosidade esperada (He), heterozigosidade observada (Ho) e média, para os seis locos obtidos para as populações.

Loci	$\boldsymbol{Na}$	$\boldsymbol{P}\%$	$\boldsymbol{N_m}$	$\boldsymbol{n_e}$	$\boldsymbol{H}$	$\boldsymbol{He}$	$\boldsymbol{Ho}$
L1	10	60	1,67	6,10	2,07	0,83	0,90
L3	11	54	2,00	8,42	2,26	0,88	0,90
L4	11	54	1,83	8,78	2,27	0,88	0,88
L6	6	100	1,00	6,04	1,78	0,83	0,76
L7	10	60	1,06	8,14	2,19	0,87	0,91
L9	6	100	1,00	6,00	1,78	0,83	0,83
Média	9	71	1,42	7,25	2,06	0,85	0,86

O Índice de Shannon-Wiener (H) (Tabela 3), constatou que os dados a nível de loci possuem uma ampla diversidade entre eles, sendo de 1,78 para os locos L6 e L9 e 2,27 para L4. Indicando que á maior riqueza de genes e consequentemente maior número de alelos no loco L4.

A diversidade dos loci foi mensurada através da heterozigosidade esperada (He) e heterozigosidade observada (Ho), onde os valores da heterozigosidade observada apresentou valores maiores que a heterozigosidade esperada, o que caracterizou uma maior presença de indivíduos heterozigotos. Para Sánchez (2008) a diversidade genética pode ser mensurada através da heterozigosidade, pois esta representa toda variação existente em cada loci nas populações.

Tabela 4: Análise de variância molecular (AMOVA), entre duas populações naturais simuladas sobre avanço no tempo.

FV	GL	SQ	QM	CV	V (%)
Entre populações	3	7,19	2,40	-0,01	0
Dentro de populações	102	261,95	2,57	2,57	100
Total	105	269,15	2,56	2,56	100
Estatistica ØST	-0,003				

Os resultados apresentados na (Tabela 4), demonstraram que as variações presentes dentro e entre as populações são de 100% e 0%, respectivamente. Isto, sugere que os fatores demográficos (distribuição de indivíduos no espaço), genéticos (genoma) e evolutivos (seleção, mutação e reprodução sexuada) foram capazes de gerar distinção apenas dentro destas. Resultado semelhante foi encontrado por Silva et al. (2020) em trabalho com duas populações de indivíduos de Imbé (*Philodendron adamantinum* Mart. ex Schott),onde verificou-se que a diversidade entre as populações foi de 5,68% e dentro 94,32%. Para Zimback et al.(2004), em populações naturais a variação maior dentre as mesmas, podem estar ligadas a maior proporção de indivíduos arbóreos. A detecção da variabilidade dentre e entre populações, pode ser considerado importante para sobrevivência das mesmas, pois explicam de forma resumida como as populações estão estruturadas na demografia (Fonseca et al.,2020).

Blanco et al.(2008) em trabalho para avaliar a diversidade genética em onze populações naturais de araticunzeiro (*Annona crassiflora* Mart.) com o auxilio de seqüências não codificantes do DNA de cloroplastos, verificou nas analises que 7,3% da diversidade foi encontrada entre as populações e 92,7% dentre as mesmas. Concluindo que a fragmentação ambiental influência na diferenciação das populações por dificultarem os eventos migratórios.

No que concerne a estatística H de Nei (Tabela 5), verificou-se que as populações apresentaram, no geral, baixo índice de endogamia (F_{is}) para os loci, variando entre -0,071 para o loco L1 e o mais alto nível de 0,086 para o loco L6, com amplitude de 0,054 entre os valores de máximo e mínimo observados. Para Muniz et al.(2008) o índice de endogamia tem influência sobre a diversidade existente na população, por caracterizarem de forma resumida as relações de frequências genéticas.

O índice de fixação total (F_{it}) apresentou uma média -0,004 indicando baixo grau de fixação deste nas populações sugerindo baixo desvio do equilíbrio de Hardy-Weinberg na população total. De acordo com Faria (2017), a verificação de equilíbrio é importante para considerar hipóteses genéticas, para simulação de dados através dos sistemas de acasalamento.

Os valores obtidos para divergência entre as populações (F_{st}) foram abaixo de -0,003 sugerindo desta forma, que os loci apresentaram baixa diferenciação genética de acordo com os parâmetros de avaliação de Nei (1973,1977). Para Hilsdorf (2013), ao avaliar o Fst em populações, este pode ser indicador de baixa heterozigosidade em subpopulações.

Tabela 5: Estatística H de Nei (1973,1977) para 6 loci encontrados nas populações simuladas. Sendo obtidos valores para índice de fixação dentro de populações (F_{is}), índice de fixação total (F_{it}) e divergência entre as populações (F_{st}).

Loci	F_{it}	F_{st}	F_{is}
L1	-0,073	-0,002	-0,071
L3	-0,014	0,010	-0,024
L4	0,011	0,003	0,008
L6	0,084	-0,002	0,086
L7	-0,033	-0,001	-0,032
L9	-0,001	-0,017	0,016
Média	-0,004	-0,001	-0,003

Para verificação da dissimilaridade existente nas populações, foi exucutado a distância de Nei (tabela 6). Pois, segundo Moraes & Derbyshire (2003) a utilização da distância de Nei, torna-se apropriada para explicar separações dos níveis, que ocorrem substituições gênicas. Com isso, pode-se observar que a menor distância assumiu valor de 0.091, estando relacionada as populações 2 e 4. E a maior distância calculada foi observada para 0,116 entre a população 3 e 4.

Tabela 6: Matriz de distância de Nei et al, (1983), entre 4 populaçoes simuladas.

Populações	1	2	3	4
1	0	0,107	0,110	0,112
2		0	0,111	0,091
3			0	0,116
4				0

Figura 3: Análise de agrupamento (UPGMA) a partir da distância de Nei et al, (1983) para quatro populações naturais.

Na visualização gráfica (Figura 3), foi possível identificar pelo ponto de corte, os dois grupos formados, sendo o grupo 1 composto pela população 3 e o grupo 2 pelas populações 1, 2 e 4, agregando um maior número de subgrupos. O ponto de fusão, que definiu o número de grupos, foi calculado de acordo com os critérios de Mingoti (2005), obtendo-se 0,1115. Demonstrando que as populações mais divergentes são a 3 e 4.

Conclusão

A simulação de dados genéticos populacionais consiste em uma saída viável, para cenários complexos e realista, além de permitir mensurar todos os efeitos para responder as hipóteses levantadas pelo pesquisador.

Novos estudos podem ser realizados, a fim de comparar esses dados estimados com os reais, com o intuito de quantificar a eficiência dos mesmos para subsidiar estratégias na preservação de recursos genéticos vegetais por meio dessa ferramenta de simulação de dados.

Referências

AGUIAR, H. N. **Índice de seleção utilizando dados simulados de tilápias do Nilo.** Dissertação (mestrado)- Universidade Federal de Viçosa, Viçosa, MG, 2006.

ANDRELLO, M.; MANEL, S. MetaPopGen: an r package to simulate population genetics in large size metapopulations. **Mol. Ecol. Resour**., v.15, n.5, p.1153-1162, 2015.

BLANCO, A.J.V.; PEREIRA, M.F.; COELHO, A. S. G.; CHAVES, L. J. Diversidade genética em populações naturais de araticunzeiro (*Annona Crassiflora* Mart.) por meio da análise de seqüências de cpDNA. **Pesq Agropec Trop**. v.37, n.3, p.169-175, 2007.

BROWN, A.D.H.; WEIR, B.S. **Measuring genetic variability in plant population**. In: TANKESLEY, S. D., ORTON, T. J. (Eds.). Isoenzymes in plant genetics and breeding. Part A. Amsterdam: Elsevier Science, 1983, p. 219-239.

COLE, C.T. Genetic variation in rare and common plants. **Annual Reviews Ecology Systems**. v. 34, p. 213-237, 2003.

CROW, J.F. & KIMURA, M. 1970. **An introduction to population genetics theory**. Harper & Row, New York.

CRUZ, C.D.; FERREIRA, F. M.; PESSONI, L. A. **Biometria aplicada ao estudo da diversidade genética**. Viçosa : UFV, 2020, 614p.

DAL BEM, E. A.; BITTENCOURT, J. V. M.; MORAES, M. L.T.; SEBBENN, A. M. Cenários de corte seletivo de árvores na diversidade genética e área basal de populações de *Araucaria angustifolia* com base em modelagem Ecogene, **Scientia Forestalis**, n. 106, p. 453-466, 2015.

ESTIGARRIBIA, F.; APARÍCIO, W. C.; GALVÃO, F. G.; PEREIRA, L. C. B.; GAMA, R. C. Estrutura da vegetação de fragmentos florestais no Campus da Universidade Federal do Amapá – Brasil, **Biota Amazônia,** Macapá, v. 7, n. 3, p. 17-22, 2017.

ESTOPA, R. A.; SOUZA, A. M. S.; MOURA, M.C. O; BOTREL, C. G.; MENDONÇA, E. G.; CARVALHO, D. Diversidade genética em populações naturais de candeia (Eremanthus erythropappus (DC.) MacLeish). **Scientia Florestalis**, Lavras. v..70, p. 97-106, 2006.

EXCOFFIER, L.; SMOUSE, P. E.; QUATTRO, J. M. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. **Genetics**. v. 131, p. 479-491, 1992.

FARIA, G. M. P. **Aplicação da seleção genômica ampla em populações autógamas e alógamas**. Tese (Doutorado). Genética e Melhoramento-Universidade Federal de Viçosa, Viçosa, 2017. 87 p.

FERREIRA, F. M. **Diversidade em populações simuladas com base em locos multiálelicos**. Tese (Doutorado). Genética e Melhoramento - Universidade Federal de Viçosa, Viçosa, 2007.117 p.

FISHER, R. A. The logic of inductive inference. **J. Roy. Stat. Soc.** v.98, p. 39-54, 1935.

FONSECA, V. L.; SANTOS, J. S.; MASCENA, J. R. L.; ROCHA, N. N.C.; OLIVEIRA, C. S. T.; MOREIRA, R. F. C.; FREITAS, M. C.; FONTELES, S. B. A. Análise da diversidade genética do robalo peva (*Centropomus parallelus*) na resex de Canavieiras, Bahia. **Braz. J. Anim. Environ. Res.**, Curitiba, v. 3, n. 3, p. 2180-2197, 2020.

GUIMARÃES, C. T. ; SCHUSTER, I.; MAGALHÃES, J.V.;SOUZA JÚNIOR, C.L. Marcadores moleculares no melhoramento de plantas. In: Borém, A.; Caixeta, E. T. (Eds). **Marcadores moleculares**. Viçosa: UFV, 2006. cap. 4, 107-144.

GUSSON, E.; SEBBENN, M. A.; KAGEYAMA, P. Diversidade e estrutura genética espacial em duas populações de *Eschweilera ovata*. **Scientia Forestalis**, n.67, p. 123- 135, 2005.

HILSDORF, A.W.S**. Marcadores moleculares e a caracterização dos recursos genéticos de peixes: Desenvolvimento sustentável da aquicultura e da pesca de espécies nativas de água doce no Brasil**. 2013. 159 f. Tese (Livre Docência). Pirassununga. 2013.

MARTINS, P. S. Estrutura populacional, fluxo gênico e conservação “in situ”, **IPEF,** n.35, p.71-78, 1987.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Ed. UFMG. 297p. 2005.

MORAES, P. L. R.; DERBYSHIRE, M. T. V. C. Diferenciação genética e diversidade em populações naturais de Cryptocarya aschersoniana Mez (Lauraceae). **Biota Neotropica**, v. 3, n.1 2003.

MUNIZ, J. A.; CAMARGO, M. S.; FERREIRA, D. F.; VEIGA, R. D Métodos de estimação do coeficiente de endogamia em uma população diplóide com alelos múltiplos. **Ciência e Agrotecnologia**, Lavras, v. 32, n. 1, p. 93-102, 2008.

NEI, M. Analysis of gene diversity in subdivided populations. **Proceedings of the National Academy of Sciences of the United States of America**. Washington, v. 70, p. 3321-3323, 1973.

NEI, M. F-statistics and analysis of gene diversity in subdivided populations. **Annual Human Genetics**. v. 41, p.225-233, 1977.

NEI, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. **Genetics**, Ottawa, v. 89, p.438-443, 1978.

NEI, M.; TAJIMA, F.; TATENO, Y. Accuracy of estimated phylogenetic trees from molecular data. **Journal of Molecular Evolution**. v.19, p. 153-170, 1983.

NEI, M. Molecular evolutionary genetics. **Columbia University Press**, New York. 1987.

OLIVEIRA, C.G.; CUNHA, E. E.; CARNEIRO, P. L. S.; EUCLYDES, R. F.; MALHADO, C. H. M. Comparação de métodos de seleção em populações simuladas de frangos de corte. **Pesq. agropec. bras.**, Brasília, v.40, n.10, p.969-974, out. 2005.

SILVA, L. F.; MELO JÙNIOR, A. F.; OLIVEIRA, D. A.; MENEZES, E. V.; ROYO, V. A.; BRANDÂO, M. M. Cerrado rupestre do espinhaço: diversidade genética de espécie endêmica. **Revista Cerrados**, v. 18, n. 01, p. 300-330, 2020.

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

VASCONCELOS, M. E. C.; RODA, E. A. Análise multivariada de dados moleculares de Rizóbio-Phaseolus isolados de nódulos de feijoeiro. **Ciências Exatas e Tecnológicas**, v.5, p. 53-57, 2006

WEIR, B. S. **Genetic Data Analysis II**. Sinauer Associates, Inc., MA, USA, 1996. 445p.

WIMMER, V.; ALBRECHT, T.; AUINGER, H.J.; SCHOEN, C.C. synbreed: a framework for the analysis of genomic prediction data using R. **Bioinformatics**, v.28, n.15 , p.2086-2087, 2012.

WRIGHT, S. The interpretation of population structure by Fstatistics with special regard to system of mating. **Evolution**, Lawrence, v. 19, p.395-342, 1965.

WRIGHT, S. **Variability within and among natural populations**. Chicago: The University of Chicago Press, 1951. 580 p.

WRIGHT, S. **Variability within and among natural populations**. Vol. 4, The University of Chicago Press, Chicago, 1978, 580p.

ZANELLA, Lisiane, **Analise da interferência antrópica na fragmentação da mata atlântica e modelos de simulação da paisagem na microrregião da Serra da Mantiqueira do estado de Minas Gerais**, Dissertação (Mestrado)-Universidade Federal de Lavras, Lavras, 2011.

Capítulo II

Predição de variáveis quantitativas através de Redes Neurais Artificias para genótipos de tabaco *(Nicotiana tabacum* L.)

Predição de variáveis quantitativas através de Redes Neurais Artificias para genótipos de tabaco (*Nicotiana tabacum* L.)

Autor: Luciana Lima dos Reis
Orientador: Prof. Dr. Ricardo Franco Cunha Moreira
Co-orientador: Prof. Dr. Jair Wyzykowski

Resumo: A utilização da predição de dados, por meio de ferramentas capazes de estimar a colheita, tem sido amplamente aplicada nas produções agrícolas. Com vista nisto, a cultura do tabaco (*Nicotiana tabacum* L.) tem buscado métodos não-destrutíveis para a predição de rendimento, verificando assim a estimativa de produtividade dos seus genótipos. O presente estudo teve como objetivo a predição da produtividade em 15 genótipos de tabaco através das Redes Neurais Artificiais (RNAs). Os dados foram obtidos através de mensuração em uma população de tabaco da variedade Bahia, submetidos a análise das RNAs com entrada de dados de genótipos e 16 variáveis quantitativas, para predição da produtividade Foram treinadas 10 mil RNAs, obtendo-se as 5 melhores, as quais apresentaram o número de neurônios na camada oculta variando de 5 a 15. . As funções de ativações foram logística, exponencial, tangente hiperbólica e identidade. As RNAs apresentaram correlação para as etapas de treino, teste e validação acima de 98% e raiz quadrada do erro médio (RMSE) variando de 6,2 a 7,5 % para as redes. Os gráficos de dispersão dos resíduos não apresentaram tendenciosidade para as produtividades estimadas. As RNAs mostrou-se eficiente na aplicação, para predição da produtividade de genótipos de tabaco variedade Bahia.

Palavras-chaves: Inteligência artificial, modelagem, produção**.**

Introdução

Os dados fenotípicos são resultado da combinação do genoma, fatores ambientais atuantes e a resposta adaptativa dos indivíduos e populações (MORALES, 2000). Estes conjuntos de fatores combinados ajudam a verificar a variabilidade populacional e possíveis estratégias de conservação com a ajuda de modelos matemáticos (JANGARELLI, 2014; BRIEUC et al.,2018).

A verificação das variáveis quantitativas nas populações ajuda na caracterização das populações e agregam conhecimento sobre os indivíduos existentes, contribuindo para a caracterização da diversidade e conservação das espécies (LIMA et al., 2011). No entanto, no setor agrícola as variáveis quantitativas são utilizadas para verificar estimativas de produção para diversas culturas (ANDRADE JUNIOR et al., 2006).

Com vista nisto, tem-se utilizado de modelos capazes de predizer as estimativas de colheita e comparação de variedades de uma cultura, com intuito de antecipar a estimativa de produção de culturas (SOARES et al., 2015). A predição das variáveis quantitativas é vantajosa, por quantificar os impactos sofridos pela interações genéticas e edafoclimáticas, as quais estão submetidas a cultura (BOOTE et al.,1996).

Dentre as diversas culturas existentes, pode-se destacar o tabaco (*N. tabacum* L.) variedade Bahia, por possuir uma ampla distribuição dos seus plantios e desenvolvimento econômico no Recôncavo Baiano (MESQUITA & OLIVEIRA, 2003; LIMA, 2016). Devido à importância do tabaco na economia, é que se tem interesse sobre o conhecimento das características de produtividade, pois esta é imprescindível para o abastecimento do mercado consumidor (LIMA, 2016).

Na cultura do tabaco, a produtividade das folhas é uma das variáveis mais importantes na verificação dos genótipos, por estar ligado diretamente à produção de cigarros, charutos e cigarrilhas, sendo que seu conhecimento prévio, ajudará no planejamento da produção e comercialização dos mesmos(LIMA, 2016).

A predição da produtividade por meio de análises estatísticas, pode ser utilizada no intuito de explorar uma ampla gama de possibilidades combinatórias para verificação das características quantitativas e qualitativas relevantes (YOUNG, 2008; FERNANDES et al.,2019). No entanto, a utilização da predição

da produtividade ainda possui muitas dificuldades devidas á fatores complexos, como a genética e condições ambientais (GUIMARÃES, 2019). Com isto, é necessário pesquisar técnicas mais avançadas para realizar a predição da produtividade.

As RNAs são heurísticas que se baseiam na interação biológica de neurônios do cérebro humano para modelagem (HAYKIN, 2009). Por ser um modelo de aprendizagem e obter rapidez das respostas dos dados analisados, as RNAs têm sido classificadas como alternativa viável para a predição de dados, as quais apresentam acurácia elevada e podem ser utilizadas como uma fase de planejamento para cenários futuros com os fenótipos analisados (TEODORO et al., 2015).

Como todos os modelos de análise de dados, as RNAs também possuem vantagens e desvantagens na sua utilização. Diante disto, podem ser relacionadas algumas vantagens na utilização no processo de modelagem como: alto poder de aprendizagem e generalização dos dados, obtenção de qualidade superior nas análises quando comparado ao método de regressão, rápida implementação nos dados, além de operarem de forma paralela e distribuída (HAYKIN, 2001; AMBRÓSIO, 2002).

No entanto as RNAs não possuem só vantagens na sua aplicação. Uma das desvantagens que podem ser destacadas é o desconhecimento da influência das variáveis em relação à resposta obtida através das redes (BARROSO et al., 2013). Outra desvantagem esta ligada ao tempo de treinamento das RNAs, pois a depender do interesse do pesquisador, o tempo de treino devido ao grande número de redes testadas pode ser relativamente longo (AMBRÓSIO, 2002; BARROSO et al., 2013). No entanto, as vantagens podem superar as desvantagens, viabilizando a utilização das RNAs na determinação de variáveis quantitativas.

Diante do exposto, o presente trabalho tem como objetivo modelar a produtividade de genótipos de tabaco variedade Bahia, por meio das RNAs.

Material e Métodos

Os dados foram obtidos com mensuração das variáveis quantitativas (Tabela 1) para 15 genótipos de tabaco variedade Bahia (Tabela 2), oriundos de plantio da

empresa ERMOR TABARAMA TABACOS DO BRASIL LTDA, na cidade de Cruz das Almas-BA.

Tabela 1: Relação das variáveis quantitativas para genótipos de tabaco, variedade Bahia.

Variáveis quantitativas	**Medida**
Produtividade	$kgha^{-1}$
Dias do transplante ao florescimento	dias
Altura total da planta	cm
Comprimento da inflorescência	cm
Número de folhas	un
Diâmetro do caule, zona mediana	cm
Índice cilíndrico	mm
Largura da 3ª folha	cm
Comprimento da 3ª folha	cm
Largura da 5ª folha	cm
Comprimento da 5ª folha	cm
Ângulo de inserção da 7ª folha	grau
Comprimento dos internódios	cm
Comprimento da flor	cm
Diâmetro do tubo da flor	mm
Engrossamento do tubo da flor	mm
Comprimento da corola	cm

Fonte : Lima, 2016.

Tabela 2: Relação dos 15 genótipos de tabaco variedade Bahia, oriundo da empresa ERMOR TABARAMA TABACOS DO BRASIL LTDA.

Código	Acessos	Tipos	Procedência
BA1	TB-MFZS	BAHIA	ERMOR TABARAMA
BA2	TB-DNC	BAHIA	ERMOR TABARAMA
BA3	TB-BB DMF	BAHIA	ERMOR TABARAMA
BA4	TB-RM	BAHIA	ERMOR TABARAMA
BA5	TB-BEM	BAHIA	ERMOR TABARAMA
BA6	TB-GUI	BAHIA	ERMOR TABARAMA
BA7	TB-B49	BAHIA	ERMOR TABARAMA
BA10	TB-MFINA	BAHIA	ERMOR TABARAMA
BA11	TB-AJS	BAHIA	ERMOR TABARAMA
BA12	TB-PSJ	BAHIA	ERMOR TABARAMA
BA13	TB-BB L	BAHIA	ERMOR TABARAMA
BA14	TB-BSP	BAHIA	ERMOR TABARAMA
BA15	TB-LENCOIS	BAHIA	ERMOR TABARAMA
BA16	TB-BB Z P	BAHIA	ERMOR TABARAMA
BA19	TB-R M	BAHIA	ERMOR TABARAMA

Fonte : Lima, 2016.

A predição foi realizada através do ajuste das RNAs (Figura 1), seguindo das etapas de treino, teste e validação, sendo estas ajustadas no programa STATISTICA 10.0 (STATSOFT, 2010). O banco de dados foi dividido para atendimento às 3 etapas, sendo 50% dos dados para treino, 15% para teste e 35 % para validação das RNAs.

No treinamento, utilizou-se como variáveis independentes os genótipos, variável qualitativa e todas as variáveis quantitativas da tabela 1, com exceção da

produtividade.

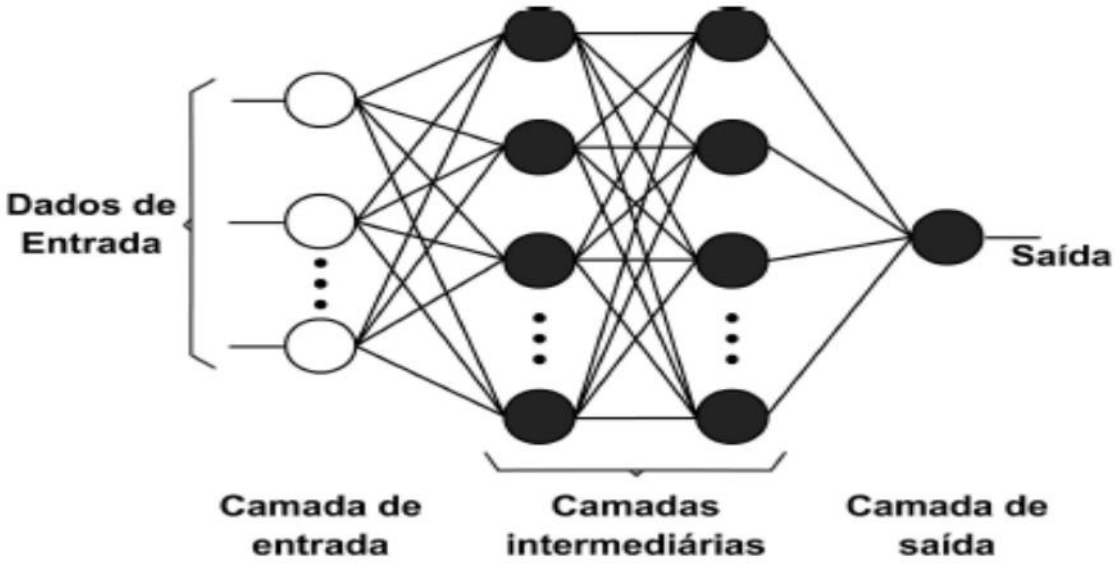


Figura 1: Demonstração simplificada do funcionamento das RNAs (Fonte: Fiorin et al.,2011).

Na camada de entrada, foram destinados 32 neurônios, um para cada variável preditora. A camada oculta foi formada por até metade do número de neurônios da camada de entrada, resultando em 16 neurônios. Foram utilizadas as função logística, exponencial, tangente hiperbólica e identidade como função de ativação, sendo as mesmas implementadas as redes de forma individual a cada arquitetura de RNA.

Foram treinadas 10 mil redes do tipo perceptrons de múltiplas camadas (MLP). Sendo o resultado apresentado através da seleção das cinco melhores RNAs. As RNAs treinadas foram avaliadas por meio da raiz quadrada do erro médio (RMSE %) (MEHTÄTALO et al. 2006). E da correlação (MARTINS et al., 2016), nas etapas de treinamento, teste e validação, através da seguintes equações:

$$RMSE\% = \frac{100}{\bar{P}} \sqrt{\frac{\sum_{i=1}^{n}(P_i - \bar{P}_i)^2}{n}}$$

Em que:

$RMSE\%$: Raiz quadrada do erro médio

$\bar{\bar{P}}$: Média das produtividades totais observados

n : número total de observações

$$r_{y\hat{y}} = \frac{cov(y\hat{y})}{\sqrt{s^2\,(y)s^2(\hat{y})}}$$

Em que:

$r_{y\hat{y}}$: Correlação da produtividade observada e estimada

s^2: variância

cov: covariância

y: produtividade observada

$\hat{y}$: produtividade estimada

Resultados e Discussão

As RNAs treinadas apresentaram arquitetura com 5 a 15 neurônios na camada oculta (Tabela 2). Segundo Braga et al. (2000), o aumento de nêuronios na camada oculta, pode ocasionar a memorização de dados no processo de treinamento, influenciando diretamente no melhor ajuste da variável resposta.

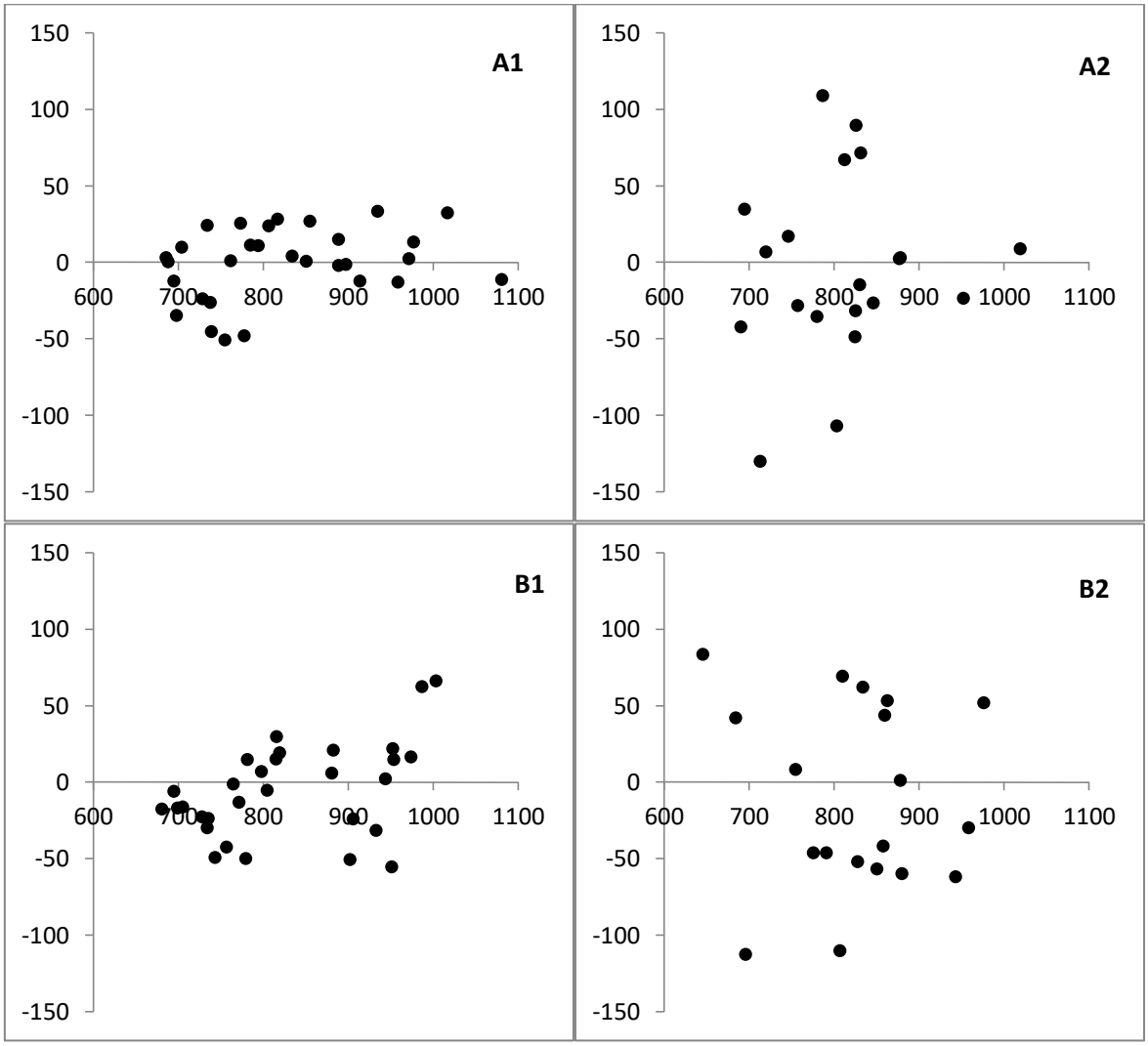
Tabela 2: Redes neurais artificiais (RNAs) selecionadas para estimar o rendimento de genótipos de tabaco variedade Bahia.

		Correlação%			RMSE%		
RNAs	**Arquitetura**	**Treino**	**Teste**	**Validação**	**Treino**	**Teste**	**Validação**
RNA 1	32-5-1	98,10	62,21	83,63	2,823	9,031	7,236
RNA 2	32-15-1	97,10	65,25	82,65	3,745	9,053	7,567
RNA 3	32-9-1	100	69,35	93,59	0,000	9,927	6,205
RNA 4	32-5-1	99,51	73,04	84,07	1,384	7,652	7,141
RNA 5	32-9-1	96,91	61,58	81,98	3,655	9,491	7,746

Legenda: (RMSE %) Raiz Quadrada do erro médio.

De acordo com a correlação e RMSE% (Tabela 2), todas as 5 RNAs selecionadas apresentaram resultados satisfatórios para predição de produtividade em tabaco. No entanto, observou-se que a RNA 3 apresentou os menores valores RMSE% e maiores valores de correlação. Em trabalho para avaliar a produtividade de grãos de milho (Zea mays L.), Soares et al. (2016) observaram que a RNA com menor erro médio de validação e treinamento, apresentou melhores valores para predição do rendimento. Desta forma, o resultado encontrado corrobora com a interpretação apresentado por Braga et al. (2000) e Masters (1993).

O gráfico de dispersão dos resíduos com os rendimentos estimados (Figura 2) (C1) mostra que, embora a RNAs 3 tenha sido exata no treinamento, a validação apresentou tendênciosidade nas estimativas da produtividade.As demais RNA's não apresentaram tendenciosidade dos dados para treino e validação para as demais redes.



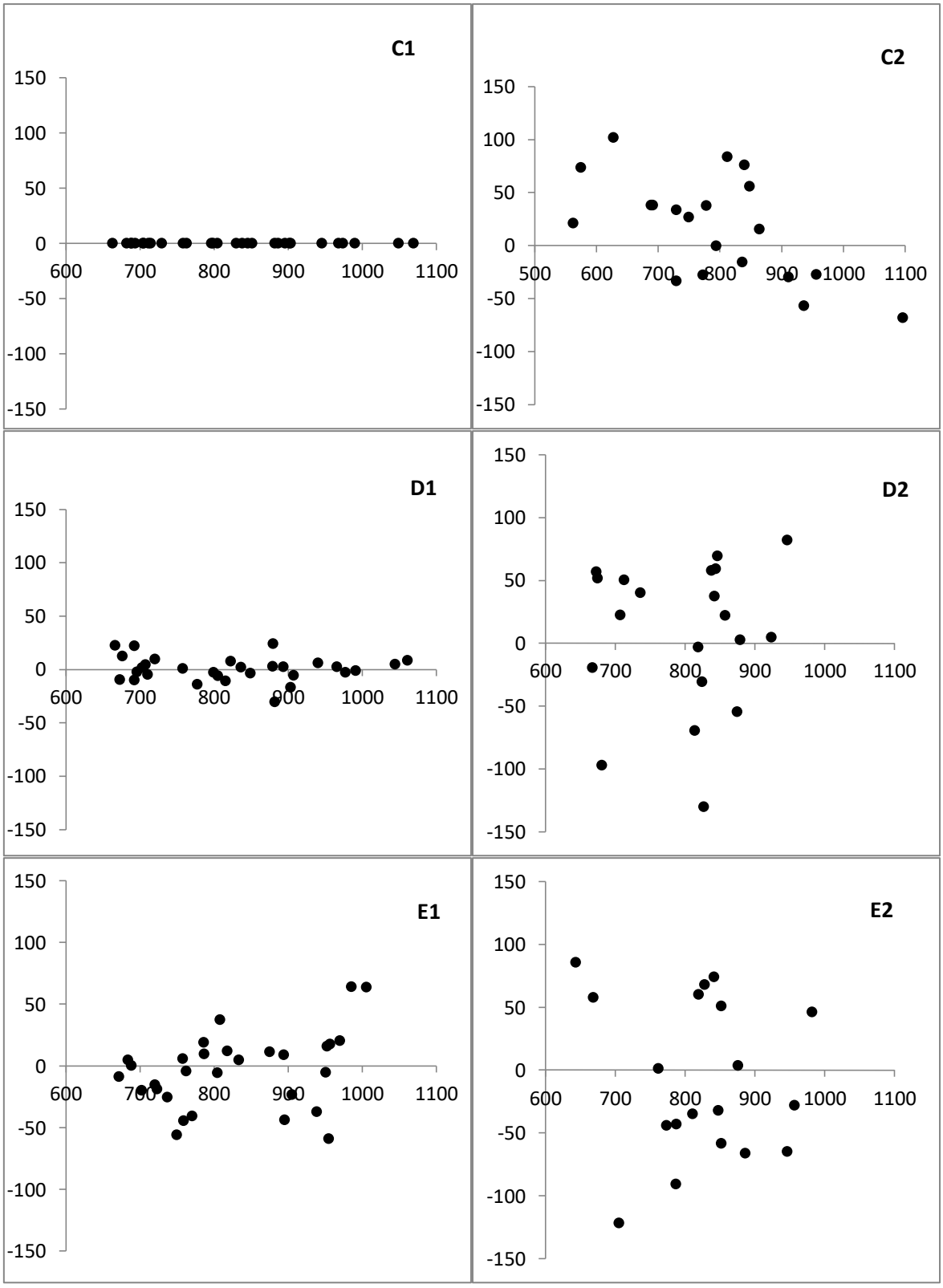


Figura 2 – Rendimento real e estimado, com dispersão gráfica dos resíduos de treino e validação para 5 RNAs. Em que: Figura A1= RNA 1 etapa treino; Figura A2= RNA etapa validação; Figura B1= RNA 2 etapa treino; Figura B2= RNA2 etapa validação; Figura C1= RNA 3 etapa treino; Figura C2= RNA 3 etapa validação; Figura D1= RNA 4 etapa treino; Figura D2= RNA 4 etapa validação; Figura E1= RNA 5 etapa treino; Figura E2= RNA 5 etapa validação.

Os resultados apresentados, nos quais a RNA com exatidão na etapa de treinamento teve viés tendêncioso na validação pode ser entendido como uma ocorrência de overfitting, processo em que a RNAs compreende muito bem os dados de treinamento mas não performa de maneira satisfatória na validação (LEAL et al., 2015;MARTINS et al.,2016). Assim, reforça-se a necessidade de se avaliar o aprendizado das RNAs treinadas com dados não utilizados na etapa de treinamento.

O uso da distribuição de resíduos também se mostrou importante na avaliação de uma RNA. Afinal, na etapa de validação, a RNA com melhores estatisticas de precisão apresentou vies nas estimativas da produtividade.

A utilização de RNAs apresentaram resultados satisfatórios para a predição de rendimento. Assim, pode-se afirmar que as RNAs são ferramentas promissoras para análise das relações existentes entre as variáveis, por não requisitarem os pressupostos estatísticos para análises, como regressão (RECKNAGEL 2001; NUNES & GÖRGENS 2016).

Conclusão

As redes do tipo MLP são eficazes na predição de produtividade de genótipos de tabaco variedade Bahia. Esta mostrou-se uma ferramenta de alta acurácia, que possui potencial para utilização na avaliação de produtividade.

A variação de neurônios na camada oculta interfere diretamente nos resultados na obtenção das redes. Desta forma torna-se importante, em futuros trabalhos, realizar o treinamento das RNAs com arquiteturas distintas, a fim de verificar as possíveis variações estatísticas que podem ser geradas. Podendo ser utilizada posteriormente como modelo para predição de produtividade futura de genótipos de tabaco variedade Bahia.

Referências

ANDRADE JÚNIOR, A. S.; FIGUEREDO JÚNIOR, L. G. M.; CARDOSO, M. J.; RIBEIRO, V. Q. Parametrização de modelos agrometeorológicos para estimativa de produtividade da cultura do milho na região de Parnaíba, Piauí. **Revista Ciência Agronômica**, v.37, p.130-134, 2006.

AMBRÓSIO, P.E. **Redes neurais artificiais no apoio ao diagnóstico diferencial de leões intersticiais pulmonares.** Dissertação de mestrado. Universidade de São Paulo, Ribeirão Preto, 2002, 85p.

BARROSO, L. M. A.; NASCIMENTO, M.; NASCIMENTO, A. C. C.; SILVA, F. F.; FERREIRA, R.P. Uso do método de Eberhart e Russell como informação a priori para aplicação de redes neurais artificiais e análise discriminante visando a classificação de genótipos de alfafa quanto à adaptabilidade e estabilidade fenotípica. **Rev. Bras. Biom.**, São Paulo, v.31, n.2, p.176-188, 2013.

BATTEY, C.J; RALPH, P. L.; KERN, A. D. Predicting Geographic Location from Genetic Variation with Deep Neural Networks. **bioRxiv,** Dezembro, 2019.

BOOTE, K.J. et al. Potential uses and limitations of crop models**. Agronomy Journal**, v.88, p.704-716, 1996.

BRAGA, A.P. et al. **Redes neurais artificiais: teoria e aplicações**. Rio de Janeiro: LTC, 2000. 250 p.

BRASILEIRO, B.P.; MARINHO, C.D.; COSTA, P.M. de A.; CRUZ, C.D.; PETERNELLI, L.A.; BARBOSA, M.H.P. Selection in sugarcane families with artificial neural networks. **Crop Breeding and Applied Biotechnology**, v.15, p.72-78, 2015. DOI: 10.1590/1984-70332015v15n2a14.

BRIEUC, M. S.O.; WATERSA, C. D.; DRINANA, D. P.; NAISHA K. A. A Practical Introduction to Random Forest for Genetic Association Studies in Ecology and Evolution. **Molecular Ecology Resources.** v.18, p.755–766. 2018

BHERING, L.L.; CRUZ, C.D.; PEIXOTO, L. de A.; ROSADO, A.M.; LAVIOLA, B.G.; NASCIMENTO, M. Application of neural networks to predict volume in eucalyptus. **Crop Breeding and Applied Biotechnology**, v.15, p.125-131, 2015. DOI: 10.1590/1984-70332015v15n3a23.

FERNANDES, M. M.; SOUSA, F. L.; SILVA, J. P. M.; ARAÚJO, E. F.; FERNANDES, M. R. M.; NÓBREGA, R. S. A. Redes Neurais Artificiais na estimação de variáveis biométricas de mudas de espécies florestais produzidas em diferentes substratos. **Revista de Ciências Agroveterinárias**. v.18, n. 1, 2019.

FIORIN D, V. et al. Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. **Revista Brasileira de Ensino de Física**, v. 33, n. 1, 1309 ,2011.

GUIMARÃES, E. S. **Aprendizado de Máquina aplicado à predição da produtividade da cultura da soja utilizando dados de clima e solo.** Dissertação (Mestrado) - Universidade de São Paulo, São Carlos, 2019, 75p.

HAYKIN, S. **Redes Neurais: princípios e práticas**. Porto Alegre: ARTMED, 2001. 900 p.

HAYKIN, S. **Neural networks and learning machines**. 3rd ed. New York: Prentice Hall, 2009. 936p.

JANGARELLI, M. Abordagem multivariada para endogamia e valor fenotípico utilizando diferentes estratégias de cruzamento. **Pesq. Agropec. Trop.**, Goiânia, v. 44, n. 1, p. 79-87, 2014.

LEAL, F. A.; MIGUEL, E. P.; MATRICARDI, E. A. T.; PEREIRA, R. S. Redes neurais artificiais na estimativa de volume em um plantio de eucalipto em função de fotografias hemisféricas e número de árvores. **Rev. Bras. Biom.**, v.33, n.2, p.233-249, 2015

LIMA, A. T. B. et al. Molecular characterization of cajá, Spondias mombin (Anacardiaceae), by RAPD markers. **Genetics and Molecular Research,** v.10, n. 4, p. 2893-2904, 2011.

LIMA, R. S. **Descritores morfoagronômicos e divergência genética entre genótipos de tabaco da variedade Bahia no Recôncavo Baiano.** Dissertação (Mestrado) - Universidade Federal do Recôncavo da Bahia, Cruz das Almas, 2016, 73p.

MARTINS, E.R.; BINOTI, M. L. M. S.; LEITE, H. G.; BINOTI, D. H. B.; DUTRA, G. C. Configuração de redes neurais artificiais para estimação do afilamento do fuste de árvores de eucalipto. **Agrária**, v.11, n.1, p.33-38, 2016.

MASTERS, T. **Practical neural network recipes in C + +**. San Diego: Academic, 1993.

MEHTÄTALO, L.; MALTAMO, M.; KANGAS, A. The use of quantile trees in the prediction of the diameter distribution of a stand. **Silva Fennica**, v.40, n.3, p.501-516, 2006.

MORALES, E. Estimating phylogenetic inertia in Tithonia (Asteraceae): a comparative approach. **Evolution**, Lawrence, v. 54, n. 2, p. 475-484, 2000.

NUNES, M. H.; GÖRGENS, E. B. Artificial Intelligence Procedures for Tree Taper Estimation within a Complex Vegetation Mosaic in Brazil. **PloS one** 11: 1-16.3. 493 p. 2016.

RECKNAGEL F. **Applications of machine learning to ecological modelling**. Ecological Modelling 146: 303-310. 2001.

SILVA, B. Z. **Filtragem robusta de SNPs utilizando redes neurais em DNA gnômico completo**. Dissertação de mestrado. Universidade Federal de Juiz de Fora, Juiz de Fora, 2013. 101p.

SOARES, F. C. et al. Predição da produtividade da cultura do milho utilizando rede neural artificial. **Ciência Rural**, v.45, n.11, nov, 2015.

STATSOFT. INC. STATISTICA (data analysis software system). version 10. 2010. <http:\\www.statsoft.com.br>

TEODORO, P. E.; BARROSO, L. M. A.; NASCIMENTO, M.; TORRES, F.E. ; SAGRILO, E.; SANTOS, A.; RIBEIRO, L. P. Redes neurais artificiais para identificar genótipos de feijão-caupi. **Pesq. agropec. bras.,** Brasília, v.50, n.11, p.1054-1060, 2015.

YOUNG, P.A. The based culture model: constructing the model of culture. **Educational & Technology Society**, v.11, n.2, p.107- 118, 2008.

Considerações Finais

A utilização de inteligência computacional e modelos matemáticos nos processos de predição de dados, sejam eles de natureza genotípicas ou fenotípicas, são importantes, pois agregam conhecimentos no desenvolvimento e inovação para a pesquisa cientifica.

A simulação de dados buscou atender a junção do conhecimento sobre estágios de vida de indivíduos x ambiente x genética, para agregar conhecimento sobre a dinâmica que ocorre em sistemas florestais naturais, tornando assim, a simulação mais próxima dos cenários reais observados nestes habitats.

As RNAs apresentaram resultados importantes na estimativa da variável em estudo, podendo o modelo ser utilizado no futuro para determinação de rendimento com precisão, em vista que o modelo de RNA proposto, sendo uma nova alternativa viável para determinação desta e outras variáveis em diversas culturas de importância agronômica.

O trabalho buscou contribuir na utilização de alternativas de baixo custo e de fácil aprendizagem como parte de planejamento para o desenvolvimento de pesquisas e crescimento intelectual.