ISABELA DE CASTRO SANT'ANNA

REDES NEURAIS ARTIFICIAIS PARA PREDIÇÃO GENÔMICA NA PRESENÇA DE INTERAÇÕES EPISTÁTICAS

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

VIÇOSA
MINAS GERAIS - BRASIL
2018

Ficha catalográfica preparada pela Biblioteca Central da Universidade Federal de Viçosa - Câmpus Viçosa

T
S232r
2018

Sant'Anna, Isabela, 19-
Redes Neurais Artificiais para Predição Genômica na Presença de Interações Epistáticas. / Isabela Sant'Anna. – Viçosa, MG, 2018.
x, 93f. : il. (algumas color.) ; 29 cm.

Inclui anexo.
Orientador: Cosme Damião Cruz.
Tese (doutorado) - Universidade Federal de Viçosa.
Inclui bibliografia.

1. Seleção Genômica. 2. Melhoramento genético. 3. . I. Universidade Federal de Viçosa. Departamento de Biologia Geral. Doutorado em Genética e Melhoramento. II. Título.

CDD 22. ed. 572.8

ISABELA DE CASTRO SANT'ANNA

REDES NEURAIS ARTIFICIAIS PARA PREDIÇÃO GENÔMICA NA PRESENÇA DE INTERAÇÕES EPISTÁTICAS

Tese apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Doctor Scientiae*.

APROVADA: 23 de fevereiro de 2018.

Leonardo Siqueira Glória

Fabyano Fonseca e Silva

Camila Ferreira Azevedo

Moysés Nascimento

Cosme Damião Cruz
(Orientador)

Esta Tese é dedicada a você.

AGRADECIMENTOS

A Deus por todas as oportunidades recebidas em minha vida, por iluminar minhas escolhas e apontar meus caminhos.

À Universidade Federal de Viçosa, minha casa, instituição a qual me formei no COLUNI, no curso de Ciências Biológicas, no mestrado em Genética e Melhoramento e agora mais uma vez muito obrigada pela formação e oportunidades em minha vida.

Ao Conselho Nacional do Desenvolvimento Cientifico e Tecnológico (CNPq), pela concessão da bolsa de estudos e pela oportunidade de fazer o Doutorado Sanduiche.

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

Ao professor Moysés Nascimento, pela sua co-orientação, dedicação, ensinamentos, amizade e as valiosas contribuições para realização desse trabalho.

Aos meus co-orientadores: Marcos Deon Vilela de Resende, pelos ensinamentos, e Professor Matias Kirst, pela orientação durante o ano de doutorado Sanduiche.

Ao professor Fabyano Fonseca, Camila Ferreira Azevedo, Leonardo Siqueira Glória, pela amizade e disponibilidade em participar da banca contribuindo para melhoria deste trabalho.

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

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

Aos colegas do BIOINFO, por tornar agradáveis as horas de trabalho, pelos conselhos, pelos cafés, rizadas, pela amizade e aprendizado. Um agradecimento especial a Alexandre, Antônio, Cristiano, Daiana, Francyse, Ivan, Iara, Luciano, Luiza, Marciane, Rafael, Renato, Vinícius e Ricardo pelas valiosas grandes contribuições para realização desse trabalho. A Gabi e Laís por formarem comigo um trio inseparável sempre presente nos momentos de alegria, de desafios pessoais e profissionais, compartilhando aprendizados, cadernos, discussões, cafés, almoços, doces e carinho. Aos colegas do Licae pela amizade e pelas importantes reuniões de muito aprendizado estatístico e pelas divertidas conversas na hora do café.

À University of Florida e ao Forest Genomics lab por me receber durante o período de doutorado sanduíche. Em especial a Marcio Resende e Rodrigo Furtado pelos valiosos ensinamentos.

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

Aos meus amigos de Ervália, de Viçosa, do COLUNI 04/06, da Biologia, da UFV, da Universidade da Florida e de Gainesville pelos infinitos conselhos e torcida.

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

A minha irmã Lívia pela incentivo, apoio, torcida, amizade e carinho.

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

Aos meus avós Conceição e Quim pelo amor, carinhos, ensinamentos, minha eterna saudade *(in memoriam)*. Aos meus avós Hélio e Zelia pelo amor incondicional e a torcida.

Ao meu namorado Henrique por todo companheirismo, amor, apoio por me acalmar e descontrair nas horas difíceis durante todo esse tempo de doutorado.

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

BIOGRAFIA

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

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

Em 2004, iniciou o ensino médio no Colégio de Aplicação-COLUNI (UFV) na cidade de Viçosa-mg que foi concluído em 2006.

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

Em setembro de 2012, iniciou o Mestrado em Genética e Melhoramento pela Universidade Federal de Viçosa que foi concluído em 2014.

Em março de 2014, iniciou o doutorado em Genética e Melhoramento pela Universidade Federal de Viçosa.

SUMÁRIO

RESUMO..... vii
ABSTRACT..... ix
INTRODUÇÃO GERAL..... 1
REVISÃO DE LITERATURA……..... 3
1. Uso de simulação para estudos de genética quantitativa ……..... 4
1.1 Princípios de genética quantitativas …….. 4
2. A genômica no melhoramento genético 10
3. A seleção genômica (SG) no melhoramento genético…….. 11
3.1 Ridge Regression BLUP (RR-BLUP)..... 12
3.2 Demais Métodos Estatísticos de Predição Genômica..... 14
4. Redes Neurais Artificiais no melhoramento genético 16
4.1 Perceptron Múltiplas Camadas- RNA-MLP 18
4.2. Redes Neural Função de Base Radial..... 21
5. REFERENCIAS BIBLIOGRÁFICAS..... 23
CAPITULO1 32
APLICAÇÃO DE REDES NEURAIS DE BASE RADIAL (RNA-RBF) E SELEÇÃO GENÔMICA NA PREDIÇÃO DE VALORES GENÉTICOS.....
1. INTRODUÇÃO..... 33
2. MATERIAL E MÉTODOS 35
3. RESULTADOS E DISCUSSÃO..... 43
4. CONCLUSÃO..... 48
5. REFERÊNCIAS BIBLIOGRÁFICAS..... 52
CAPITULO 2. ……. 57
PREDIÇÃO GENÔMICA DE CARACTERES QUANTITATIVOS POR MEIO DE REDES NEURAIS ARTIFICIAIS APÓS REDUÇÃO DA DIMENSIONALIDADE……...
1. INTRODUÇÃO..... 62
2. MATERIAL E MÉTODOS..... 64
3. RESULTADOS E DISCUSSÃO..... 75
4. CONCLUSÕES..... 86
5. REFERÊNCIAS BIBLIOGRÁFICAS..... 87
CONCLUSÕES GERAIS..... 92
ANEXOS……... 93

RESUMO

SANT'ANNA, Isabela de Castro, D.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Redes Neurais Artificiais para Predição Genômica na Presença de Interações Epistáticas.** Orientador: Cosme Damião Cruz. Coorientadores: Matias Kirst e Marcos Deon Vilela de Resende.

A identificação de genótipos com desempenho superior é um dos principais objetivos da maioria dos programas de melhoramento de plantas. No entanto, a capacidade de atingir esse objetivo é limitada pelo alto custo da fenotipagem e realização de experimentos. Neste contexto, a Seleção Genômica (SG) foi proposta para estimar o valor genético (VGG) de indivíduos que ainda não foram fenotipados por meio de informações de marcadores distribuídos em todo o genoma. No entanto, a maioria das modelagens da SG expressam o valor fenotípico como função apenas do efeito aditivo do valor genotípico o que dificulta, muitas vezes, uma representação mais realística da arquitetura genética de caracteres quantitativos, sendo a inclusão de efeitos dominância e interações epistáticas fatores cruciais para aumentar a acurácia da predição. O papel da epistasia na arquitetura genética de caracteres complexos tem sido discutido desde o surgimento da genética quantitativa e, embora seja visto por diferentes perspectivas, o reconhecimento sobre sua importância é crescente. Nas populações, a variância genética total é dividida em componentes de variância aditivo, de dominância e de epistasia, que dependem dos efeitos dos locos e das frequências dos alelos presentes na população. Assim, se a frequência do alelo epistático varia entre as populações, o efeito do gene de interesse pode significativo em uma população, mas não em outra, e o efeito pode até mesmo ser inverso sobre o caráter em ambientes diferenciados. Neste contexto, as Redes Neurais Artificias (RNAs) tornam-se alternativas de análise promissoras por capturar relações não lineares entre os marcadores a partir dos próprios dados, o que a maioria dos modelos comumente utilizados na SG não conseguem. Entretanto, a inclusão de todos os marcadores no genoma no modelo aumenta as chances de existência de alta correlação entre eles e representa um enorme desafio computacional, que acarreta menor precisão no treinamento da RNA, que utilizam boa parte de seus recursos para representar porções irrelevantes do espaço de busca, dificultando o aprendizado. Assim, um modelo mais realístico deveria incluir apenas os SNPs (Single Nucletiode polymorphism) ao caráter de interesse. Para minimizar os efeitos da dimensionalidade sobre a modelagem de SG usando RNA foi proposta, no presente trabalho, a utilização de métodos de redução

de dimensionalidade do tipo Sonda e *Stepwise* para fins de seleção de um subconjunto de marcadores que serão utilizados na predição do valor genético. Após a seleção de marcadores, foi avaliada a eficiência do método de seleção genômica RR-BLUP e das redes neurais artificias do tipo de base radial (RNA-RBF) e Perceptron de Múltiplas camadas (RNA-MLP) na predição do valor genético em população natural com desequilíbrio gamético. Para isso, foi simulada uma população F_1 oriunda da hibridação de genitores divergentes, com 500 indivíduos, genotipados com 1000 marcadores do tipo SNP. As características fenotípicas foram determinadas adotando-se três modelos: aditivo, aditivo-dominante e epistático, atendendo duas situações de dominância: parcial e completa com caracteres quantitativos admitindo herdabilidades (h^2) de 30 e 60%, controlados cada um por 100 locos, considerando dois alelos por loco, totalizando 12 cenários distintos. Para avaliar a capacidade de predição, o modelo RR-BLUP e RNA-RBF foram treinados utilizando 80% dos indivíduos da população e procedimento de validação cruzada com cinco repetições. Para tanto foram obtidos o quadrado da correlação entre o valor genômico predito (GEBV) e o valor genotípico/fenotípico para medir a acurácia seletiva (R^2) e a raiz do erro do quadrado médio (REQM), para medir a acurácia preditiva. Os resultados obtidos pela validação genotípica no primeiro capitulo mostraram que o uso de redes neurais permite capturar as interações epistáticas levando a uma melhora na predição do valor genético e, principalmente, a grande redução da raiz do erro médio quadrado (REQM), o que indica maior confiabilidade da predição do valor genômico. No entanto, a partir dos resultados obtidos por validação fenotípica foi evidente que a acurácia de predição poderia ser melhorada ao introduzir a seleção de marcadores. Consequentemente, no segundo capítulo de trabalho, após aplicar os métodos de redução de dimensionalidade, sonda e *Stepwise*, acurácia de predição aumentou. Por exemplo, para a h^2= 0.3 no cenário aditivo, o R^2 de validação foi de 59.1% para rede neural (RNA-RBF), 57% (RNA-MLP) e 57% para RR-BLUP e, no cenário epistático, os valores de R^2 foram de 50%, 47 e 41%, respectivamente. Adicionalmente, ao analisarmos REQM, a diferença entre os desempenhos das técnicas é ainda maior. Para o cenário 1, as estimativas foram de 91 (RR-BLUP) e 5 para ambas as redes neurais e, no cenário mais crítico que incluía epistasia e dominância, de 427(RR-BLUP) e 20 para as redes neurais. Os resultados obtidos mostram que a utilização de redes neurais permite capturar as interações epistáticas levando a um aumento na acurácia da predição do valor genético e, principalmente, redução do erro quadrático médio, o que indica maior confiabilidade da predição do valor genômico.

ABSTRACT

SANT'ANNA, Isabela de Castro, D.Sc., Universidade Federal de Viçosa, February, 2018. **Artificial Neural Network for Genomic Prediction of genetics values with espistatics interactions.** Adviser: Cosme Damião Cruz. Co-advisers: Marcos Deon Vilela de Resende and Matias Kirst.

The identification of elite individual is a critical component of most plant breeding programs. However, the ability to achieve this goal is limited by the high cost of phenotyping and conducting experiments. In this context the genomic selection was proposed to use all marks presents in the genome to estimate the genomic breeding value of individuals (GEBV) without the need to phenotyping. However, most applications of GS includes only the additive portion of the genetic value, and a more realistic representation of the genetic architecture of quantitative traits should have the inclusion of dominance and epistatics interaction. The role of epistasis in the genetic architecture of quantitative traits has been debated since first formulations of quantitative genetic theory, and different perspectives regarding the importance of epistasis arise. In populations, the total genetic variance is partitioned into components that are attributable to additive, dominance and epistatic variance, which depend on allele frequencies. If the allele frequency of the interacting locus varies among populations, the effect of the target locus can be significant in one population but not in another, or can even be of the opposite sign. In this context, Artificial Neural Networks (ANNs) has a great potential because they can capture non-linear relationships between markers from the data themselves, which most of the models commonly used in the GS can not. However, the inclusion of all markers in the prediction model increases the chances of a high correlation between the marks and represents a huge challenge that add less precision and a great computational demand for ANNs training that use a good part of their resources to represent irrelevant portions of the search space and compromising the learning process. Thus, a more realistic model should include only SNPs that are related to the traits of interest. Because of this, it was proposed to use dimensionality reduction methods, applied to the prediction of genetic values, for the purpose of selecting a subset of markers by means of specific procedures such as Sonda or Stepwise regressions. In this way, the objective of this work is to evaluate the efficiency of genome enabled prediction by using RR-BLUP (GS) and artificial neural networks as radial basis function neural network (RBFNN), and Multi-layer Perceptron (RNA-MLP) in the prediction of the genetic value in a natural population with linkage disequilibrium without (chapter 1) and with (chapter 2) the dimensionality reduction. For this, an F_1 population from the hybridization of

divergent parents with 500 individuals genotyped with 1,000 SNP-type markers was simulated. The phenotypic traits were determined by adopting three different gene action models: additive, additive-dominance and epistasis, attending two dominance situations: partial and complete with quantitative traits admitting heritabilities (h^2) ranging from 30 to 60%, each is controlled by 50 loci, considering two alleles per loco, totaling 12 different scenarios. To evaluate the predictive ability of RR-BLUP and the neural networks a cross-validation procedure with five replicates were trained using 80% of the individuals of the population. Two dimensionality reduction methods Stepwise and Sonda were used to calculated the square of the correlation between predicted genomic value (GEBV) and genotype/phenotype value was used to measure predictive reliability(R^2) and the predictive mean-squared error root (MSER). In the chapter one of this work the results showed that the use of neural networks allows capturing the epistasic interactions leading to an improvement in the accuracy of the prediction of the genetic value and, mainly, a large reduction of the mean square error root (MSER) that indicates greater reliability of the prediction of the genomic value. But from the results using phenotypic validation it was clearly that is possible to make further improvements on the accuracy by introducing the variable selection. Consequently, in the second chapter, after applied the dimensionality reduction methods, the the accuracy increased. For example, for $h^2 = 0.3$ in the additive scenario, the validation R^2 was 59% for neural network (RBFNN), 57% (RNA-MLP) and 57% for RR-BLUP, and in the epistemic scenario R^2 values were 50%, 47 and 41%, respectively. Additionally, when analyzing the mean-squared error root the difference in performance of the techniques is even greater. For additive scenario, the estimates were 91 (RR-BLUP) and 5 for both neural networks and, in the most critical scenario, 427 (RR-BLUP) and 20 for neural networks. The results show that the use of neural networks allows capturing the epistasis interactions leading to an improvement in the accuracy of the prediction of the genetic value and, mainly, a large reduction of the mean square error root that indicates greater reliability of the prediction of the genomic value.

1. INTRODUÇÃO GERAL

Um dos grandes desafios do melhoramento genético atual é entender a variação da herança genética de caracteres quantitativos, QTL (*Quantitative trait loci*), que são condicionados por grande número de genes com pequeno efeito (Risch, 2000) cuja interação resulta, muitas vezes, em não linearidade nas relações entre fenótipos e genótipos (Gianola e de los Campos, 2008; Yamamoto et al., 2008; Mackay, 2014).

O avanço das tecnologias de genotipagem e o aparecimento dos marcadores do tipo SNP, que são abundantemente distribuídos nos genomas de muitas espécies, permitiu o surgimento da Seleção Genômica (SG) que foi proposta por Meuwissen et al. (2001) como uma forma de aumentar a eficiência seletiva, diminuir os custos via redução da frequência de fenotipagem e aumentar os ganhos anuais reduzindo o ciclo de geração (Rutkoski et al., 2010). A SG permite a estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma de forma que a maioria dos alelos de interesse esteja associado a esses marcadores (em desequilíbrio de ligação) para explicar grande parte da variação genética e predizer o valor genético dos indivíduos que ainda não tiveram seus fenótipos coletados (Resende et al., 2014).

O valor genético de indivíduos pode ser atribuído em suas porções relativas aos efeitos aditivos dos alelos, aos efeitos da dominância, que expressam as interações intra-alélicas, e aos efeitos epistáticos, que expressam as interações inter-alélicas. As inclusões dessas interações entre alelos podem ser exploradas pelos métodos de SG através de modelagens adequadas que incorporem a complexidade da arquitetura genética dos caracteres envolvidos. No entanto, a maioria das modelagens da SG usa apenas modelos de regressão que contemplam, em seu modelo, apenas a porção aditiva do valor genético, o que dificulta, muitas vezes, uma representação mais realística da arquitetura genética dos caracteres quantitativos, sendo a inclusão de dominância e interações epistáticas fatores importantes para aumentar a acurácia da predição (Lee et al., 2008; Akidemir et al., 2017).

Inúmeros estudos evidenciam que agregar múltiplas interações intra e interalélicas, ainda que em pequenos efeitos, é importante para explicar ao máximo a herdabilidade de caracteres quantitativos em estudos de prediçao em todo o genoma (McKinney e Pajewski 2012). Segundo Holland (2006), a epistasia têm importância inquestionável em características quantitativas como sugerido por numerosos estudos clássicos de genética quantitativa realizados em plantas. Entretanto, a inclusão de interações epistáticas é ainda mais difícil devido aos efeitos multiplicativos entre os

alelos envolvidos na determinação do caráter, o que leva a uma superparametrização dos modelos agravando ainda mais problemas estatísticos e de processamento de dados, já que o número de marcadores sempre é muito superior ao número dos indivíduos Gianola et al., (2006).

Nesse contexto, uma metodologia disponível para a seleção de genótipos superiores que tem sido empregada nos programas de melhoramento vegetal e animal, ainda que de forma tênue, são as redes neurais artificiais (RNA). Diferentemente das outras modelagens estocásticas, as RNA utilizam os princípios de aprendizado a partir de um conjunto amplo de informação do desempenho do genótipo que pode ser direta ou indiretamente mensurada. Assim, as redes neurais funcionam como cérebro humano, captam toda informação disponível para gerar um critério de tomada de decisão e podem acomodar a arquitetura de características complexas.

As RNAs podem capturar relações não lineares que a maioria dos modelos comumente utilizados na SG não conseguem (Haykin, 2009; Gianola et al., 2006; Long et al., 2010 e Howards et al., 2014) já que a arquitetura das características é inferida dos próprios dados utilizados em seu treinamento, há conexão entre informações acumuladas em diferentes neurônios e diferentes camadas, não necessitam do conhecimento de distribuições a priori como os métodos bayesianos e não há necessidade de atender pressuposições sobre as distribuições dos dados e dos resíduos.

Os desafios estatísticos relacionados à alta dimensionalidade onde o número de marcadores (m) é muito maior que o número de observações (n) ($m > n$) e a alta correlação existente entre os marcadores provocada pelo desequilíbrio de ligação (Crossa et al., 2017) representa enorme desafio computacional já que existem centenas de milhares de marcadores disponíveis no genoma (Long et al., 2011). Para as técnicas fundamentadas em inteligência computacional a grande quantidade de marcadores disponível no genoma acarreta menor precisão e grande demanda computacional para o treinamento da RNA que utilizam boa parte de seus recursos para representar porções irrelevantes do espaço de busca, dificultando o aprendizado.

Segundo Long et al. (2011), um modelo mais realístico incluiria SNPs apenas relacionadas a característica de interesse. Uma solução é escolher um subconjunto de SNPs para o treinamento de dados pois com a redução do espaço de buscas as (RNAs) melhora o processo de aprendizado e aumenta o poder preditivo de modelo (Long et al., 2010).Os métodos propostos para redução de dimensionalidade incluem métodos baseados em regularização tais como penalizações (RR-BLUP), redução via combinações lineares independentes (Horn e Camp, 2004; Long et al., 2010; Woolaston et al., 2007;

Azevedo, et al., 2013; 2014; James et al., 2013; Resende et al., 2014), mínimos quadrados parciais (Moser et al., 2007; Tier et al., 2007), e seleção de um subconjunto de marcadores por meio de procedimento específicos como sondagens ou regressão *Stepwise* (Habier et al., 2007; Piyasatian et al., 2007).

Para tanto, o objetivo deste trabalho é propor abordagem de RNA para fins de Seleção Genômica na presença de interações epistáticas. Além disso, após a utilização de métodos de redução da dimensionalidade pretendemos demonstrar que os resultados obtidos pela predição do valor genético com informações reduzidas melhoram a predição e preservam as mesmas conclusões biológicas quando se utiliza um conjunto de dados maior e que, na situação de menor dimensionalidade, ainda podemos acrescentar a utilização de redes neurais artificiais que envolvam topologias mais complexas para fins de predição. Dessa forma, pretendemos observar se a presenças das interações epistáticas afetam as predições genômicas, com a expectativa de que a abordagem das redes neurais artificiais possa capturar padrões complexos de interação não lineares. Além disso, descrevemos também uma nova estratégia de seleção de variáveis, para fins de redução de dimensionalidade, aplicável a predição de valores genômicos. A abordagem RNA foi comparada ao método RR-BLUP.

2. REVISÃO DE LITERATURA

2.1. Uso de simulação para estudos de genética quantitativa

Os métodos clássicos de melhoramento genético são responsáveis pelo progresso contínuo da grande maioria das cultivares disponibilizadas no Brasil e no mundo, (Vencovsky e Ramalho, 2000). Entretanto, essa metodologia convencional é relativamente lenta tendo em vista a demanda crescente de produtos melhorados e a introdução de um gene, ou um conjunto de genes, por estes métodos convencionais exige repetidos cruzamentos e ciclos de seleção, o que torna este processo restrito às espécies de reprodução sexuada.

O aumento da eficiência nos programas de melhoramento genético é um processo constante devido à pressão para aumentar a produtividade e a estabilidade das espécies cultivadas. A maioria das alternativas existentes precisam da realização de experimentos em várias condições, demandando muito tempo e recurso técnico e financeiro. Neste contexto, a utilização dos recursos computacionais faz-se uma grande aliada na obtenção de bancos de dados confiáveis para estudos ligados ao melhoramento genético permitindo

que os modelos, métodos e estratégias mais eficazes sejam testados em condições práticas.

A utilização da simulação computacional tem sido de grande utilidade em estudos genéticos sob vários contextos, como o de populações, do indivíduo ou do próprio genoma. Segundo Cruz e Sant'Anna (2016), a simulação computacional demanda dos geneticistas o desenvolvimento de modelos biológicos que retratem, da melhor maneira possível, os fenômenos de interesse, e dos programadores as rotinas para o processamento adequado, apesar de impor restrições, para que a influência de certos fatores possa ser avaliada.

A simulação de populações para estudos da estrutura genética é importante e vem sendo utilizada atualmente em vários ramos da genética, como em análise discriminante em estruturas de populações (Sant'Anna, 2014), estudos de genoma (Price et al., 2006), detecção das interações gene-a-gene e variância genética (Bhattacharya et al., 2010).

2.2. Princípios de genética quantitativas

O melhoramento de plantas se baseia na manipulação de caracteres quantitativos e, para isso, é preciso estimar os parâmetros populacionais e estudar formas de manipulá-los seja pelo controle dos acasalamentos, pela endogamia ou pela própria seleção. Os objetivos continuam sendo essencialmente os mesmos, ou seja, identificar, acumular e perpetuar genes favoráveis (Cruz et al., 2014). Estas são características da genética quantitativa que a torna tão relevante para o melhoramento de plantas.

A Genética Quantitativa é a área da genética que permite estudar os caracteres quantitativos, sua herança e os componentes determinantes de sua variação. Caracteres quantitativos são, em geral, controlados por vários genes e muito influenciados pelo ambiente, exibindo, desta maneira, variações contínuas (às vezes descontinuas), ao passo que os caracteres qualitativos são de herança monogênica e tem pouca ou nenhuma influência do ambiente (Falconer e Mackay, 1996).

Por essa razão o estudo de caracteres de herança quantitativa não se baseia somente no indivíduo, o que força o pesquisador a basear seus estudos a grupos maiores de indivíduos, adotando-se ainda, um modelo biométrico. Assim, se o efeito do ambiente pode tanto aumentar quanto diminuir a manifestação fenotípica de um caráter, a média de um conjunto de indivíduos será uma medida mais confiável, pois os efeitos do ambiente tendem a se cancelar. Outra medida usada para definir uma população, além da media, é a variância. Logo, no estudo da herança de caracteres quantitativos avaliam-se quais frações da média e da variância são herdáveis (Cruz, 2012).

2.2.1. Modelo para estudos genéticos de caracteres quantitativos

Os estudos genéticos de caracteres quantitativos são realizados adotando-se o modelo básico de acordo com a equação a seguir:

$$F=G + M$$

em que,

F: é o valor fenotípico, medido nos indivíduos;

G: é o valor genotípico;

M: é o desvio causado pelo ambiente;

Segundo Falconer (1987), o valor fenotípico do indivíduo é determinado pelo genótipo e pela fração influenciada pelo ambiente, logo, entende-se que o valor fenotípico é o valor observado quando uma característica quantitativa é mensurada em um indivíduo, e por esse motivo, observações como média e variância devem ser baseadas nessa medida.

De acordo com Cruz (2012), o valor genotípico é definido como sendo a constituição genética do indivíduo e é de grande importância em processos de seleção e identificação de superioridade genética. O valor genotípico, para um determinado loco, pode ser desdobrado em uma fração herdável, denominada valor genético aditivo ou simplesmente valor aditivo, e a outra não herdável por processos sexuados, correspondente aos valores atribuídos aos desvios de dominância.

A partir da mensuração dos valores fenotípicos, torna-se possível estimar a variância fenotípica (σ_F^2) e, sob determinadas condições, desdobrá-la nos componentes de variância genética ou genotípica (σ_G^2) e em variância ambiental (σ_M^2), definida pela equação abaixo:

$$\sigma_F^2= \sigma_G^2+ \sigma_M^2 +2 \text{ COV (G,M)}$$

Como a avaliação dos genótipos é feita sob os princípios básicos da experimentação, a ação do genótipo e os desvios devido aos efeitos ambientais atuam de forma independente, de modo que a covariância entre elas é nula, ou seja:

$$\sigma_F^2= \sigma_G^2+ \sigma_M^2$$

2.2.2. Estimativas dos parâmetros genéticos

Segundo Cruz (2012), com as estimativas da média e da variância, torna-se possível a obtenção dos parâmetros genéticos necessários para avaliar o potencial genético para a prática do melhoramento e seleção.

Um dos parâmetros genéticos mais importantes é a herdabilidade. Segundo Falconer (1987), este parâmetro expressa a proporção da variabilidade observada devido aos efeitos aditivos dos genes, isto é, representam a proporção herdável da variabilidade total.

Essa fração da variância é expressa pela razão entre a variância genética e a variância fenotípica como é mostrado na equação abaixo:

$$h^2 = \frac{\sigma_G^2}{\sigma_F^2}$$

2.3.1.Decomposição da variância genética

O valor genotípico pode ser desdobrado em três partes: a fração herdável, também denominada valor genético aditivo, e a fração não herdável por processos sexuados, correspondente aos desvios de dominância devido às interações intra-alélicas e aos efeitos epistáticos devido às interações interalélicas (Allard, 1964). A partir destes valores, pode-se estimar a variância genotípica dada por:

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$$

em que σ_A^2 é a variância aditiva; σ_D^2é a variância atribuída aos desvios de dominância (devido às interações intra-alélicas) e σ_I^2 é a variância atribuída aos efeitos epistáticos (devido às interações interalélicas).

Dessa forma, o valor fenotípico total, no modelo aditivo-dominante considerando g locos, expresso por um determinado indivíduo i pertencente à população é estimado a partir da expressão abaixo:

$$Y_i = \mu + \sum_{j=1}^{g} p_j \alpha_j + E_i$$

Em que:

$\alpha_j = a_i + d_i$ e d_i/a_i = gmd (grau médio de dominância) , sendo $\mu + a_j$, $\mu + d_j$ e $\mu - a_j$ os valores genotípicos associados as classes AA, Aa e aa, considerando igual a 1, 0 ou -1, respetivamente quando codificadas, e p_j a contribuição do loco j para a manifestação da característica considerada, no trabalho como tendo distribuição binomial.

O valor de d_i foi estabelecido a partir do grau médio da dominância manifestado em cada característica.

Na presença de epistasia, além da contribuição dos locos individuais, por meio de seus efeitos aditivos e dominantes é preciso considerar um segundo somatório que representa os efeitos multiplicativos que correspondem as interações epistáticas entre pares de locos. Sendo assim, o valor fenotípico total, no modelo epistático, expresso por um determinado indivíduo pertencente à população e estimado a partir da expressão apresentada a seguir. Neste modelo simplificado, a epistasia considerada traduz a ação de g genes em uma cadeia de uma via metabólica considerando a existência apenas interações mais significativas, no sentido único da cadeia, de par em par de locos consecutivos.

$$Y_i = \mu + \sum_{j=1}^{g} p_j \alpha_j + \sum_{j=1}^{g-1} p_j \alpha_j \, \alpha_{j+1} + E_i$$

Em que:

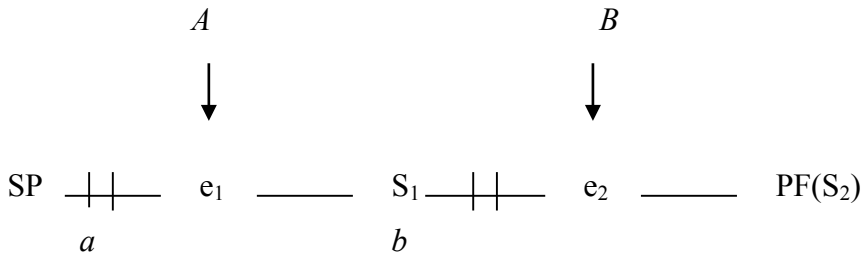
$\alpha_j = a_i + d_i$ e d_i/a_i = gmd, sendo $\mu + a_j$, $\mu + d_j$ e $\mu - a_j$ os valores genotípicos associados as classes AA, Aa e aa,considerando igual a 1, 0 ou -1, respetivamente quando codificadas, e p_jé a contribuição do loco j para a manifestação da característica.

2.3.2. Importância da epistasia

2.3.2.1. Epistasia no contexto biológico

Num contexto biológico, as interações do tipo epistáticas ocorrem quando dois ou mais genes determinam a produção de enzimas que catalisam diferentes etapas de uma mesma via biossintética. Vias biossintéticas são aquelas em que as enzimas produzidas por determinados genes atuam, de maneira que uma substância inicial (substância precursora) é desdobrada em substratos até dar origem a um produto final, que pela ação do meio resultará num determinado fenótipo para aquele caráter.

A epistasia (do grego epi, sobre, e stasis, parada, inibição) envolve a supressão gênica inter-alélica, ou seja, os alelos de um loco gênico encobrem, ou suprimem, a expressão de outro alelo pertencente a outro loco gênico (não-alelo). Isto pode ser evidenciado, de uma forma simples, na via biossintética ilustrada a seguir:



nesta via identificamos alguns elementos básicos:

SP: substância precursora

S_1: substância intermediária

PF (ou S_2): produto final ou nova substância intermediária

e_1 e e_2: enzimas

Todos os produtos da via, por ação do ambiente, poderão produzir determinado fenótipo. Assim, como ilustração consideraremos o seguinte mecanismo de ação sobre uma característica que representa a coloração das flores de uma planta, sob ação dos genes *A/a* e *B/b:*

A-: permite a expressão da cor

aa: inibe a cor (branco)

B-: manifesta a cor vermelha

bb: manifesta a cor amarela

Verifica-se, portanto, que o fenótipo apresentado por um indivíduo de genótipo *B*- dependerá da ação do gene *A*/*a*. Se houver o padrão genotípico *A*-, o fenótipo será vermelho, pois a enzima produzida pelo alelo B atuará catalisando a transformação de S_1 em S_2. Entretanto, na presença do padrão genotípico *aa*, a ação do gene *B* será suprimida, pois apesar de produzir a enzima e_2, não haverá substrato para sua ação catalítica. O mesmo fato ocorre em relação à condição genotípica *bb*. Pode-se, portanto, afirmar que a condição aa inibe, ou suprime, a ação do loco *B/b*, ou ainda, que aa é epistático em relação ao gene *B/b.*

O alelo (ou gene) que mascara a expressão do outro é denominado de epistático a esse alelo (ou gene). O alelo (ou gene) cuja ação suprimida é denominado de hipostático. No exemplo em consideração, o alelo *a* é o epistático e o gene *B/b* o hipostático.

Quando se verifica epistasia entre dois locos gênicos, o número de fenótipos entre os descendentes será menor que quatro. A proporção 9:3:3:1 se modifica dando origem a

uma nova proporção que é combinação da 9:3:3:1.

2.3.2.2. Epistasia no contexto de Genética Quantitativa

Quando mais de um loco é envolvido no controle gênico de uma característica é possível considerar que eles possam interagir. Segundo Falconer (1987), as interações envolvendo mais de dois locos, ao mesmo tempo, contribuem muito pouco para a variância, de forma que elas possam ser ignoradas. Sendo assim, num contexto de genética quantitativa, a epistasia foi proposta por Fisher (1918) como qualquer interação entre locos que resulta em desvio da combinação aditiva de dois loci em relação à sua contribuição para um fenótipo e pode ser subdivida em três fatores de acordo com a existência ou não de dominância na caraterística de interesse. Conforme mencionado anteriormente, a variância genotípica é dada por: $\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2$.

A variância devido a epistasia σ_I^2 que representa a interação entre locos pode ser do tipo aditiva x aditiva (quando ocorre a interação entre dois valores aditivos, σ_{AA}^2), aditiva x dominante (quando ocorre interação entre valores aditivos de um loco e desvio provocado pela dominância em outro, σ_{AD}^2) e dominante x dominante (interação entre dois desvios causados por dominância , σ_{DD}^2). Assim muitos vezes, os componentes $\sigma_D^2 + \sigma_I^2$ são conhecidos como variância não-aditiva.

Assim :

$$\sigma_I^2 = \sigma_{AA}^2 + \sigma_{DA}^2 + \sigma_{DD}^2$$

Utilizando um modelo biométrico (Cruz et al., 2014) considerando apenas dois locos gênicos, com dois alelos cada podemos ilustrar a existência de epistasia.

Num modelo envolvendo dois locos, com os alelos i e j no primeiro loco (por exemplo i = A e j = a) e k e l no segundo loco (por exemplo k = B e l = b. Também será admitido, inicialmente, que a população esteja em equilíbrio de ligação gênica. Assim, os valores genotípicos são expressos por:

$$Y_{ijkl} = (\alpha_i + \alpha_j + \delta_{ij}) + (\alpha_k + \alpha_l + \delta_{kl}) + \alpha\alpha + \alpha\delta + \delta\delta$$

sendo:

$\alpha\alpha = \alpha_i\alpha_k + \alpha_i\alpha_l + \alpha_j\alpha_k + \alpha_j\alpha_l$ (aditiva x aditiva)

$\alpha\delta = \alpha_i\delta_{kl} + \alpha_j\delta_{kl} + \alpha_k\delta_{ij} + \alpha_l\delta_{ij}$(aditiva x dominante)

$\delta\delta = \delta_{ij}\delta_{kl}$ (dominante x dominante)

$E = \alpha\alpha + \alpha\delta + \delta\delta$

em que: μ: é a média da população;

α_i α_j são os efeitos gênicos aditivos dos alelos do loco um.

δ_{ij} :é o efeito atribuído ao desvio da dominância entre os alelos do loco um;

α_k α_l: são os efeitos gênicos aditivos dos alelos do loco dois;

δ_{kl} é o efeito atribuído ao desvio da dominância entre os alelos do loco dois;

$\alpha\,\alpha\alpha\alpha$: é o efeito da interação epistática do tipo aditiva x aditiva;

$\alpha\delta$: é o efeito da interação epistática do tipo aditiva x dominante;

$\delta\delta$: é o efeito da interação epistática do tipo dominante x dominante.

Como consequência, os valores genotípicos associados às constituições genotípicas de um loco dependem das constituições genotípicas do outro loco. Assim, a existência de epistasia pode ser evidenciada nas seguintes situações apresentadas a seguir, em que se compara os efeitos de quaisquer dois genótipos referentes a um dos locos em duas situações diferentes, ou seja, na presença de dois outros genótipos relativos ao segundo loco:

$e_{11}= (Y_{ii.kk} - Y_{ij.kk}) - (Y_{ii.kl} - Y_{ij.kl}) =$(AABB-AaBB) – (AABb - AaBb)$\neq 0$

$e_{12}= (Y_{ij.kk} - Y_{jj.kk}) - (Y_{ij.kl} - Y_{jj.kl}) =$(AaBB-aaBB) – (AaBb - aaBb)$\neq 0$

$e_{21}= (Y_{ii.kl} - Y_{ii.ll}) - (Y_{ii.kl} - Y_{ij.kl}) =$(AABb-AaBb) – (AAbb - Aabb)$\neq 0$

$e_{22}= (Y_{ii.kl} - Y_{jj.kl}) - (Y_{ij.ll} - Y_{jj.ll}) =$ (AaBb-aaBb) – (Aabb – a abb)$\neq 0$

ou, de forma mais geral,

$e = (Y_{ii.kk} - Y_{jj.kk}) - (Y_{ii.ll} - Y_{jj.ll}) =$ (AABB+aabb) – (aaBB + AAbb)$\neq 0$

3.0. A genômica no melhoramento genético

O desenvolvimento do melhoramento genético da maioria das espécies é predominantemente realizado por meio das técnicas clássicas que se fundamentam na execução de hibridações artificiais para gerar populações segregantes, que são conduzidas por um método específico e submetidas a vários ciclos de gerações, visando à identificação de genótipos superiores (Cruz et al., 2014). Quando estes genótipos superiores são desenvolvidos, encontra-se grande dificuldade na obtenção de novos genótipos que possam superar aqueles que já são utilizados pelo agricultor. Assim, a melhor maneira é a integração das técnicas clássicas de melhoramento com as aquelas advindas da biotecnologia.

A identificação de marcadores moleculares ligados a genes controlados de características oligogênicas e, ou, quantitativas nos programas de melhoramento genético da maioria das espécies estudadas, tem aumentado a eficiência e reduzido o período de desenvolvimento de cultivares superiores.

A Genômica tornou-se uma aliada de grande relevância para os melhoristas do desenvolvimento dos seus programas de melhoramento. De acordo com Schuster e Cruz (2008), a Genômica é definida como uma ciência que estuda o genoma de forma completa, pela integração de várias áreas tradicionais da genética, como a Genética Mendeliana, a Citogenética, a Genética Molecular, a genética de Populações e a Genética Quantitativa, incluindo também a ciência da computação e os sistemas automatizados.

3.1. Marcadores moleculares

A seleção de características agronômicas monitoradas por marcadores moleculares baseia-se no princípio de que se um gene, ou um bloco de genes, encontra-se ligado a um marcador genético de fácil identificação, então, esse marcador pode ser usado para selecionar a característica de interesse em um programa de melhoramento.

Atualmente os marcadores de DNA são mais utilizados pelos pesquisadores pelo fato desses marcadores serem capazes de estabelecer relações de ligação entre eles e os locos de interesse. De acordo com Tanksley et al. (1989), essa ligação permite, ao pesquisador, inferir sobre a presença dos locos, mediante o ensaio de alguns desses marcadores, facilitando trabalhos de transferência e de mapeamento.

Recentemente, uma nova classe de marcadores moleculares, denominada SNPs (Single Nucleotide Polymorphisms), tem sido muito utilizada pela comunidade científica. Estes marcadores genéticos se baseiam na detecção de polimorfismos resultantes da alteração e uma única base no genoma. Em comparação com marcadores SSR, análise de SNP pode ser feito sem exigir uma separação por tamanho de DNA e, portanto, pode ser automatizado em formatos de ensaio de alta produtividade. Desta forma, estes marcadores podem ser utilizados em programas de melhoramento possibilitando a seleção precoce de indivíduos portadores do gene de interesse.

3.2. A seleção genômica (SG) no melhoramento genético

O avanço das tecnologias de genotipagem e o aparecimento dos marcadores do tipo SNP, que são abundantemente distribuídos nos genomas de muitas espécies,

permitiu o surgimento da Seleção Genômica (SG) que foi proposta por Meuwissen et al. (2001) como uma forma de aumentar a eficiência seletiva, diminuir os custos via redução da frequência de fenotipagem e aumentar os ganhos anuais reduzindo o ciclo de geração (Rutkoski et al., 2010).

A SG permite a estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma de forma que a maioria dos alelos de interesse esteja associado a esses marcadores (em desequilíbrio de ligação) para explicar grande parte da variação genética e predizer o valor genético dos indivíduos que ainda não tiveram seus fenótipos coletados (Resende et al., 2014). Na SG, primeiramente, os efeitos dos marcadores são estimados baseados em dados fenotípicos de uma população conhecida como população de estimação ou treinamento. Uma vez que seus efeitos são modelados e estimados, estes são testados em uma população de validação.

Na população de validação, utiliza-se um conjunto de dados menor do que aquele da população de estimação e contempla indivíduos genotipados e fenotipados para a característica de interesse. Esta amostra independente é utilizada para testar e verificar as acurácias das equações de predição de valores genéticos genômicos. Como a população de validação é independente e não é envolvida na predição dos efeitos dos marcadores, os erros associados aos valores preditos e observados são também independentes, de forma que toda a correlação entre esses valores seja de natureza genética e indique a capacidade preditiva da SG (Goddard e Hayes, 2007; Resende, 2008).

A implementação da seleção genômica impõe desafios estatísticos e computacionais como a dimensionalidade do modelo, colinearidade entre marcas e a complexidade das características quantitativas. Para isso, vários métodos têm sido propostos, que diferem entre si pelo tipo de suposição sobre o modelo genético associado ao caráter quantitativo. Entre eles, o método RR-BLUP (*Ridge Regression-Best Linear Unbiased Prediction*) estima simultaneamente os efeitos de todas as marcas(Meuwissen et al., 2001), sendo estas consideradas efeitos aleatórios com variância comum, ou seja, assumem que todos os marcadores contribuem igualmente para a variação genética (ausência de genes de efeitos maiores).

3.2.1. Ridge Regression BLUP (RR-BLUP)

O método RR-BLUP consiste em uma análise de regressão em ridge ou regressão de cumeeira, baseado em modelos mistos de seleção (Henderson, 1973). Em sua metodologia, o RR-BLUP assume que os marcadores moleculares são variáveis regressoras com efeitos aleatórios no modelo (Meuwissen et al., 2001). Modelos de RR-

BLUP utilizam ainda um procedimento denominado regularização ou *shrinkage* (Cruz et al., 2013), no qual os estimadores associados ao modelo produzem o parâmetro de penalização ou *shrinkage* λ.

$$\lambda = {\sigma_e^2}\Big/{\sigma_{gi}^2} = {\sigma_e^2}\Big/{(\sigma_g^2/n_Q)}, \text{ com } n_Q = 2\sum_i^n p_i(1-p_i)$$

Em que:

σ_{gi}^2: variância genética aditiva associada ao i-ésimo loco;

σ_g^2: variância genética aditiva do caráter;

σ_e^2: variância residual

n_Q: número de locos controladores da característica.

O modelo linear misto geral que define o RR-BLUP é dado, conforme Resende et al. (2008), tal como descrito abaixo:

$$y = Wb + Xm + e$$

Em que:

y: vetor de observações fenotípicas;

b: vetor de efeitos fixos (média geral) com matriz de incidência W;

m: vetor dos efeitos aleatórios dos marcadores com matriz de incidência X;

X: Matriz de incidência composta pelos valores 0, 1 e 2 (ou codificados para -1, 0 e 1, respectivamente) para o número de alelos do marcador dos genótipos mm, Mm e MM, respectivamente;

W: Matriz de incidência (no caso, um vetor caracterizado por uma coluna de 1's);

e: vetor de resíduos aleatórios.

O valor acurado de cada indivíduo, para fins de uso como critério de seleção, obtido pelo método RR-BLUP/GWS é denominado *genomic estimated breeding value* (GEBV), que expressa o valor predito para o j-ésimo indivíduo por meio das informações dos marcadores moleculares incluídos no modelo ajustado (Resende et al., 2014) com a propriedade de melhor explorar a variação genética genômica (VGG) da população.

$$GEBV = \hat{y}_j = \hat{\mu} + \sum_i X_{ij}\hat{m}_i$$

Assume-se que os efeitos de QTL seguem distribuição normal com variância constante associada aos locos controladores da característica quantitativa. Desse modo, cada loco explicará a porção σ_g^2/n_Q de variação genética, em que σ_g^2 é a variação genética total e n_Q é o número de locos (Resende et al., 2014).

Quando λ não é conhecido, a escolha arbitrária do mesmo leva ao método de regressão *Ridge Regression* (RR). Se o parâmetro de regressão for associado a $\lambda = \sigma_e^2/\sigma_{gi}^2 = \sigma_e^2/(\sigma_g^2/n)$ tem-se a regressão aleatória BLUP para o efeito do segmento cromossômico *i*, em que σ_{gi}^2 é a variância genética associada ao loco ou segmento *i* e σ_g^2 e σ_e^2 são a variância genética do caráter e variância residual, respectivamente. A quantidade *n* é desconhecida a priori, mas pode ser inferida por via iterativa ou sintonia, escolhendo-se aquele que maximiza a correlação entre valor fenotípico e valor genético predito na validação cruzada (Resente et al., 2014).

O valor genético genômico global do indivíduo *j* é dado por $VGG = \hat{y}_j = \sum_i Z_i \hat{a}_i$, em que Zi equivale a -1, 0 ou 1 para os genótipos A_1A_1, A_1A_2 e A_2A_2, respectivamente, para marcadores bialélicos e codominantes. As equações de predição apresentadas acima foram modeladas com σ_g^2 comum. Assim, a variação genética explicada por cada loco é dada por σ_g^2/n, em que σ_g^2 é a variação genética total e n é o número de marcadores utilizados em cada um dos testados. Essa estratégia foi adotada por Meuwissen et al. (2001), Muir(2007), Bernardo (2007) e Kolbehdari et al. (2007), dentre outros.

3.2.2. Demais Métodos Estatísticos de Predição Genômica

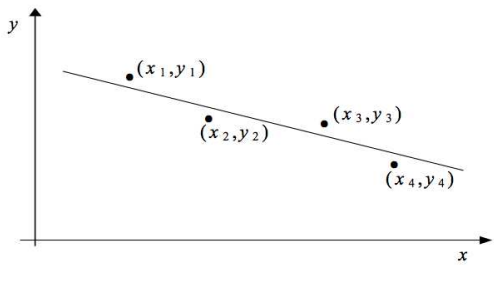
Os modelos estatísticos fornecem diferentes suposições sobre o número e os efeitos dos marcadores associados aos QTL (*Quantitative trait loci),* possivelmente, afetando as respectivas acurácias de GEBV. De forma geral, os métodos apresentam abordagens que diferem na suposição sobre o modelo genético associado ao caráter quantitativo. O método RR-BLUP assume o modelo infinitesimal com muitos locos de pequenos efeitos, ou seja, considera que os efeitos de QTL apresentam distribuição normal com variância constante ao longo dos segmentos cromossômicos (Resende et al., 2010). Por outro lado, os métodos bayesianos, como o Bayes B apresentam grande flexibilidade. Neste método muitos efeitos de marcadores são assumidos como zero, ou próximo a zero, a priori, reduzindo o tamanho do genoma, o que permite o enfoque em regiões do genoma onde realmente existem QTLs. Quando os efeitos de QTL apresentam distribuição exponencial, sendo poucos desses efeitos com valor zero, o melhor estimador dos efeitos

alélicos é o método BLASSO. Desse modo, observa-se que o melhor método de predição genômica é aquele que reflete melhor a natureza biológica do caráter quantitativo sob análise, em termos de efeitos gênicos (Resende et al., 2008).

Desse modo, na implementação da GS existem alguns desafios estatísticos computacionais, como a definição de métodos de predição genômica que possibilitem melhor tratamento dos dados genômicos, considerando a sua dimensionalidade, a colinearidade entre marcadores e a complexidade dos caracteres quantitativos. A escolha de métodos estatísticos para a predição dos efeitos de marcadores também pode afetar a acurácia de GEBV. Assim, o desafio enfrentado está diretamente ligado às pressuposições acerca do modelo avaliado, tais como dimensionalidade das matrizes envolvidas, multicolinearidade entre os marcadores moleculares e a complexidade dos caracteres quantitativos em estudo, com a inclusão das interações intra e inter-alélicas.

Segundo Resende et al. (2014), os principais métodos de seleção genômica podem ser divididos em três classes: regressão explícita, implícita e com redução dimensional. Os métodos da classe de regressão explícita podem ser divididos em dois grupos: (i) métodos de estimação penalizada, como RR-BLUP (Meuwissen et al., 2001; LASSO (Tibshirani, 1996); (ii) métodos de estimação bayesiana, tais como BayesA, BayesB (Meuwissen et al., 2001). Os métodos de regressão com redução dimensional, por sua vez, compreendem os de componentes independentes, quadrados mínimos parciais e de componentes principais (Solberg et al., 2009). Na classe de regressão implícita, destacam-se método semi-paramétrico RKHS (*Reproducing Kernel Hilbert Spaces*) (Gianola e de los Campos, 2009), e as redes neurais artificiais(Gianola et al., 2011) e a regressão Kernel não-paramétrica (Gianola et al., 2006).

Em uma regressão paramétrica a forma do relacionamento funcional entre as variáveis dependentes e independentes é pré-estabelecida, embora possa existir parâmetros cujos valores são desconhecidos, mas passíveis de serem estimados a partir do conjunto de treinamento. Um exemplo é o ajuste de uma reta a uma distribuição de pontos y .



y = ax + b

Figura1: Representação de uma regressão paramétrica em que a,b são valores desconhecidos e y a variável dependente.

Em uma regressão não-paramétrica ocorre ausência (completa ou quase completa) de conhecimento a priori a respeito da forma da função que está sendo estimada. Sendo assim, mesmo que a função continue a ser estimada a partir do ajuste de parâmetros livres, o conjunto de formas que a função pode assumir (classe de funções que o modelo do estimador pode prever) é muito amplo. Como consequência, vai existir um número elevado de parâmetros (por exemplo, quando comparado ao número de dados de entrada-saída para treinamento), os quais não mais admitem uma interpretação física isolada(Von Zubene e Attux, 2001).

Todos os modelos de regressão, que não são puramente paramétricos, são denominados não-paramétricos ou semi-paramétricos. As redes neurais artificiais (RNAs) para treinamento supervisionado pertencem à classe de modelos de regressão não paramétricos. Sendo assim, os pesos não apresentam um significado físico particular em relação ao problema de aplicação. Além disso, estimar os parâmetros de um modelo não-paramétrico (por exemplo, pesos de uma rede neural artificial) não é o principal objetivo do aprendizado supervisionado e sendo este o espaço de aproximação (ou ao menos a saída para certos valores desejados de entrada).

4. Redes Neurais Artificiais no melhoramento genético

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

Dentre as vantagens da utilização da RNAs, ressaltam-se duas: primeiramente, a sua estrutura não linear, capaz de captar complexas características entre o conjunto de dados de entrada (Galvão et al., 1999); e em segundo lugar, a sua capacidade de não requerer informação detalhada dispensando a definição de modelos e estabelecimentos de distribuições acerca dos dados permitindo que qualquer tipo de informação, relacionada ao indivíduo ou à população em estudo, possa ser contabilizada em sua estrutura (Haykin, 2001). Dessa forma, a utilização das redes neurais tem se mostrado mais promissora devido a possibilidade de um desempenho superior aos modelos convencionais utilizados na solução de problemas (Braga et al., 2007).

Nesse contexto, uma linha de pesquisa que tem apresentado destaque em estudos de melhoramento genético é a inteligência computacional, por meio das abordagens de Redes Neurais Artificiais fundamentadas no modelo Perceptron Múltiplas Camadas (RNA-MLP) e Redes de Base Radial (RBF) (Long et al., 2007; 2010; 2011; Gianola et al., 2011; González-Camacho et al., 2012; 2016; Pérez-Rodríguez, et al., 2012; Nascimento et al., 2013; Silva et al., 2014; 2017; González-Recio et al., 2014; Sant'anna et al., 2015; Gloria et al.,2016 e Akidemir et al., 2017). Espera-se com uso da abordagem de redes neurais captar relações genéticas importantes como os efeitos das interações intra e interalélicas e minimizar os efeitos ambientais que atuam como agente perturbador do processo de predição do verdadeiro valor genético.

As Redes Neurais Artificiais do tipo Rede de Base Radial (RNA-RBF) e Redes Perceptron de Multiplas Camadas (RNA-MLP) são exemplos de redes com camadas *feedforward* não-lineares. Ambas são aproximadores universais. Entretanto a RNA-MLP utiliza hiperplanos para particionar espaço de entradas (camada escondida e a Rede RNA-RBF utiliza hiper-elipsóides para particionar o espaço de entradas conforme pode ser visto na Figura 2 e um comparativo entre estas abordagens é apresentado, de forma resumida, na Tabela 1.

Tabela 1: Comparação resumida entre uma RNA-MPL e uma RNA-RBF.

RNA-MLP	RNA-RBF
• Neurônios das camadas intermediárias e da saída tem funções semelhantes. • Entrada da função de ativação é o produto interno da entrada de vetores e pesos. • Separa padrões de entrada com hiperplanos Contrói aproximadores globais para mapeamento entrada-saída • Pode ter uma ou mais camadas intermediárias.	• Neurônios das camadas intermediárias e da saída tem funções diferentes. • Entrada da função de ativação é a distância eucliadina entre os vetores da entrada e de pesos. • Separa padrões de entrada com hiperelipsóides • Contrói aproximadores locais para mapeamento entrada-saída • Geralmente possui apenas uma camada intermediária.

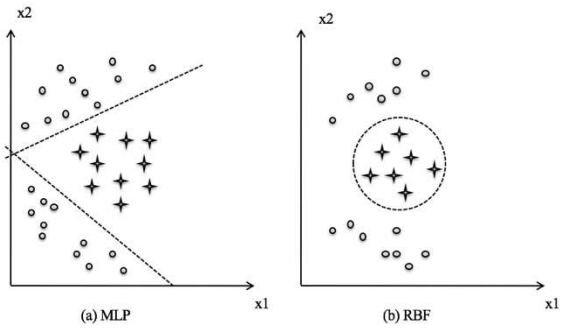


Figura 2. Particionamento dos dados de entrada por uma MLP (esquerda) e RBF (direita).

4.1. Perceptron Múltiplas Camadas- RNA-MLP

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

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

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

Na Figura 3 está ilustrada uma arquitetura de rede neural na qual podem ser identificadas as camadas de entrada, as camadas intermediárias e a camada de saída que deve retornar valores preditos para as variáveis de interesse, cuja resposta pode, de forma semelhante aos dados de entrada, ser uni ou multivariada. Essa Figura representa o primeiro modelo de redes de múltipla camada, o Perceptron multicamadas que surgiu e tornou as redes capazes de resolver problemas não linearmente separáveis.

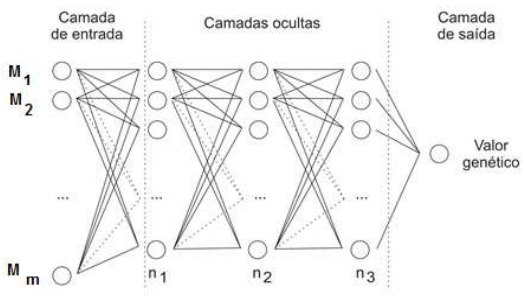


Figura 3. Representação dos três tipos de camadas existentes em redes neurais de múltiplas camadas, característica do modelo Perceptron Múltiplas Camadas.

A escolha da arquitetura utilizada pelas redes neurais, bem como os valores iniciais dos pesos utilizados no processo de treinamento ainda é feita de maneira empírica, e determina o tempo necessário para o treinamento (Silva et al., 2010). Normalmente, os valores iniciais dos pesos da rede são números aleatórios uniformemente distribuídos, em um intervalo definido. A escolha errada destes pesos pode levar a uma saturação prematura. Nguyen e Widrow (1990) encontraram uma função que pode ser utilizada para determinar valores iniciais melhores que valores puramente aleatórios.

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

A rede RNA-MLP possui neurônios com função não linear, possuem uma ou mais camadas ocultas ou intermediárias e possui alto grau de conectividade entre seus elementos processadores. Esta conectividade é definida pelos pesos sinápticos. As camadas intermediárias da rede são como detectores de características, as quais serão representadas através dos pesos sinápticos.

Na fase de treinamento da rede, o algoritmo adotado foi o algoritmo de *backpropagation* é realizado mediantes duas etapas principais: fase 1 – *forward* – e fase 2 – *backward.* Na fase 1, as entradas e seus respectivos pesos sinápticos são propagados, camada a camada ao longo da rede, até que a saída de rede seja produzida e comparada à saída desejada. Por fim, calcula-se o erro da rede. A fase 2 de retropropagação (*backward*), por sua vez, começa na camada de saída. O erro é propagado para as camadas anteriores, permitindo que os pesos sinápticos sejam recalculados de acordo com a regra Delta até que se retorne à primeira camada da rede.

$$\Delta w_{(t)} = \alpha \Delta w_{(t-1)} + \eta \delta_{(t)} y_{(t)}$$

Em que:

α: é a constante de *momentum* com $0 < \alpha < 1$;

δ: é o gradiente local;

η: taxa de aprendizagem;

y: a saída da rede.

Δw_t=Erro obtido pela rede neural na iteração t;

$\Delta w_{(t-1)}$= Erro obtido pela rede neural na iteração anterior $(t-1)$

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

O processo de aprendizagem de uma rede neural ocorre por meio do ajuste dos seus pesos sinápticos de acordo com a resposta da rede aos dados de entrada. O modo como é realizado este ajuste é que determina o tipo do aprendizado da rede que pode ser supervisionado quando a rede aprende utilizando exemplos fornecidos por um supervisor externo, ou não-supervisionado, quando utiliza apenas os dados de entrada (Haykin, 2001).

4. 1.2. Algoritmo da Retropropagação (“Backpropagation”)

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

O algoritmo de retropropagação se baseia no aprendizado supervisionado por correção de erros. Basicamente, a aprendizagem por retropropagação de erro consiste em dois passos através das diferentes camadas da rede: um passo para frente, Feed-forward (a propagação), e um passo para trás, *Feed-backward* (retropropagação) (Haykin, 2001). Primeiro, um padrão é apresentado à camada de entrada da rede. A resposta de uma unidade é propagada como entrada para as unidades na camada seguinte, até a camada de saída, onde é obtida a resposta da rede e o erro é calculado. No segundo passo, o erro é propagado a partir da camada de saída até a camada de entrada, e os pesos das conexões das unidades das camadas internas vão sendo modificados conforme o erro é retropropagado (Silva et al.,2010).

O erro de uma rede neural pode ser calculado como a diferença entre a saída gerada

pela rede e a saída desejada fornecida. Os erros são calculados sucessivamente até que sejam minimizados a um valor satisfatório, definido *a priori*. Sendo assim, pode ser visualizada uma curva de erros, a qual está diretamente relacionada à natureza do modelo de neurônio utilizado. Nem sempre é possível alcançar o menor valor de erro ou o mínimo global atingindo o que chamamos de mínimo local. Caso este erro alcançado seja desfavorável, é necessário recomeçar processo de aprendizado.

4.2. Redes Neural Função de Base Radial

As redes funções de base radial (RBF) podem ser utilizadas em problemas de aproximação de funções, predição e classificação de padrões (Silva et al., 2010). Uma função de ativação de base radial é caracterizada por apresentar uma resposta que decresce (ou cresce) em relação à distância a um ponto central.

A arquitetura das Rede RNA-RBF é do tipo *feedforward* e consiste de uma camada de entrada, que conecta a RNA ao seu ambiente (agrupa os dados de entrada em clusters), apenas uma camada oculta, que aplica uma transformação não-linear do espaço de entrada para um espaço oculto de alta dimensionalidade (geralmente são utilizadas funções de ativação de base radial gaussianas), e a camada de saída, que aplica uma transformação linear no espaço oculto fornecendo uma saída para a rede (Braga et al., 2007). Funções radiais representam uma classe especial de funções cujo valor diminui ou aumenta em relação à distância de um ponto central (Braga et al., 2007). Para o trabalho, utilizou-se topologia conforme apresentado na Figura 4.

.

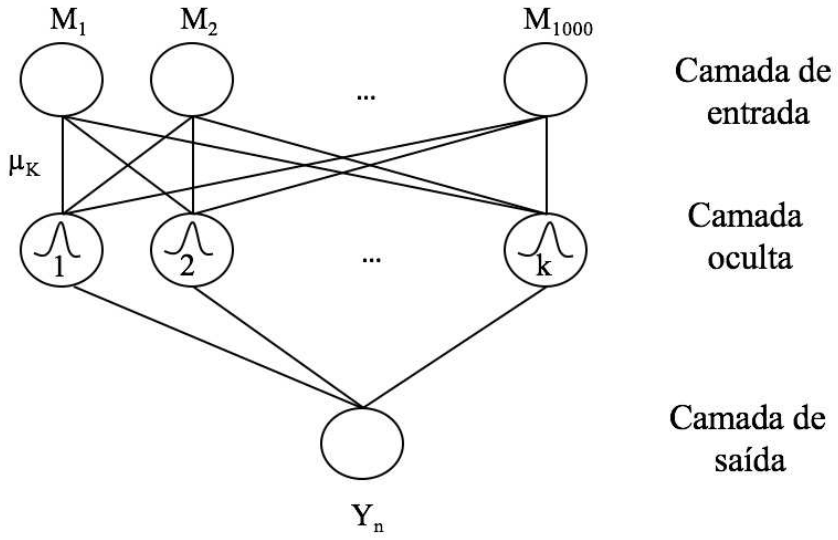


Figura 4. Arquitetura e topologia de uma Rede Funções de Base Radial com n, igual a n entradas, K neurônios na camada intermediária e uma saída (Y_n) .

A consequência imediata do uso de funções de ativação de base radial está na forma como as entradas são processadas pelos neurônios da camada intermediária. Ao invés da ativação interna de cada neurônio da camada intermediária se dar pelo emprego do produto escalar entre o vetor de entradas e o vetor de pesos, como no caso do Perceptron, ela é obtida a partir de uma norma ponderada da diferença entre ambos os vetores.

O treinamento das redes RNA-RBF é constituida por duas etapas: Na primeira etapa, ocorre os ajustes dos pesos dos neurônios na camada intermediária (por meio da alocação das funções de base radial) adotando-se um metodo de aprendizagem não-supervisionado, dependente apenas das características dos dados de entrada (Silva et al., 2010). A segunda etapa do treinamento ocorre de forma supervionada, em que ocorre os ajustes dos pesos na camada de saida utilizando a regra delta generalizada (Haykin et al, 2009). De forma resumida o processo de aprendizado da rede RNA-RBF desenvolvida transforma um problema de classificação não-linear em um problema linear, e é dividido nas seguintes etapas:

a) seleção dos centros **(*c*):** um subconjunto dos dados de treinamento é atribuído aos vetores centro das funções de base radial;

b) definição do raio de abrangência **(σ):** calcula-se a área de sensibilidade da função de base em relação ao seu centro.

c) cálculo da ativação dos neurônios ocultos (*u*): define-se o grau de ativação de cada neurônio da camada oculta utilizando distância euclidiana.

d) mapeamento do espaço não-linear (φ)**:** na camada oculta da rede, as funções gaussianas realizam a transformação dos dados de entrada não-lineares;

e) cálculo das saídas (*O*): os pesos de saída da rede são atualizados de acordo com a regra do *Perceptron simples* e utilizados na próxima iteração.

f) cálculo do erro (*e*): diferença entre a saída desejada (d) e a saída real da rede (o) referente à k-ésima observação na iteração t,

onde: $e_k(t) = d_k(t) - o_{k(t)}$

g) ajuste das sinapses (*m*): a atualização dos pesos sinápticos, associado ao i-ésimo neurônio, que ocorre somente quando o erro for diferente de zero.

$$m_{ki}(t+1) = m_{ki}(t) + \eta e_k \varphi_i(t)$$

h) condição de parada (*E*): o algoritmo atinge a convergência quando a rede não apresentar mudanças significantes nas sinapses.

A apresentação de todos os vetores de treinamento à rede define uma época de treinamento, nesta fase a condição de parada é testada e se não for satisfeita, o conjunto de treinamento é embaralhado e a rede continua seu processamento iterativamente.

5.0. REFERÊNCIAS BIBIOGRÁFICAS

Akdemir, D., Jannink, J.L. and Isidro-Sánchez, J., 2017. Locally epistatic models for genome-wide prediction and association by importance sampling. **Genetics Selection Evolution**, *49*(1), p.74.

Allard, R. W**. Princípios de melhoramento genético das plantas**. São Paulo: Edgard Blücher, 381p, 1971.

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

Bhattacharya, S. Heitmann, K. White, M.; Lukić, Z. Wagner, C. Habib, S.2010. Mass Function Predictions Beyond LCDM. **arXiv preprint arXiv:1005.2239**.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J.M., Pérez-Elizalde, S., Beyene, Y. and Dreisigacker, S., 2017. Genomic selection in plant breeding: Methods, models, and perspectives. **Trends in plant science.**

Cruz, C.D. Sant'anna, I.C. 2016. **Bioinformática e os avanços computacionais nas análises biométricas aplicadas ao melhoramento in** Desafios biométricos aplicados ao melhoramento genético.

Cruz, C. D. Carneiro, P. C. S. Regazzi, A. J. 2014. **Modelos biométricos aplicados ao melhoramento genético (volume 2).** 3ª Ed Viçosa: UFV. 668 p.

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

Cruz, C. D. 2012 **Princípios de Genética Quantitativa**. Editora UFV. Viçosa (MG). 2° reimpressão. 394p.

Cruz, C. D. **Programa Genes: Análise multivariada e simulação**. UFV. Viçosa- MG, 2006.

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

Dekkers, J.C.M; Hospital, F. **The use of molecular genetics in the improvement of agricultural populations**. Nature Reviews Genetics V. 3, p.22-32, 2002.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting quantitative traits with regression models for

dense molecular markers. **Genetics** 182, 375–385.

de los Campos, G., Gianola, D. and Rosa, G.J., 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. **Journal of Animal Science**, 87(6), pp.1883-1887.

Falconer, D.S**. Introdução à genética quantitativa**, Trad. Silva, M. A. e Silva, J.C. Viçosa, UFV. Imprensa Universitária. 279p., 1987.

Falconer, D.S., Mackay, T.F.C. **Introduction to quantitative genetics**. 4.ed. Edinburgh: Longman Group Limited, 464p., 1996.

Fernando, R. L., Habier, D., Stricker, C., Dekkers, J. C. M.; Tottir, L. R. **Genomic selection**. Acta Agriculturae Scandinavica, Section A - Animal Science. V. 57, n. 4, p. 192-195, 2007.

Fisher, R. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. **Transactions of the Royal Society of Edinburgh***, 52*(2), 399-433.

Galvão, C.O., Valença, M.J.S.,Vieira, V.P.P.B., Diniz, L. S.,Lacerda, E.G.M., Carvalho, A.C.P.L.F., Ludermir, T. B. **Sistemas inteligentes: Aplicações a recursos hídricos e ciências ambientais.** Porto Alegre: UFRGS/ABRH, 1999. 246p.

Gianola, D., Okut, H., Weigel, K.A. and Rosa, G.J., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC genetics**, 12(1), p.87.

Gianola, D., Van Kaam, J.B.C.H.M. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. **Genetics**, 178 (4), pp. 2289-2303.

Gianola, D. and De Los Campos, G., 2008. Inferring genetic values for quantitative traitsnon-parametrically. **Genetics Research,** 90(6), pp.525-540.

Gianola, D., R. Fernando and A. Stella, 2006. Genomic-assiste prediction of genetic value with semiparametric procedures. **Genetics** 173: 1761–1776.

Glória, L.S., Cruz, C.D., Vieira, R.A.M., Resende, M.D.V., Lopes, P.S., Siqueira, O.H.G.B.D., Silva, F.F. 2016. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science**, 191, p:91–96.

Goddard, M.E.; Hayes, B.J. 2007. **Genomic selection. Journal of Animal Breeding and Genetics**.V.124, n.6, p. 323–330.

González-Camacho, J.M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J.E., Mahuku, G., Babu, R. and Crossa, J., 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theoretical and Applied Genetics**, *125*(4), pp.759-771.

González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, L. and Gianola, D.,

2016. Genome-enabled prediction using probabilistic neural network classifiers. **BMC genomics**, *17*(1), p.208.

González-Recio, O., Rosa, G.J. and Gianola, D., 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science**, *166*, pp.217-231.

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

Holland J.B., 2006. Theoretical and biological foundations of plant breeding. In: Lamkey K.R., Lee M (eds) **Plant breeding:** the Arnel R Hallauer International Symposium. Blackwell Publishing, Ames.

Howard R., Carriquiry A. L., Beavis W. D. 2014.Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3: Genes, Genomes, Genetics,** 2014, 4(6): 1027–1046

Lee, S.H., van der Werf, J.H., Hayes, B.J., Goddard, M.E. and Visscher, P.M., 2008. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **PLoS genetics**, *4*(10), p.e1000231.

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A. and Avendano, S., 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of animal breeding and genetics**, 1246, pp.377-389.

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A. and Avendano, S., 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of animal breeding and genetics**, 124(6), pp.377-389.

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A., Kranis, A. and Gonzalez-Recio, O., 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genetics research**, 92(3), pp.209-225.

Long N, Gianola D, Rosa GJ, Weigel KA. 2011.Marker-assisted prediction of non-additive genetic values. **Genetica**. Jul 1;139(7):843-54.

Luidwig Junior, O.; Montgomery, E. **Redes Neurais – Fundamentos 45 e Aplicações com programas em C.** Editora Ciência Moderna, 2007, 125p.

Mackay, T.F., 2014. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. **Nature Reviews Genetics**, *15*(1), p.22.

Mackay, T.F.C. **The genetic architecture of quantitative traits**. Annual Review of Genetics. V.35, p.303-339, 2001.
Mcclelland, J. L. e Rumelhart, D. E. 1987. Learning the past tenses of english verbs: Implicit rules or parallel distributed processing. In B. MacWhinney (Ed.), **Mechanisms of Language Acquisition** (pp. 194-248). Mahwah, NJ: Erlbaum.

Meuwissen, T. H. E. **Genomic selection: marker assisted selection on genome-wide scale.** Journal of Animal Breeding and Genetics. V. 124, p. 321-322, 2007.

Meuwissen, T. H. E.; Hayes, B. J.; Goddard, M. E. **Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.** Genetics. 157: 1819-1829, 2001.

Mckinney, B. and Pajewski, N., 2012. Six degrees of epistasis: statistical network models for GWAS. **Frontiers in genetics**, *2*, p.109.

Muir, W.M. **Comparison of genomic and traditional BLUP-estimated breeding value accuracy and selection response under alternative trait and genomic parameters.** Journal of Animal Breeding and Genetics. V. 124, n. 6, p. 342–355, 2007.

Nascimento Filho, F.J. Do; Atroch, A.L., Sousa, N.R. De, Garcia, T.B., Cravo, M. Da S., Coutinho, E.F. 2001. Divergência genética entre clones de guaranazeiro. **Pesquisa Agropecuária Brasileira**, v.36, p.501-506.

Nguyen, D., and B. Widrow, 1989.The truck backer-uppcr: **An example of self-learning in neural networks.** Intenational Joint Conference on Neural Network.vol. I I , pp. 357-363, Washington, DC.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manès, Y. and Dreisigacker, S., 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3:* **Genes, Genomes, Genetics**, *2*(12), pp.1595-1605.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., Reich, D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. **Nature Genetics** v. 38, n. 8, p. 904- 909, 2006.

Resende, M. D. V. 2008.**Genômica quantitativa e seleção no melhoramento de plantas perenes e animais**. Colombo: Embrapa Florestas, p. 330.

Resende, M.D. V, Aguiar, A.M.,Abad, J.I.M., Missiaggia, A.A., Sansaloni, C., Petroli, C.; Grattapaglia, D., Resende Júnior, M.F.R**. 2010.Computação da Seleção Genômica Ampla (GWS)**. Colombo: Embrapa Florestas, p. 79.

Resende, M. D. V., F. F. e Silva, and C. F.2014. **Azevedo. Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência**. Editora Suprema, Viçosa.

Risch, N. J. 2000. Searching for genetic determinants in the new millennium. **Nature** 405, 847–856.

Rutkoski, J.E., Heffner, E.L. and Sorrells, M.E., 2011. Genomic selection for durable stem rust resistance in wheat. **Euphytica**, *179*(1), pp.161-173.

Schuster, I., Cruz, C. D. **Estatística genômica aplicada a populações derivadas de cruzamentos controlados**. 2ª Ed. Viçosa: UFV. 2008. p. 568.

Sant'anna, I.C., Tomaz, R.S., Silva, G.N., Nascimento, M.,Bhering, L.L., Cruz, C.D. Superiority of artificial neural networks for a genetic classification procedure. **Genetics**

and Molecular Research, v.14, p.9898-9906, 2015

Silva, G. N. **Redes Neurais Artificiais: novo paradigma para a predição de valores genéticos**. Viçosa: Universidade Federal de Viçosa, 2014. 106 p. (Dissertação - Mestrado em Estatística Aplicada e Biometria).

Silva, G. N. Tomaz, R. S. ; Sant'anna, I. C. Carneiro, V. Q. Cruz, C. D. Nascimento, M. 2016. Evaluation of the efficiency of artificial neural networks for genetic value prediction. Genetics and Molecular Research,v. 15.

Silva, I.N.; Spatti, D.H.; Flauzino, R.A. **Redes neurais artificiais para engenharia e ciências aplicadas- curso prático**. Editora: Artliber, p. 399, 2010.

Solberg, T. R., A. K. Sonesson, J. A. Woolliams, T. H. E. Meuwissen. 2009. Reducing dimensionality for prediction of genome-wide breeding values. **Genetics Selection Evolution**, 41:299.

Tanksley, S.D; Young, N.D; Paterson, A.H; Bonierbale, M.W. RFLP mapping in plant breeding: new tools for an old science. **Bio/Technology** 7:257–264. 1989.

Tafner, M., Xerez, M., E Rodrigues Filho, I.**Redes Neurais artificiais : introdução e princípios de neurocomputação**.Blumenau : EKO, 1995.

Tibshirani, R., 1996 Regression shrinkage and selection via the LASSO. J. R. **Stat. Soc. B** 58: 267–288.

Vencovsky, R. and Ramalho, M.A.P., 2000. Contribuição do melhoramento genético de plantas no Brasil. *Agricultura brasileira e pesquisa agropecuária. Brasília: EMBRAPA,* **Comunicação para Transferência de Tecnologia,** pp.57-98.

Yamamoto, A., Zwarts, L., Callaerts, P., Norga, K., Mackay, T.F.C. e Anholt, R.R.H., 2008. **Neurogenetic networks for startle-induced locomotion in Drosophila melanogaster.** Proceedings of the National Academy of Sciences of the USA 105, 12393–12398.

CAPITULO 1

Capitulo 1. Redes Neural Base Radial (RNA-RBF) na predição de valores genéticos

RESUMO

SANT'ANNA, Isabela de Castro, D.Sc., Universidade Federal de Viçosa, fevereiro de 2018, **Redes Neural Função de Base Radial (RNA-RBF) na predição de valores genéticos.** Orientador: Cosme Damião Cruz. Coorientadores: Matias Kirst, Marcos Deon Vilela de Resende e Moysés Nascimento.

Uma das principais contribuições da genética molecular em benefício do melhoramento de plantas é a possibilidade de utilização direta das informações de DNA na seleção. Todavia, para maior eficiência, nos estudos genéticos que envolvem os caracteres quantitativos regulados por vários genes com pequena magnitude de efeitos, as estratégias de seleção no melhoramento genético devem ser aprimoradas de forma a obter predições mais acuradas. A arquitetura dos caracteres complexos muitas vezes exige uma modelagem estatística avançada o qual enfrenta problemas em comportar as interações epistáticas que exigem modelos não lineares o que representa um novo desafio. Além dos problemas comumente causados pelo fato do número de marcadores utilizados superar, em muito, a quantidade de indivíduos na população de seleção. Desta forma, o objetivo deste trabalho é avaliar a eficiência da seleção genômica (SG) e das redes neurais artificias do tipo de base radial (RNA-RBF) na predição do valor genético em população natural com desequilíbrio gamético. Para isso, foi simulada uma população F_1 oriunda da hibridação de genitores divergentes, com 500 indivíduos, genotipados com 1000 marcadores do tipo SNP. Foram simulados dados genotípicos e fenotípicos adotando-se três modelos: aditivo, aditivo-dominante e epistático, atendendo duas situações de dominância: parcial e completa com caracteres quantitativos admitindo herdabilidades (h^2) de 30 e 60%, controlados cada um por 50 locos, considerando dois alelos por loco, totalizando 12 cenários distintos. Para avaliar a acurácia de predição, o modelo (RR-BLUP) e a rede de base radial (RNA-RBF) foram treinados utilizando 80% dos indivíduos da população e procedimento de validação cruzada com cinco repetições. O quadrado da correlação entre o valor genômico predito (GEBV) e valor genotípico/fenotípico foi utilizado para medir a acurácia seletiva (R^2) e o a raiz do erro do quadrado médio(REQM) acurácia preditiva. Por exemplo, para h^2= 0,3 no cenário aditivo, a validação R^2 foi de 31.4% para RNA-RBF, 58% para RR-BLUP e, no cenário de epistasia, os valores de R^2 foram 28.2% e 25.7%, respectivamente. Além disso, ao analisar o REQM, a diferença no desempenho das técnicas é ainda maior. Para o cenário aditivo, as estimativas foram 97.33 RR-BLUP e 5.80 para RNA-RBF, no cenário mais crítico, 329.70 RR-BLUP e 20.78 para RNA-RBF. Os resultados mostram que o uso de redes neurais permite capturar as

interações epistáticas levando a uma melhoria na acurácia da predição do valor genético e, principalmente, uma grande redução do erro quadrático médio que indica maior confiabilidade da predição do genômico valor.

ABSTRACT

SANT'ANNA, Isabela de Castro, D.Sc., Universidade Federal de Viçosa, February, 2018. **Radial Basis Function Neural Network (RNA-RBF) in the prediction of genetics values.** Adviser: Cosme Damião Cruz. Co-advisers: Marcos Deon Vilela de Resende, Matias Kirst and Moysés Nascimento.

One of the main contribution of molecular genetics for the benefit of plant breeding is the possibility of using directly DNA information in the selection of individuals. However, in order to obtain the most accurate genomic prediction in the genetic breeding of quantitative traits should consider that they can be affected by large numbers of genes, with a small magnitude of effects and usually can have nonlinear interactions between them. Modeling these interactions is extremely challenging because the number of parameters needed to accommodate epistatics interactions in a model with large numbers of markers and small number of individuals grows exponentially as more factors are considered. In this context, the neural networks have a great potential because they can capturing non-linear relationships between markers from the data themselves, which most of these models commonly used in the GS can not. The objective of this work is to evaluate the efficiency of genomic selection (GS) and radial base function network neural type (RBFNN) in the prediction of the genetic value in a natural population with linkage disequilibrium. For this, an F_1 population from the hybridization of divergent parents with 500 individuals genotyped with 1000 SNP-type markers was simulated. The phenotypic traits were determined by adopting three different gene action models: additive, additive-dominance and epistasis, attending two dominance situations: partial and complete with quantitative traits admitting heritability levels (h^2) ranging from 30 to 60%, each is controlled by 50 loci, considering two alleles per loco, totaling 12 different scenarios. To evaluate the comparison of models RR-BLUP and RBFNN a cross-validation procedure with five replicates were trained using 80% of the individuals of the population. The square of the correlation between predicted genomic value (GEBV) and genotype/phenotype value was used to measure predictive reliability(R^2) and the predictive mean-squared error root (MSER). For example, in h^2 = 0.3 in the additive scenario, the validation R^2 was 31% for RBFNN, 58% for RR-BLUP, and in the epistatics scenario R^2 values were 28% e 26%, respectively. Additionally, when analyzing the mean-squared error root the difference in performance of the techniques is even greater. For additive scenario, the estimates were 97.33 RR-BLUP and 5.80 for RBFNN, in the most critical scenario, 329.70 RR-BLUP and 20.78 for RBFNN. The

results show that the use of neural networks allows capturing the epistatics interactions leading to an improvement in the accuracy of the prediction of the genetic value and, mainly, a large reduction of the mean square error that indicates greater reliability of the prediction of the genomic value.

1. INTRODUÇÃO

Com o desenvolvimento dos marcadores moleculares e o avanço em técnicas de biologia molecular, criou-se a expectativa de que as informações genotípicas dos marcadores moleculares, uma vez correlacionados com características fenotípicas de interesse, pudessem ser amplamente utilizadas na seleção de indivíduos com maior valor genético (Crossa et al., 2017). Assim, muitos estudos têm sido feitos com o objetivo de auxiliar o melhoramento genético agregando estas informações moleculares. Dentre eles, destacam-se inicialmente abordagens sobre mapeamento genético, detecção de QTL (locos controladores de características quantitativas) e predições de ganhos por técnicas de seleção assistida por marcadores (SAM), que apresentavam como principalmente limitação a necessidade de se realizarem estudos específicos para cada família ou tipo de população e para a herdabilidade da característica de interesse (Crossa et al., 2017).

O avanço das tecnologias de genotipagem e o aparecimento dos marcadores do tipo SNP, que são abundantemente distribuídos nos genomas de muitas espécies, permitiu o surgimento da Seleção Genômica (SG) que foi proposta por Meuwissen et al. (2001) como uma forma de aumentar a eficiência seletiva, diminuir os custos via redução da frequência de fenotipagem e aumentar os ganhos anuais reduzindo o ciclo de geração (Rutkoski et al., 2010). A SG permite a estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma de forma que a maioria dos alelos de interesse esteja associado a esses marcadores (em desequilíbrio de ligação) para explicar grande parte da variação genética e predizer o valor genético dos indivíduos que ainda não tiveram seus fenótipos coletados (Resende et al., 2014).

O valor genético de indivíduos pode ser atribuído em suas porções relativas aos efeitos aditivos dos alelos, aos efeitos da dominância, que expressam as interações intra-alélicas, e aos efeitos epistáticos, que expressam as interações inter-alélicas. No entanto, a maioria das modelagens da SG usa apenas modelos de regressão que contemplam a porção aditiva do valor genético, o que dificulta, muitas vezes, uma representação mais realística da arquitetura genética dos caracteres quantitativos, sendo a inclusão de dominância e interações epistáticas fatores cruciais para aumentar a acurácia da predição (Lee et al., 2008; Akidemir et al., 2017).

A contribuição da dominância na herança de caracteres quantitativos de interesse é especialmente importante para espécies perenes e clones, com a possibilidade de

perpetuar todo o valor genotípico, e para espécies anuais onde há interesse comercial de híbridos e exploração da heterose (Almeida Filho et al., 2016). A inclusão desses efeitos em modelos de SG tem sido realizada por alguns autores como (Denis e Bouvet, 2011; Azevedo et al., 2015; Almeida Filho et al., 2016; Santos et al.,2016; Viana e Piepho, 2016; Viana et al., 2017), entretanto, muitos estudos ainda precisam ser feitos para permitir o melhor entendimento da contribuição dessas interações na predição dos valores fenotípicos.

Inúmeros estudos evidenciam que agregar múltiplas interações intra e interalélicas, ainda que em pequenos efeitos, é importante para explicar ao máximo a herdabilidade de caracteres quantitativos em estudos de associação em todo o genoma (McKinney e Pajewski 2012). Segundo Holland (2006), a epistasia têm importância inquestionável em características quantitativas como sugerido por numerosos estudos clássicos de genética quantitativa realizados em plantas e também em estudos de mapeamento de QTL como (Doebley et al.,1995; Lark et al., 1995; Wu et al., 1995; Fu e Ritland, 1996; Li et al., 1997; Pooni et al.,1987; Routman e Cheverud 1997; Yu et al., 1997) e nos demais trabalhos de (Wang et al., 1999; Viana, 2000; Viana, 2005; Dudley, 2008; Zhang et al., 2008; Dudley e Johnson, 2010; Denis e Bouvet, 2011;Viana, 2017.)

Entretanto, a inclusão de interações epistáticas é ainda mais difícil devido aos efeitos multiplicativos entre os alelos envolvidos na determinação do caráter, o que leva a uma superparametrização dos modelos agravando ainda mais problemas de processamento de dados, já que o número de marcadores sempre é muito superior ao número dos indivíduos (Gianola et al., 2006). Essas interações entre alelos podem ser exploradas pelos métodos de SG através de modelagens adequadas que incorporem a complexidade da arquitetura genética dos caracteres envolvidos. Entretanto, os métodos de SG são limitados pelos problemas de multicolinearidade e de dimensionalidade demandando mais recursos computacionais e uma modelagem mais apropriada para obtenção de predições acuradas.

Uma alternativa a esse problema são as metodologias de inteligência computacional em que a arquitetura das características é inferida dos próprios dados utilizados em seu treinamento e, dessa forma, não necessitam do conhecimento de distribuições a priori (como nos métodos bayesianos) e não há necessidade de atender pressuposições sobre as distribuições dos dados e dos resíduos. Além disso, as Redes Neurais Artificias (RNAs) podem capturar relações não lineares que a maioria dos modelos comumente utilizados na SG não consegue (Haykin, 2001; Gianola et al., 2006; Long et al., 2010 e Howards et al., 2014).

As redes neurais de função base radial (RNA-RBF) são uma classe particular de RNA que possui propriedades que a tornam atraentes para aplicações em SG como a capacidade de aprender a partir de dados utilizado em seu treinamento (Gianola et al., 2011), capacidade de aproximação universal (Park e Sandberg, 1991), apresentar solução única e a utilização de algoritmos rápidos que tornam a agilidade de uma RNA-RBF muito maior que uma RNA padrão (González-Camacho et al., 2012). Dessa forma, a RNA-RBF vem sendo utilizado com êxito no contexto de SG, no melhoramento genético por alguns autores como González-Camacho et al. (2012; 2016), Pérez-Rodríguez, et al. (2012) e Long et al. (2010; 2011).

Entretanto, muitos estudos com redes neurais ainda precisam ser realizados para entender melhor a complexidade dos caracteres quantitativos sendo que a RNA-RBF demonstra um grande potencial para capturar interações complexas ao reconhecer a importância dos efeitos da dominância e da epistasia nos dados de treinamento e incorpora-los nas predições futuras.

O objetivo do presente estudo é comparar a metodologia de seleção genômica (RR-BLUP) comumente aplicada a análise de rede de base radial e observar se a presença das interações epistáticas podem afetar o modelo e melhorar as estimativas de predição, comparar as metodologias quanto a eficiência na capturação de padrões complexos de interações não lineares.

2.MATERIAL E MÉTODOS

2.1. Estabelecimento de uma população com desequilíbrio de fase gamética

Dados genotípicos foram, originalmente, simulados para dez populações potenciais para uso como genitores em equilíbrio de Hardy-Weinberg com 2000 indivíduos em cada população. Foram geradas informações relativas a 1000 locos manifestando, em cada loco, dois alelos codominantes. Este conjunto prévio de dados foi utilizado para o cálculo de uma medida de dissimilaridade genotípica de Nei (1972) que foi posteriormente utilizada pelo método de agrupamento de projeção no plano bidimensional. Na Figura 1 é apresentado o resultado que evidencia que as populações 1 e 10 foram os mais divergentes e passam a ser chamados de P_1 e P_2 e empregadas para fins de cruzamentos e estabelecimento da população F1 com 500 indivíduos que supostamente teria maior concentração de desequilíbrio de fase gamética.

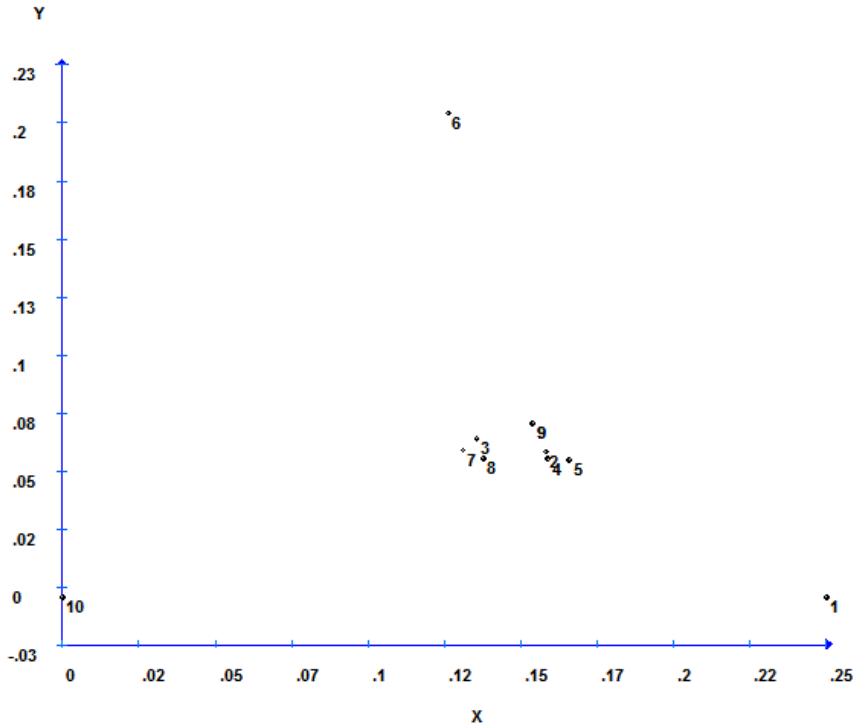


Figura 1- Projeção no plano bidimensional da matriz de dissimilaridade expresso pela Distancia de Nei, calculada a partir das informações genotípicas simuladas para seleção dos genitores.

2.2. Estabelecimento dos valores genotípicos

A população F1 foi então genotipada em relação a 1000 marcadores do tipo SNP. Para proceder às análises de seleção genômica, foram simulados valores genotípicos e fenotípicos considerando três níveis de dominância (0, 0.5 e 1), diferentes herdabilidades no sentido amplo (h^2 = 0.3 ou 0.6), e dois cenários considerando que o valor genotípico era estabelecido por um modelo aditivo (1) ou epistático (2).

As características foram estabelecidas pela ação de alelos de 50 locos, tomados ao acaso entre os 1000 previamente genotipados, com efeito aditivo diferencial e com pesos da importância do loco, sobre a variabilidade genotípica total do caráter, estabelecidos a partir de uma distribuição binomial $(p+q)^n$, em que p=q =0,5 e n= 49.

2.3. Estabelecimento dos valores fenotípicos

Os fenótipos dos indivíduos foram gerados segundo o modelo $F_i = G_i + E_i$, em que G_i é o efeito genético dado pelo somatório dos efeitos genéticos em cada loco e E_i o efeito ambiental, gerado segundo uma distribuição normal com média e variância compatível

com a herdabilidade do caráter simulado. Foram simulados doze cenários considerando que o valor genotípico era estabelecido por um modelo aditivo incluindo, ou não, interações epistáticas.

Tabela 1. Características avaliadas no estudo com seus respectivos valores de herdabilidade, modelo adotado e grau médio de dominância (Gmd).

Característica	Herdabilidade (%)	Modelo	Gmd
V1 - D0H30_Ad	30	aditivo	0
V2 - D0.5H30_Ado	30	aditivo-dominante	0.5
V3 – D1H30_Ad0	30	aditivo-dominante	1
V4 - D0H30_Ep	30	epistático	0
V5 – D0.5H30_Ep	30	epistático	0.5
V6 – D1H30_Ep	30	epistático	1
V7 - D0H60_Ad	60	aditivo	0
V8 - D0.5H60_Ado	60	aditivo-dominante	0.5
V9 - D1H60_Ado	60	aditivo-dominante	1
V10 – D0H60_Ep	60	epistático	0
V11 – D0.5H60_Ep	60	epistático	0.5
V12 - D1H60_Ep	60	epistático	1

Para o modelo aditivo o valor fenotípico do indivíduo i (F_i) foi estabelecido segundo o modelo: $F_i = G_i + E_i$, em que G_i é o efeito genético dado pelo somatório dos efeitos genéticos em cada loco e E_i o efeito ambiental, gerado segundo uma distribuição normal com média e variância compatível com a herdabilidade do caráter simulado.

O valor fenotípico total, no modelo aditivo, expresso por um determinado indivíduo i pertencente à população foi estimado a partir da expressão abaixo.

$$Y_i = \mu + \sum_{j=1}^{50} p_j \alpha_j + E_i$$

Em que:

i variava de 1, ..., 500 representando o número de indivíduos), $\alpha_j = a_i + d_i$ e d_i/a_i = gmd (grau médio de dominância) , sendo $\mu + a_j$, $\mu + d_j$ e $\mu - a_j$ os valores genotípicos associados as classes AA, Aa e aa, considerando igual a 1, 0 ou -1, respetivamente, quando codificadas, e p_jé a contribuição do loco j para a manifestação da característica considerada, no trabalho como tendo distribuição binomial. O valor de d_i foi estabelecido

a partir do grau médio da dominância manifestado em cada característica.

O valor fenotípico total, no modelo epistático, expresso por um determinado indivíduo pertencente à população foi estimado a partir da expressão abaixo. Em que o tipo de epistasia considerado representa a cadeia de uma via metabólica sendo permitida apenas interações no sentido único da cadeia de par em par .

$$Y_i = \mu + \sum_{j=1}^{50} p_j \alpha_j + \sum_{j=1}^{49} p_j \alpha_j \, \alpha_{j+1} + E_i$$

Em que:

$\alpha_j = a_i + d_i$ e d_i/a_i = gmd, sendo $\mu + a_j$, $\mu + d_j$ e $\mu - a_j$ os valores genotípicos associados as classes AA, Aa e aa,considerando igual a 1, 0 ou -1, respetivamente quando codificadas, e p_jé a contribuição do loco j para a manifestação da característica considerada, no trabalho como tendo distribuição binomial. O valor de d_i foi estabelecido a partir do grau médio da dominância manifestado em cada característica.

Nessa expressão, $Y_i = \mu + \sum_{j=1}^{50} p_j \alpha_j + \sum_{j=1}^{49} p_j \alpha_j \, \alpha_{j+1} + E_i$ o primeiro somatório da expressão se refere a contribuição dos locos individuais por meio de seus efeitos aditivos e dominantes e o segundo somatório representa os efeitos multiplicativos que correspondem as interações epistáticas entre pares de locos.

2.4. Estimação efeitos de marcas e predição do GEBVs (*Genomic estimated Breding Value*)

Para estimar os efeitos de marcas e os dos valores genômicos (GEBVs) foi utilizado a metodologia RR-BLUP conforme descrito por Meuwissen et al. (2001) onde:

$$y = Xb + Za + e,$$

em que *y* é o vetor de observações fenotípicas, *b* é o vetor de efeitos fixos, *a* é o vetor dos efeitos aleatórios dos marcadores e *e* refere-se ao vetor de erros aleatórios. *X* e *Z* são as matrizes de incidência para *b* e *a*. A estrutura de médias e variâncias no modelo em questão é definida como: $a \sim N(0,G)$, $E(y)=Xb$, $e \sim N(0,R=I)$, $Var(y)=V=ZGZ' + R$.

$G= I\frac{\sigma_g^2}{n}$

em que n é o número de marcadores dipostos no genoma, Z_{ij} é a linha da matriz de incidência que aloca o genótipo do i=ésimo marcador para cada indivíduos, 0, 1,-1 para

os genótipos AA, Aa, AA, respectivamente, para marcadores bialélicos e codominantes, e $\hat{\alpha}_i$ é o efeito estimado do i-ésimo marcador.

De posse dos efeitos de marcadores foram estimados os efeitos dos indivíduos (GEBVs) por meio do seguinte estimador:

$$G\hat{E}BVs = \hat{y}_j = \sum_i^n Z_{ij}\hat{a}_i$$

As equações de predição apresentadas acima foram modeladas assumindo a priori de que todos os locos explicam quantidades iguais da variação genética e portanto apresentam σ_g^2 comum. Assim, a variação genética explicada por cada loco é dada por (σ_g^2/n),em que σ_g^2 é a variação genética total e *n* é o número de marcadores utilizados (igual a 1000, neste estudo).

2.5. Predição do GEBV (*Genomic estimated Breeding Value*) por meio de Redes de Base Radial

A arquitetura das Redes RNA-RBF é do tipo *feedforward* e consiste de uma camada de entrada, que conecta a RNA ao seu ambiente (agrupa os dados de entrada em clusters), apenas uma camada oculta, que aplica uma transformação não-linear do espaço de entrada para um espaço oculto de alta dimensionalidade (geralmente são utilizadas funções de ativação de base radial gaussianas), e a camada de saída, que aplica uma transformação linear no espaço oculto fornecendo uma saída para a rede (Braga, 2007). Funções radiais representam uma classe especial de funções cujo valor diminui ou aumenta em relação à distância de um ponto central (Braga, Carvalho, Ludermir, 2007). Para o trabalho, utilizou-se topologia conforme apresentado na Figura 2.

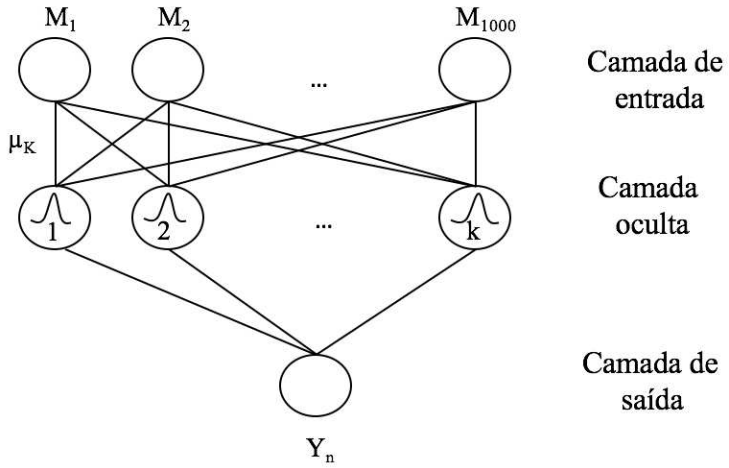


Figura 2. Arquitetura e topologia de uma Rede Funções de Base Radial com n, igual a 1000, entradas, K neurônios na camada intermediária (variando de 1 a 200) e uma saída (Y_n) que variava de 1 a 400 no processo de treinamento e de 1 a 100 no processo de validação.

O treinamento das redes RNA-RBF é constituida por duas etapas: Na primeira etapa, ocorre os ajustes dos pesos dos neurônios na camada intermediária (por meio da alocação das funções de base radial) adotando-se um metodo de aprendizagem não-supervisionado, dependente apenas das características dos dados de entrada (Silva et al., 2010). A segunda etapa do treinamento ocorre de forma supervionada, em que ocorre os ajustes dos pesos na camada de saida utilizando a regra delta generalizada (Haykin et al, 2009). De forma resumida o processo de aprendizado da rede RNA-RBF desenvolvida transforma um problema de classificação não-linear em um problema linear, e é dividido nas seguintes etapas:

a) seleção dos centros **(*c*):** um subconjunto dos dados de treinamento é atribuído aos vetores centro das funções de base radial;

b) definição do raio de abrangência **(σ):** calcula-se a área de sensibilidade da função de base em relação ao seu centro. Neste trabalho foi utilizado valores do raio entre 5 e 100, sendo o raio médio utilizado de 25. O valor do raio foi estipulado com base em tentativas em que os intervalos selecionados proporcionava os menores valores de erro e maiores valores de R^2, sendo repetido sempre que o limite superior ou inferior proporcionasse o melhor valor.

c) cálculo da ativação dos neurônios ocultos (*u*): define-se o grau de ativação de cada neurônio da camada oculta utilizando distância euclidiana.

d) mapeamento do espaço não-linear (φ)**:** na camada oculta da rede, as funções gaussianas realizam a transformação dos dados de entrada não-lineares;

e) cálculo das saídas (*O*): os pesos de saída da rede são atualizados de acordo com a regra do *Perceptron simples* e utilizados na próxima iteração.

f) cálculo do erro (*e*): diferença entre a saída desejada (d) e a saída real da rede (o) referente à k-ésima observação na iteração t,

onde: $e_k(t) = d_k(t) - o_{k(t)}$

g) ajuste das sinapses (*m*): a atualização dos pesos sinápticos, associado ao i-ésimo neurônio, que ocorre somente quando o erro for diferente de zero.

$$m_{ki}(t+1) = m_{ki}(t) + \eta e_k \varphi_i(t)$$

h) condição de parada (*E*): o algoritmo atinge a convergência quando a rede não apresentar mudanças significantes nas sinapses.

A apresentação de todos os vetores de treinamento à rede define uma época de treinamento, nesta fase a condição de parada é testada e se não for satisfeita, o conjunto de treinamento é embaralhado e a rede continua seu processamento iterativamente.

2.6 Validação

Validação Genotípica

Nesse tipo de validação, o treinamento é realizado com as informações fenotípicas de todos os indivíduos mensurados e utilizados na estimação dos valores GEBVs, mas utiliza na validação informações do mesmos indivíduos porém relacionando com seus valores genotípicos ou valores reais. Esse tipo de validação só e possível em estudos realizados com dados simulados e indica se o efeito perturbador ambiental na fase de treinamento compromete a identificação de genótipos superiores na fase de validação.

Validação cruzada

A validação cruzada foi realizada pela reamostragem de um grupo de indivíduos via procedimento *k-fold*. (Bengio e Grandvalet, 2004). A metodologia do *k-fold* baseia-se na divisão do conjunto de C dados amostrais em g grupos de tamanho igual a k, de forma que C = gk. Em cada um dos g grupos, k indivíduos são retirados para a formação da população de validação. No caso, tomou-se k=100 e, portanto, a população de validação usou todos os C=500 indivíduos, e a população de estimação, 400 indivíduos e 5 repetições. Em cada repetição, cem indivíduos foram removidos da população e utilizados para a formação da população de validação, e os outros indivíduos restantes foram utilizados na população de estimação dos efeitos dos marcadores. Este tipo de procedimento de validação foi executado por meio do software R em interface com o programa genes (Cruz, 2016).

2.7. Comparação entre os modelos avaliados

Para avaliar as eficiências do modelos utilizados RR-BLUP e RNA-RBF foram utilizados as seguintes estatísticas: capacidade preditiva do modelo (CP), confiabilidade

ou acurácia seletiva do modelo, raiz do erro quadrático médio ($REQM$) e o o viés (β).

a) Capacidade Preditiva

A acurácia da SG depende da proporção da variação genética genômica (VGG) do indivíduo explicada pelos marcadores (Resende et al., 2014). Quando a população de validação é independente da população utilizada para predição de efeitos dos marcadores, a correlação entre o efeito estimado dos marcadores e o valor fenotípico dos indivíduos é predominantemente de natureza genética (Resende et al., 2014). Nesse caso, a correlação passa a ser definida como sendo a capacidade preditiva da SG e é dada pela equação abaixo:

$$CP = cor(\ y \text{ e } y_{p)}$$

Sendo y e y_p os valores observados e os valores predito respectivamente.

b) Acurácia e confiabilidade seletiva

Para avaliar a eficiência da abordagem SG quantificou-se a acurácia que é conceituada como a correlação entre o valor genético verdadeiro e aquele estimado a partir das informações genotípicas (marcadores) e/ou fenotípica dos indivíduos. E o seu quadrado, também conhecido como confiabilidade, é uma medida do quadrado da correlação entre o valor estimado e os valores verdadeiros, ou seja, mede o quanto a estimativa obtida é relacionada com o valor real do parâmetro.

Na análises realizadas foi obtido o quadrado da correlação entre o valor predito e o valor utilizado no aprendizado supervisionado (valor fenotípico ou genotípico, dependendo da estratégia da análise) em analogia ao quadrado da correlação entre a média fenotípica e o valor fenotípico verdadeiro, que expressa a herdabilidade da característica.

Se y representa o valor fenotípico, dado por y = g + ε com cov(g,ε)=0, e se o valor de rede é obtido mantendo a propriedade cov(y_p,ε)=0 então podemos afirmar que:

$Cov(y_p,y) = Cov(y_p,g+ \varepsilon) = Cov(y_p,g)$

Logo

$$Cor(y_p,y) = \frac{Cov(y_p,y)}{\sigma_{yp}\sigma_y} = \frac{Cov(y_p,g)}{\sigma_{yp}\sigma_y}\frac{\sigma_g}{\sigma_g}$$

Ou

$r_{yp,y} = r_{yp,g}\frac{\sigma_g}{\sigma_y} = r_{yp,g}h$, sendo $h = \frac{\sigma_g}{\sigma_y}$

c) Acurácia preditiva

Segundo Keng et al. (2017) menores valores de erro quadrático médio indicam a acurácia e precisão de um estimador. Dessa forma, outra estatística usada na avaliação da

eficiência da SG foi a acurácia preditiva medida pela diferença quadrática entre valores verdadeiros e estimados, ou seja, apresenta pelo mínimo erro quadrático médio (REQM) de predição, dado por:

$$\text{REQM} = \frac{\sqrt{\sum(y_p - y)^2}}{n}$$

2.8. Recursos Computacionais

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

Para a etapa de simulação das populações e análises de diversidade genética das mesmas, o software GENES foi utilizado (Cruz, 2016).

Para avaliar as metodologia de RR-BLUP proposta, o software Genes também foi utilizado, porém no módulo de integração com o software R (R Core Team, 2017).

Para avaliar as metodologia de RNA-RBF proposta, o software Genes também foi utilizado, porém no módulo de integração com o software Matlab (Matlab, 2011).

3. RESULTADOS E DISCUSSÃO

3.1. Predição de Valores Genotípicos por meio de Redes Neurais do Tipo Base Radial e Seleção Genômica – Análise com validação genotípica

Nas tabelas a seguir serão apresentados os valores de acurácia de predição (REQM), acurácia seletiva (R^2), capacidade preditiva (CP), obtidos pela metodologia RR-BLUP e de Rede de Base Radial para dados genotípicos e fenotípicos obtidos em cenários aditivo, dominância e epistáticos.

a. O modelo genético aditivo

Na Tabela 2 são apresentados os valores de acurácia de predição (REQM), acurácia seletiva (R^2), capacidade preditiva, viés obtidos pela metodologia obtidos pelas pela abordagem de RNA-RBF para dados genotípicos. Para interpretação dos resultados obtidos, deve ser ressaltado que este estudo visa comparar o desempenho das técnicas RR-BLUP e RNA-RBF, fundamentada em inteligência computacional, para fins de predição de valores genéticos. Entretanto, deve-se ter em mente que, em análise comparativa de resultados, para as características que não incluíram efeito da dominância

em seu controle gênico deveria, a princípio, indicar uma igualdade entre a técnica RR-BLUP e a RNA-RBF ou até mesmo uma certa superioridade do RR-BLUP. Tal fato, decorre de se considerar, no modelo RR-BLUP, apenas os efeitos de dose alélica de cada marcador de forma que, ao ser aplicado para estudo de característica sem efeito de dominância, conseguiria captar toda a contribuição aditiva dos locos e, nesta situação, a redução na eficácia estaria apenas comprometida pelos efeitos perturbadores do ambiente refletido sobre a variância residual do modelo. Nesta condição, seria esperado que as técnicas de inteligência computacional RNA-RBF poderiam proporcionar resultados apenas equivalentes, na medida que os neurônios utilizados teriam o papel de detectar unicamente as influências das entradas (marcadores) sabidamente lineares. A menor eficiência das técnicas de inteligência computacional, nestas condições, seria atribuída ao processo iterativo, aos critérios de parada, à convergência pelo uso algoritmo de treinamento que poderia estacionar a solução em pontos locais ao invés de globais, à especificação do tamanho do raio de abrangências, dentre outros.

Pode ser verificado na Tabela 1 que houve bom desempenho, em termos de predição de valores genotípicos, nas situações em que se utilizou cenários representado alta ou baixa herdabilidade. Os resultados do R^2 obtidos por meio de treinamento em dados fenotípicos e a validação em dados genotípicos mostram que o modelo linear utilizado foi capaz de captar efeitos de marcas que recuperem o valor genético simulado em 58% para o cenário de baixa herdabilidade e 70% para o de alta. Para a metodologia de RNA-RBF as medidas de confiabilidade retratam a recuperação de 31 e 65% do valor genético simulado. Já os resultados de REQM indicam um menor erro do processo preditivo para os cenários de alta herdabilidade e em geral são menores para a predição realizada pela RNA-RBF. Através de uma análise do REQM é possível observar a importância do efeito ambiental como um agente pertubador já que o REQM é maior quando o ruído ambiental presente no treinamento também é maior h^2=30.

Tabela 1. Medidas da eficiência de predição de valores genotípicos obtidas pela abordagem RR-BLUP e RNA-RBF para características com controle gênico aditivo.

Treinamento em dados fenotípicos e validação em dados genotípicos				
Modelos	Cenários	R^2	CP	REQM
RR-BLUP	d0h30	0,58	0,76	91,85
	d0h60	0,77	0,88	81,62
RNA-RBF	d0h30	0,31	0,56	5,24
	d0h60	0,65	0,81	3,82
Treinamento e validação cruzada em dados fenotípicos				
Modelos	Cenários	R^2	CP	REQM
RR-BLUP	d0h30	0,10 ± 0,02	0,31 ± 0,03	97,35 ± 2,14
	d0h60	0,36 ± 0,07	0,60 ± 0,07	85,76 ± 2,92
RNA-RBF	d0h30	0,11 ± 0,03	0,33 ± 0,04	5,89 ± 0,12
	d0h60	0,38 ± 0,08	0,62 ± 0,07	4,52 ± 0,12

Para características predominantemente determinada por genes de ação aditiva os modelos de SG são comprovadamente eficientes (Howards et al., 2014). Nesse aspecto o resultado obtido foi também observado na literatura por outros autores que compararam modelos lineares e paramétricos como é o caso de Gianola et al. (2006), Howards et al. (2014) e Gianola et al. (2015).

A correlação entre o valor fenotípico real simulado e o valor fenotípico predito só é possível em dados simulados e é considerada por autores, como Long et al. (2010), como a medida que melhor descreve a acurácia já que que o ambiente não é um fator perturbador. Dessa forma, ao compararmos os resultados obtidos pelo treinamento e validação em dados fenotípicos (Tabela 2) a acurácia seletiva pelas duas abordagens se tornaram muito inferior a herdabilidade da característica indicando claramente a influência do ambiente como o maior fator perturbador. Nos cenários de baixa herdabilidade o RR-BLUP é inferior a RNA-RBF para todos os parâmetros observados indicando que a rede é capaz de lidar melhor com o ruído ambiental observado. Howard et al. (2014) compararam métodos paramétricos e não paramétricos, relataram a influência ambiental, ou seja o menor controle genotípico sobre o fenótipo, como o principal problema preditivo para ambos as abordagens. Entretanto, assim como no presente trabalho, Long et al. (2010), usando uma simulação de genoma completo com 2000 marcas numa população de acasalamento ao acaso de touros e novilhas, também

detectaram a RNA-RBF como o melhor método preditivo mesmo em cenários puramente aditivos.

b. O modelo aditivo-dominante

Na Tabela 2 é possível observar que a RNA-RBF e o RR-BLUP levam a resultados semelhantes no processo preditivo quando analisamos os valores de R^2 em cenários aditivo dominantes de alta e baixa herdabilidade.

Em estudos comparativos de resultados envolvendo características com controle gênico influenciado pela dominância devemos considerar as particularidades das técnicas empregadas para fins de entendimento de possíveis resultados. Assim, quando a característica é determinada por efeitos aditivos e atribuídos à dominância pode-se questionar se técnicas como RR-BLUP teria sua eficácia de predição preservada mesmo quando se utiliza modelos em que os regressores representam apenas o efeito de dose alélica, não contemplando interações intra e nem inter-alélica no modelo estatístico. Entretanto, deve-se considerar que no RR-BLUP o efeito estimado representa o efeito de substituição gênica que, segundo teoria de genética quantitativa, é determinado tanto pelo valor genotípico do homozigoto (a) quanto do heterozigoto (d) e, assim, a presença da dominância no controle da característica mesmo com a ausência de uma matriz de incidência da dominância no modelo estatístico não representaria grande problema. Por outro lado, tem-se a expectativa de que a técnica RNA-RBF sofreria menos efeito da influência da dominância sobre a característica pois recorreria aos valores agregados em seus neurônios, em sua camada oculta, para maximizar a acurácia de predição.

Resultado semelhantes ao relatado na Tabela 2 também foi encontrado no trabalho de Long et al. (2010) usando uma simulação de genoma completo com 2000 marcas, numa população de acasalamento ao acaso de touros e novilhas que em cenário aditivo dominante, informando a ocorrência de resultados semelhantes para o método de seleção genômic Bayes A e a RNA-RBF. Entretanto, ao analisarmos os valores do REQM, tem-se para o RR-BLUP que, quanto maior a influencia da dominância, maior o valor do REQM obtido. Entretanto, para a abordagem utilizando RNA-RBF a presença de indivíduos heterozigotos incrementa o treinamento de forma que o REQM da RNA-RBF apresenta os menores valores tanto em cenários de alta ou baixa herdabilidade para validação fenotípica ou genotípica.

Tabela 2. Medidas da eficiência de predição de valores genotípicos obtidas pela abordagem RR- BLUP e RNA-RBF para características com controle gênico aditivo-dominante.

Treinamento em dados fenotípicos e validação em dados genotípicos				
Modelos	Cenários	R^2	CP	REQM
RR-BLUP	d0.5h30	0,54	0,73	118,41
	d1h30	**0,08**	0,29	157,17
	d0.5h60	0,71	0,84	108,54
	d1h60	0,63	0,79	134,40
RNA-RBF	d0.5h30	0,31	0,55	5,47
	d1h30	0,25	0,50	14.55
	d0.5h60	0,60	0,77	4,03
	d1h60	0,61	0,77	4,27
Treinamento e validação cruzada em dados fenotípicos				
		R^2Val.	CP	REQM
RR-BLUP	d0.5h30	0,12 ± 0,07	0,34 ± 0,08	122,65 ± 3,02
	d1h30	0,00 ± 0,01	-0,04 ± 0,04	267,59 ± 21,72
	d0.5h60	0,30 ± 0,07	0,54 ± 0,07	111,50 ± 3,87
	d1h60	0,19 ± 0,05	0,44 ± 0,06	137,09 ± 3,85
RNA-RBF	d0.5h30	0,12 ± 0,06	0,34 ± 0,07	6,17 ± 0,14
	d1h30	0,02 ± 0,01	0,04 ± 0,13	16,78 ± 1,19
	d0.5h60	0,34 ± 0,07	0,58 ± 0,06	4,67 ± 0,19
	d1h60	0,18 ± 0,04	0,42 ± 0,05	5,40 ± 0,16

Embora a dominância não tenha sido contemplada no modelo linear utilizado, segundo alguns autores Viana et al. (2006), Denis et al.(2011), Almeida-Filho et al.(2016), Santos et al.(2016), sua incorporação não leva a uma melhora da acurácia do processo preditivo de características complexas condicionadas por muitos genes e além disso, segundo Toro e Varona (2010), os modelos de aditivo-dominante superaram os modelos aditivos apenas na primeira geração preditiva no trabalho realizado com caracteres poligênicos em população simulada. Dessa forma, para dados genotípicos a correlação entres valores fenotípicos e genotípicos é mais afetada pelo ambiente, ou seja, tem valores menores nos cenários de baixa herdabilidade e não é tão afetada nos cenários em que a dominância aumenta.

As correlações obtidas entre os valores fenotípicos observados e os valores genômicos preditos demonstram que a RNA-RBF desempenho igual ou superior ao RR-BLUP. Resultado semelhante foi encontrado por Long et al. (2010) usando uma simulação de genoma completo com 2000 marcas numa população de acasalamento ao acaso de touros e novilhas que em cenário aditivo dominante também encontrou

resultados semelhantes para o métodos Bayes A e a RNA-RBF. Ao analisarmos o REQM (Tabela 3), a RNA-RBF apresenta os menores valores tanto em cenários de alta ou baixa herdabilidade para validação fenotípica ou genotípica.

c. O modelo epistático

Os resultados obtidos na Tabela 3 retratam a superioridade das RNA-RBF em cenários com alta influencia epistática na presença ou não de dominância tanto pela análise do R^2 quanto pela análise do REQM.

No estudo de características complexas em que o valor genotípico final é o resultado da ação individual e interativa entre os alelos deve-se esperar baixa eficiência de modelos estatísticos que contemplem apenas efeitos aditivos dos regressores. O atenuante para este modelo é admitir que tais efeitos multiplicativos, apesar de serem relevantes, possam ser comparativamente menos importantes que os efeitos aditivos e o ajuste do modelo, mesmo que distante do ótimo, possa ser útil no melhoramento genético. A inclusão de efeitos multiplicativos em abordagem como RR-BLUP poderia demandar modelagem de alta complexidade e uso de recurso computacional mais sofisticado para processamento de dados. Assim, a opção de uso do RR-BLUP apenas com matriz de incidência contemplando efeito de dose alélica pode ser uma simplificação indesejável se, de fato, efeitos multiplicativos assumirem importância de maior magnitude. Diante deste cenário, tem-se expectativa de melhor desempenho da RNA-RBF por contar com informações de camadas ocultas que acrescentariam valiosas informações no processo de predição. Abordagem por RNA-RBF que fazem uso de função de ativação gaussiana deve ter boa aderência aos dados de características complexas, em especial com as particularidades utilizadas neste estudo em que o erro tem distribuição normal e os efeitos dos marcadores tem importância definida por uma distribuição binomial simétrica $(p+q)^n$, sendo p=q=0.5 e n relativamente grande.

Resultados semelhantes ao descritos na Tabela 4 tem sido encontrados pela maioria dos autores que utilizam modelos não paramétricos como é o caso de Gianola et al. (2006), Long et al. (2010), Gianola et al. (2011), Long et al. (2011), González-Camacho et al. (2012), Pérez-Rodríguez et al. (2012), Howards et al. (2014) e González-Camacho et al. (2016). Essa superioridade dos modelos não paramétricos se deve ao aprendizado baseado nos dados que não é previamente modelado de forma linear. Assim, a ineficiência do RR-BLUP se deve a modelagem linear realizada o que também já era esperado, entretanto a constatação de que a tanto a epistasia, quanto a dominância afetam o processo preditivo indica uma necessidade de incorporar essas novas metodologias com

o objetivo de aprimorar o processo preditivo contrariando muitos autores como Hill et al. (2008; 2010) e Crow (2010), que acreditavam que, grande parta da epistasia, é de natureza aditiva x aditiva e por isso majoritariamente seria acomodada na própria variância aditiva.

Tabela 3. Medidas da eficiência de predição de valores genotípicos obtidas pela abordagem RR- BLUP e RNARBF para características com controle epistático

Treinamento em dados fenotípicos e validação em dados genotípicos				
Baixa herdabilidade				
Modelos	Cenários	R^2	CP	REQM
RR-BLUP	d0h30e	0,13	0,37	155,66
	d0.5h30e	0,20	0,44	221,14
	d1h30e	0,26	0,51	329,69
RNARBF	d0h30e	0,32	0,56	14,48
	d0.5h30e	0,27	0,52	16,58
	d1h30e	0,28	0,53	20,78
Alta Herdabilidade				
		R^2	CP	REQM
RR-BLUP	d0h60e	0,31	0,56	148,64
	d0.5h60e	0,36	0,60	209,79
	d1h60e	0,46	0,68	335,00
RNA-RBF	d0h60e	0,56	0,75	10,74
	d0.5h60e	0,56	0,75	12,39
	d1h60e	0,56	0,75	15,31
Treinamento em dados e validação cruzada em dados fenotípicos				
Baixa Herdabilidade				
		R^2	CP	REQM
RR-BLUP	d0h30e	0,01 ± 0.02	0,07 ± 0,09	278,61 ± 23,91
	d0.5h30e	0,02 ± 0.06	0,15 ± 0,06	366,06 ± 19,67
	d1h30e	0,06 ± 0.00	0,20 ± 0,14	575,41 ± 23,59
RNA-RBF	d0h30e	0,03 ± 0,02	0,14 ± 0,07	16,85 ± 1,17
	d0.5h30e	0,05 ± 0,02	0,22 ± 0,05	18,51 ± 0,54
	d1h30e	0,06 ± 0,05	0,18 ± 0,19	23,84 ± 1,65
Alta Herdabilidade				
Modelos	Cenários	R^2	CP	REQM
RR-BLUP	d0h60e	0,03 ± 0,02	0,16 ± 0,05	197.378 ± 12.08
	d0.5h60e	0,08 ± 0,04	0,28 ± 0,07	274.193 ± 21.94
	d1h60e	0,13 ± 0,05	0,35 ± 0,07	434.492 ± 26.43
RNA-RBF	d0h60e	0,06 ± 0,03	0,24 ± 0,06	13.55 ± 0.30
	d0.5h60e	0,10 ± 0,02	0,32 ± 0,03	15.48 ± 0.46
	d1h60e	0,13 ± 0,03	0,36 ± 0,04	18.83 ± 0.75

Os resultados obtidos por validação fenotípica (Tabela 4) indicam que o efeito ambiental é o principal agente perturbador do processo preditivo. Ainda assim, tanto a

dominância quanto a epistasia, quando presente, impacta os resultados da SG pela redução da acurácia seletiva (redução nos valores de R^2) e da acurácia de predição (aumento nos valores de REQM). Em cenários epistáticos os valores obtidos pelas duas metodologias se tornam ainda mais discrepantes, sendo maior com aumento da dominância o que pode ser entendido devido ao fato das interações epistáticas terem interações do tipo aditiva x aditiva, aditiva x dominante e também dominante x dominante e que a rede de base radial pode ser capaz de capturar esse tipo de comportamento nos dados de treinamento. A analise do REQM obtidos por meio de validação fenotípica indicam uma boa acurácia preditiva já que o REQM máximo observado para ambas as herdabilidades foi 20. Entretanto os valores de R^2 obtidos nesse tipo de validação são muito inferiores a herdabilidade paramétrica dos cenários simulados e retratam uma acurácia seletiva semelhante para qualquer uma das metodologias testadas.

Segundo González-Camacho et al. (2012), a baixa performance da RNA-RBF pode ser causada pela presença de muitos marcadores sem importância para a determinação caráter em estudo indicando mais uma vez a complexidade enfrentada pelo pesquisador ao lidar com características complexas com forte influência de efeitos ambientais e efeitos não lineares em sua constituição genética.

Long et al. (2010) realizaram estudos baseados na simulação de genoma completo com 2000 marcas numa população de acasalamento ao acaso de touros e novilhas em três cenários: aditivo, dominante e epistático. Neste trabalho dois modelos de RNA-RBF foram utilizados sendo que, no primeiro, havia pesos específicos para cada SNP e, no segundo, todos os SNPs apresentavam a mesma importância. Na maioria dos casos, o modelo com pesos específicos foi melhor do que o com peso comum para cada SNP.

Neste trabalho foram utilizados, sem prejuízo na extrapolação ou generalização, a análise de apenas 1000 marcadores moleculares. Este número pode ser considerado relativamente baixo tendo em vista a possibilidade da genotipagem em alta escala em trabalhos rotineiros do melhoramento genético. Entretanto, mesmo com este número reduzido já é possível perceber a alta demanda computacional em especial pela técnica biométrica de inteligência computacional, como a RNA-RBF. Assim, uma forma de amenizar os problemas enfrentados por pesquisadores diante das problemáticas computacionais e estatísticos, gerada pela utilização de grande número de marcadores em todo o genoma e um número de indivíduos limitado, são os métodos de seleção de variáveis. Tais métodos permitirão análise mais eficiente, em tempo e demanda computacional, e de uma forma mais apropriada para resolver o problema da dimensionalidade e multicolinearidade para que os modelos a serem testados tenha

melhores condições de acomodar ou amenizar os efeitos perturbadores tais como o ambiente, a dominância e a epistasia que incidem no processo de predição do fenótipo desejado.

4.0. CONCLUSÃO

- Os métodos de RNA-RBF possuem capacidade de predição similar ou melhor, quando a ação gênica era puramente aditiva

- Na presença de uma complexa relação genótipo-fenótipo (isto é, não linearidade e não aditividade), os modelos RNA-RBF superaram um modelo aditivo linear, RR-BLUP, na predição de valores genéticos totais de caracteres quantitativos usando marcadores SNP.

- Ao lidar com um número grande número de marcadores, a demanda computacional no RNA-RBF é intensiva sugerindo a utilização de uma seleção de variáveis como um solução interessante para melhoria do processo preditivo a ser abordados.

5.0.REFERÊNCIAS

Almeida Filho, J.E., Guimarães, J.F.R., e Silva, F.F., de Resende, M.D.V., Muñoz, P., Kirst, M. and Resende Jr, M.F.R., 2016. The contribution of dominance to phenotype prediction in a pine breeding and simulated population. **Heredity**, *117*(1), p.33.

Akaike H., 1974. A New Look at the Statistical Model Identification. **IEEE Transactions on Automatic Control** ,19, p.716–723.

Azevedo, C.F., De Resende, M.D.V., E Silva, F.F., Viana, J.M.S., Valente, M.S.F., Resende, M.F.R. And Muñoz, P., 2015. Ridge, Lasso and Bayesian additive-dominance genomic models. **BMC genetics,** *16*(1), p.105.

Bernardo, R., 2002. **Breeding for quantitative traits in plants** (No. 576.5 B523). Stemma Press.

Braga, A., Carvalho, A., Ludermir, T. ,2000. **Redes Neurais Artificiais: Teoria e Aplicações**, Livro Técnico e Científico, Rio De Janeiro. pp. 05-55.

Chen, S., Cowan, C.F. and Grant, P.M., 1991. Orthogonal least squares learning algorithm for radial basis function networks. **IEEE Transactions on neural networks**, *2*(2), pp.302-309.

Crow, J.F., 2010. On epistasis: why it is unimportant in polygenic directional selection. **Philosophical Transactions of the Royal Society B: Biological Sciences**, *365*(1544), pp.1241-1244.

Denis M., Bouvet J.M., 2012. Efficiency of genomic selection with models including dominance effect in the context of Eucalyptus breeding. **Tree Genet Genomes** 9:p37–51.

Doebley J., Stec A., Gustus C., 1995. Teosinte branched l and the origin of maize: evidence for epistasis and the evolution of dominance. **Genetics** 141:333–346.

Dudley J.W., 2008. Epistatic interactions in crosses of Illinois high oil 9 Illinois low oil and of Illinois high protein 9 Illinois low protein. **Crop Science** 48:59–68.

Dudley J.W., Johnson G.R., 2010. Epistatic models improve between years prediction and prediction of testcross performance in corn. **Crop Science** 50:763–769.

Felipe, V.P., Silva, M.A., Valente, B.D. and Rosa, G.J., 2015. Using multiple regression, Bayesian networks and artificial neural networks for prediction of total egg production in European quails based on earlier expressed phenotypes. **Poultry science,** *94*(4), pp.772-780.

Fu, Y.B, Ritland K., 1996. Marker-based inferences about epistasis for genes influencing inbreeding depression. **Genetics** 144: 339–348.

Gianola, D., 2015. Genomic-assisted prediction of breeding values: indications of its effectiveness.

Gianola D., Okut H., Weigel K.A., Rosa G.J.M., 2011. Predicting complex quantitative traits with neural networks: a case study with Jersey cows and wheat. **BMC Genetics.** 12:87.

Gianola D., van Kaam J.B.C.H.M., 2008. Reproducing kernel Hilbert space regression methods for genomic-assisted prediction of quantitative traits. **Genetics** 178:2289–2303.

Gianola, D. and De Los Campos, G., 2008. Inferring genetic values for quantitative traitsnon-parametrically. **Genetics Research,** 90(6), pp.525-540.

Gianola D., Fernando R., Stella A., 2006. Genomic-assisted prediction of genetic values with semiparametric procedures. **Genetics** 173:1761–1776.

Gonzalez-Recio O., Gianola D., Long N., Wiegel K., Rosa G.J.M., Avendano S., 2008 Non parametric methods for incorporating genomic information into genetic evaluation: an application to mortality in broilers. **Genetics** 178:2305–2313.

Hastie T., Tibshirani R., Friedman J., 2009. **The elements of statistical learning**, 2nd edn. Springer, New York .

Haykin, S., 2001. **Redes neurais: princípios e prática.** 2ed. Porto Alegre: Bookman.

Haykin, S., 2009. **Neural Networks and Learning Machines**. 3ed. Ontario:McMaster University Hamilton.

Hill, W.G., Goddard M.E., Visscher P.M., 2008. Data and theory point to mainly additive genetic variance for complex traits**. PLoS Genetics** 4(2).

Holland, J.B., 2001. Epistasis and plant breeding. **Plant Breed Rev** 21:27–92.

Holland J.B., 2006. Theoretical and biological foundations of plant breeding. In: Lamkey K.R., Lee M (eds) **Plant breeding:** the Arnel R Hallauer International Symposium. Blackwell Publishing, Ames.

Howard, R., Carriquiry, A.L., Beavis, W.D., 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **Genetics** 4:6.

Kang, H., Zhou, L. and Liu, J., 2017. Statistical considerations for genomic selection. **Frontiers of Agricultural Science and Engineering**, *4*(3), pp.268-278.

Lark, K.G., Chase, K., Adler, F., Mansur, L.M. and Orf, J.H., 1995. Interactions between quantitative trait loci in soybean in which trait variation at one locus is conditional upon a specific allele at another. **Proceedings of the National Academy of Sciences**, *92*(10), pp.4656-4660.

Lee, S.H., van der Werf, J.H., Hayes, B.J., Goddard, M.E. and Visscher, P.M., 2008. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **PLoS genetics**, *4*(10), p.e100023.

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A., Kranis, A. and Gonzalez-Recio, O., 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genetics research**, *92*(3), pp.209-225.

Long, N., Gianola, D., Rosa, G.J. and Weigel, K.A., 2011. Marker-assisted prediction of non-additive genetic values. **Genetica**, *139*(7), pp.843-854.

Matlab (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.

Mckinney, B. and Pajewski, N., 2012. Six degrees of epistasis: statistical network models for GWAS. **Frontiers in genetics**, *2*, p.109.

Meuwissen T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic values using genome-wide dense marker maps. **Genetics**, *157*(4), pp.1819-1829.

Nei, M., 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. **Genetics**, *89*(3), pp.583-590.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manès, Y. and Dreisigacker, S., 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. **G3:Genes, Genomes, Genetics**, *2*(12), pp.1595-1605.

Pooni, H.S., Coombs, D.T., Virk, P.S. and Jinks, J.L., 1987. Detection of epistasis and linkage of interacting genes in the presence of reciprocal differences. **Heredity**, *58*(2), p.257.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.R-project.org/

Resende, M.D.V., Silva, F.F., Azevedo, C.F., 2014. **Estatística Matemática, Biométrica e Computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados**

(REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição, Sobrevivência. Editora Suprema, Viçosa.

Routman, E.J. and Cheverud, J.M., 1997. Gene effects on a quantitative trait: two-locus epistatic effects measured at microsatellite markers and at estimated QTL. **Evolution**, *51*(5), pp.1654-1662.

Rutkoski, J.E., Heffner, E.L. and Sorrells, M.E., 2011. Genomic selection for durable stem rust resistance in wheat. **Euphytica**, *179*(1), pp.161-173.

Santos, V. S., Filho S. Martins, M. D. Resende, C. F. Azevedo, P. S. Lopes, S. E. Guimarães, and F. F. Silva., 2016. Genomic prediction for additive and dominance effects of censored traits in pigs. **Genetics and molecular research,**15: 4.

Silva, I. N., Spatti, H. D., Flauzino, R. A. **Redes Neurais Artificiais: para engenharia e ciências aplicadas**. São Paulo: Artliber, 2010. 399p.

Toro, M.A. and Varona, L., 2010. A note on mate allocation for dominance handling in genomic selection. **Genetics Selection Evolution**, *42*(1), p.33.

Viana, J.M.S. and Piepho, H.P., 2017. Quantitative genetics theory for genomic selection and efficiency of genotypic value prediction in open-pollinated populations. **Scientia Agricola,** *74*(1), pp.41-50.

Viana, J.M.S., Piepho, H.P. and e Silva, F.F., 2016. Quantitative genetics theory for genomic selection and efffciency of breeding value. **Scientia Agricula,** *73*(3), pp.243-251.

Viana, J.M.S., 2005. Dominance, epistasis, heritabilities and expected genetic gains. **Genetics and Molecular Biology**, *28*(1), pp.67-74.

Viana, J.M.S., 2000. Components of variation of polygenic systems with digenic epistasis. **Genetics and Molecular Biology**, *23*(4), pp.883-892.

Wang, C.S., Rutledge, J.J. and Gianola, D., 1994. Bayesian analysis of mixed linear models via Gibbs sampling with an application to litter size in Iberian pigs. **Genetics Selection Evolution**, *26*(2), p.91.

Wu, P., Zhang, G., Huang, N. and Ladha, J.K., 1995. Non-allelic interaction conditioning spikelet sterility in an F 2 population of indica/japonica cross in rice. **Theoretical and Applied Genetics**, *91*(6-7), pp.825-829.

Yu, S.B., Li, J.X., Xu, C.G., Tan, Y.F., Gao, Y.J., Li, X.H., Zhang, Q. and Maroof, M.S., 1997. Importance of epistasis as the genetic basis of heterosis in an elite rice hybrid. **Proceedings of the National Academy of Sciences**, *94*(17), pp.9226-9231.

Zhang, K., J. Tian, L. Zhao, and S. Wang, 2008 Mapping QTLs with epi- static effects and QTL environment interactions for plant height using a doubled haploid population in cultivated wheat. J. Genet. Genomics 35 (2): 119–127.

Zhang, Y.M. and Xu, S., 2005. A penalized maximum likelihood method for estimating epistatic effects of QTL. **Heredity**, *95*(1), p.96.

CAPITULO 2

Capitulo 2. Predição Genômica de Caracteres Quantitativos por meio de Redes Neurais Artificiais após redução da dimensionalidade

RESUMO

SANT'ANNA, Isabela de Castro, D.Sc., Universidade Federal de Viçosa, fevereiro de 2018. **Predição Genômica de Caracteres Quantitativos por meio de Redes Neurais Artificiais após redução de dimensionalidade.** Orientador: Cosme Damião Cruz; Coorientadores: Matias Kirst e Marcos Deon Vilela de Resende.

Um dos grandes desafios do melhoramento genético atual é entender a variação da herança genética de caracteres quantitativos, QTLs (*Quantitative trait loci*), que são condicionados por um grande número de genes com pequeno efeito cuja interação resulta, muita vezes, em não linearidade nas relações entre fenótipos e genótipos. Com o surgimento da Seleção Genômica (GS) tornou-se possível estimar o valor genético de indivíduos (VGG) que ainda não foram fenotipados. No entanto, a maioria das modelagens da SG contempla apenas a porção aditiva do valor genético, o que dificulta, muitas vezes, uma representação mais realística da arquitetura genética dos caracteres, sendo a inclusão de dominância e interações epistáticas fatores cruciais para aumentar a acurácia da predição. A inclusão dessas interações é dificultada e conduz à superparametrização do modelo pelo grande número de marcadores presentes em todo o genoma e o reduzido número de indivíduos genotipados. Neste contexto, as Redes Neurais Artificias (RNAs) tornam-se alternativas de análise potenciais por capturar relações não lineares entre marcadores a partir e fenótipos o que a maioria dos modelos comumente utilizados na SG não consegue. Entretanto, a inclusão de todos os marcadores no modelo aumenta a de existência de alta correlação entre as marcas e representa um enorme desafio que acarreta menor precisão e grande demanda computacional para o treinamento da RNA que utilizam boa parte de seus recursos para representar porções irrelevantes do espaço de busca, dificultando o aprendizado. Assim, um modelo mais realístico deveria incluir SNPs apenas relacionadas a característica de interesse. Diante do exposto, foi proposto o uso métodos de redução de dimensionalidade, aplicados a predição de valores genéticos, para fins de seleção de um subconjunto de marcadores por meio de procedimento específicos como sondagens ou regressões *Stepwise*. Desta forma, o objetivo deste trabalho é avaliar a eficiência da seleção genômica ampla (GS) e das redes neurais artificias do tipo de base radial (RBF) e Perceptron de Múltiplas camadas (RNA-MLP) na predição do valor genético em população natural com desequilíbrio gamético. Para isso, foi simulada uma população F_1 oriunda da hibridação de genitores divergentes, com 500 indivíduos, genotipados com 1000 marcadores do tipo SNP. As características fenotípicas foram determinadas adotando-se três modelos: aditivo, aditivo-dominante e epistático, atendendo duas situações de dominância: parcial

e completa com caracteres quantitativos admitindo herdabilidades (h^2) de 30 e 60%, controlados cada um por 100 locos, considerando dois alelos por loco, totalizando 12 cenários distintos. Para avaliar a capacidade de predição, o modelo RR-BLUP e RNA-RBF foram treinados utilizando 80% dos indivíduos da população e procedimento de validação cruzada com cinco repetições. Foi utilizado dois métodos de redução de dimensionalidade do tipo Stepwise e Sonda. O quadrado da correlação entre o valor genômico predito (GEBV) e valor fenotípico foi utilizado para medir a confiabilidade preditiva. Para a h^2= 0.3 no cenário aditivo, o R^2 de validação foi de 59% para rede neural (RNA-RBF), 57%(RNA-MLP) e 57% para RR-BLUP, e no cenário epistático os valores de R^2 foram de 50%, 47 e 41%, respectivamente. Adicionalmente, ao analisarmos a raiz erro quadrático médio a diferença é do desempenho das técnicas é ainda maior. Para o cenário 1, as estimativas foram de 91 (RR-BLUP) e 5 para ambas as redes neurais e, no cenário mais crítico, de 427(RR-BLUP) e 20 para as redes neurais. Os resultados obtidos mostram que a utilização de redes neurais juntamente com a seleção de variáveis permitem capturar as interações epistáticas levando a melhora na acurácia da predição do valor genético e, principalmente, grande redução do erro quadrático médio que indica maior confiabilidade da predição do valor genômico. E que a redução de dimensionalidade melhora as acurácias obtidas em todos os cenários utilizados.

ABSTRACT

SANT'ANNA, Isabela de Castro, D.Sc., Universidade Federal de Viçosa, February, 2017. **Genomic Prediction of Quantitative traits by Artificial Neural Networks after using dimensionality reduction methodology.** Adviser: Cosme Damião Cruz. Co-advisers: Marcos Deon Vilela de Resende, and Matias Kirst.

One of the major challenges of genetic breeding today is to understand the genetic variation of quantitative traits, QTLs (Quantitative trait loci), which are conditioned by a large number of genes with small effects whose interaction, often, results in non-linearity in relations between phenotypes and genotypes. With the advent of the Genomic Selection (GS), it became possible to estimate the genetic value of individuals (VGG). However, most applications of GS includes only the additive portion of the genetic value, and a more realistic representation of the genetic architecture of quantitative traits should have the inclusion of dominance and epistatic interactions. These are crucial factors to increase the accuracy of the prediction. The inclusion of these interactions is computationally challenging and leads to the superparametrization of the models that are already in high dimensionality because the large number of markers in the genome and the smallest number of individuals. In this context, Artificial Neural Networks (ANNs) has a great potential because they can capturing non-linear relationships between markers from the data themselves, which most of the models commonly used in the GS can not. However, the inclusion of all markers in the prediction model increases the chances of a high correlation between the marks and represents a huge challenge that add less precision and a great computational demand for ANNs training that use a good part of their resources to represent irrelevant portions of the search space and compromising the learning process. Thus, a more realistic model should include only SNPs that are related to the traits of interest. Because of this, it was proposed to use dimensionality reduction methods, applied to the prediction of genetic values, for the purpose of selecting a subset of markers by means of specific procedures such as Sonda or Stepwise regressions. In this way, the objective of this work is to evaluate the efficiency of genome enabled prediction by using RR-BLUP (GS) and artificial neural networks as radial basis function neural network (RBFNN), and Multi-layer Perceptron (RNA-MLP) in the prediction of the genetic value in a natural population with linkage disequilibrium. For this, an F1 population from the hybridization of divergent parents with 500 individuals genotyped with 1,000 SNP-type markers was simulated. The phenotypic traits were determined by adopting three different gene action models: additive, additive-dominance and epistasis,

attending two dominance situations: partial and complete with quantitative traits admitting heritabilities (h^2) ranging from 30 to 60%, each is controlled by 50 loci, considering two alleles per loco, totaling 12 different scenarios. To evaluate the predictive ability of RR_BLUP and the neural networks a cross-validation procedure with five replicates were trained using 80% of the individuals of the population. Two dimensionality reduction methods Stepwise and Sonda were used. The square of the correlation between predicted genomic estimated value (GEBV) and phenotype value was used to measure predictive reliability. For $h^2 = 0.3$ in the additive scenario, the validation R^2 was 59% for neural network (RBFNN), 57% (RNA-MLP) and 57% for RR-BLUP, and in the epistemic scenario R^2 values were 50%, 47 and 41%, respectively. Additionally, when analyzing the mean-squared error root the difference in performance of the techniques is even greater. For additive scenario, the estimates were 91 (RR-BLUP) and 5 for both neural networks and, in the most critical scenario, 427 (RR-BLUP) and 20 for neural networks. The results show that the use of neural networks and variable selection techniques allows capturing the epistasis interactions leading to an improvement in the accuracy of the prediction of the genetic value and, mainly, a large reduction of the mean square error that indicates greater reliability of the prediction of the genomic value.

1. INTRODUÇÃO

Um dos grandes desafios do melhoramento genético atual é quantificar e entender a variação da herança genética de caracteres quantitativos que são condicionados por um grande número de genes, QTLs (*Quantitative trait loci*), com pequeno efeito (Risch, 2000) cuja interação resulta, muita vezes, em não linearidade nas relações entre fenótipos e genótipos (Gianola e de los Campos, 2008; Yamamoto et al., 2008; Mackay, 2013).

Compreender a relação entre a variação fenotípica e a variação da sequência de DNA para os caracteres quantitativos de interesse possibilitará predições cada vez mais acuradas do valor genético. O avanço das tecnologias de genotipagem e a desenvolvimento de marcadores moleculares abundantes no material genético culminou com o surgimento da Seleção Genômica (SG) (Meuwissen et al., 2001). A SG permite estimação simultânea dos efeitos genéticos de marcadores dispersos em todo o genoma, sob o pressuposto de que a maioria dos alelos de interesse esteja associado a esses marcadores (em desequilíbrio de ligação) e possa explicar grande parte da variação genética e predizer o valor genético dos indivíduos que ainda não tiveram seus fenótipos coletados (Resende et al., 2014).

O valor genético de indivíduos pode ser atribuído a porções relativas aos efeitos aditivos dos alelos, aos efeitos da dominância, que expressam as interações intra-alélicas, e aos efeitos epistáticos, que expressam as interações inter-alélicas. No entanto, a maioria das modelagens da seleção genômica usa apenas modelos de regressão que contemplam a porção aditiva do valor genético, o que dificulta, muitas vezes, uma representação mais realística da arquitetura genética dos caracteres quantitativos, sendo a inclusão de dominância e interações epistáticas fatores cruciais para aumentar a acurácia da predição (Lee et al., 2008; Akidemir et al., 2017).

A inclusão das interações alélicas na predição do valor genético é dificultada pelo elevado número de marcadores presentes em todo o genoma, e contemplados no modelo, e um reduzido número de indivíduos genotipados (Gianola et al., 2006). Neste contexto, em vez de apresentar pressupostos restritivos sobre a relação entre genótipos e fenótipos, pode-se considerar abordagens não-paramétricas, por exemplo, como as Redes Neurais Artificiais (RNA) (Long et al., 2011a). As RNAs podem capturar relações não lineares que a maioria dos modelos comumente utilizados na SG não conseguem (Gianola et al., 2006; Haykin, 2009; Long et al., 2010 e Howards et al., 2014). Nas RNAs a arquitetura das características é inferida dos próprios dados utilizados em seu treinamento, não necessitam do conhecimento de distribuições a priori como os métodos bayesianos e

não há necessidade de atender pressuposições sobre as distribuições dos dados e dos resíduos.

Os desafios estatísticos relacionados à alta dimensionalidade é devido ao fato do número de marcadores (p) ser muito maior que o número de observações (n) (p> n) e, também, a alta correlação existente entre os marcadores (Crossa et al., 2017) que representa enorme desafio computacional já que existem centenas de milhares de marcas disponíveis no genoma (Long et al., 2011b). Para as técnicas fundamentadas em inteligência computacional a grande quantidade de marcadores disponível no genoma, que representam as entradas das redes, acarreta menor acurácia do valor predito e grande demanda computacional para o treinamento da RNA que utilizam boa parte de seus recursos para representar porções irrelevantes do espaço de busca, dificultando o aprendizado. Segundo Long et al. (2011a), um modelo mais realístico incluiria SNPs apenas relacionadas a característica de interesse. Uma solução é escolher um subconjunto de SNPs para o treinamento de dados pois, com a redução do espaço de buscas, as RNAs melhoram o processo de aprendizado e aumentam o poder preditivo de modelo (Long et al., 2010).

Os métodos propostos para redução de dimensionalidade incluem métodos baseados em regularização tais como penalizações (RR-BLUP), redução via combinações lineares independentes (Horn e Camp, 2004; Long et al., 2010; Woolaston et al., 2007; Azevedo, et al., 2013;2014; James et al., 2013; Resende et al., 2014), mínimos quadrados parciais (Moser et al., 2007; Tier et al., 2007), e seleção de um subconjunto de marcadores por meio de procedimento específicos como sondagens ou regressão *Stepwise* (Habier et al., 2007; Piyasatian et al., 2007).

Para tanto, o objetivo deste trabalho, em comparação com o capitulo anterior, é demonstrar que os resultados obtidos pela predição do valor genético com informações reduzidas preservam as mesmas conclusões quando se utiliza um conjunto de dados maior e que ainda podemos acrescentar, com grande facilidade, a utilização de redes neurais artificiais que envolvam topologias mais complexas. Dessa forma, pretendemos observar se a presença das interações epistáticas afeta as predições genômicas, com a expectativa de que a abordagem das redes neurais artificiais possa capturar padrões complexos de interação não lineares. Além disso, descrevemos também uma nova estratégia de seleção de variáveis aplicável ao melhoramento genético, denominado método das Sondas.

2.MATERIAL E MÉTODOS

2.1. Estabelecimento de uma população com desequilíbrio de fase gamética

Conforme descrito no capítulo anterior foi utilizada uma população F1 proveniente de genitores geneticamente divergentes e que, supostamente, teria maior concentração de marcadores em desequilíbrio de fase gamética.

2.2. Estabelecimento dos valores genotípicos

A população F1 foi então genotipada em relação a 1000 marcadores do tipo SNP. Para proceder às análises de seleção genômica, foram simulados valores genotípicos e fenotípicos considerando três níveis de dominância (0, 0.5 e 1), diferentes herdabilidades (h^2 = 0.3 ou 0.6), e dois cenários considerando que o valor genotípico era estabelecido por um modelo aditivo (1), aditivo-dominante (2) ou epistático (3), conforme descritos no capítulo anterior.

As características foram estabelecidas pela ação de alelos de 50 locos, tomados ao acaso entre os 1000 previamente genotipados, com efeito gênico diferencial e com pesos da importância do loco, sobre a variabilidade genotípica total do caráter, estabelecidos a partir de uma distribuição binomial $(p+q)^n$, em que p=q =0,5 e n= 49.

2.3. Estabelecimento dos valores fenotípicos

Os fenótipos dos indivíduos foram gerados segundo o modelo $F_i = G_i + E_i$, em que G_i é o efeito genotípico dado pelo somatório dos efeitos genotípico de cada loco, acrescido de efeito multiplicativo entre os efeitos de pares locos quando existia epistasia, e E_i o efeito ambiental, gerado segundo uma distribuição normal com média zero e variância compatível com a herdabilidade do caráter simulado. Foram simulados doze cenários considerando que o valor genotípico era estabelecido por um modelo aditivo incluindo, ou não, interações epistáticas.

Tabela 1. Características avaliadas no estudo com seus respectivos valores de herdabilidade, modelo adotado e grau médio de dominância (Gmd).

Característica	Herdabilidade (%)	Modelo	Gmd
V1 - D0H30_Ad	30	aditivo	0
V2 -D0.5H30_Ado	30	aditivo-dominante	0.5
V3- D1H30_Ado	30	aditivo-dominante	1
V4 - D0H30_Ep	30	epistático	0
V5 -D0.5H30_Ep	30	epistático	0.5
V6 -D1H30_Ep	30	epistático	1
V7 - D0H60_Ad	60	aditivo	0
V8 - D0.5H60_Ado	60	aditivo-dominante	0.5
V9 - D1H60_Ado	60	aditivo-dominante	1
V10 - D0H60_Ep	60	epistático	0
V11 -D0.5H60_Ep	60	epistático	0.5
V12 - D1H60_Ep	60	epistático	1

2.4. Métodos de Redução de Dimensionalidade Avaliados

Nos estudos de associação entre características quantitativas um grande numero de variáveis independes (marcadores associados aos locos de interesse) são utilizados na predição da variável dependente Y(características quantitativas). Entretanto, em estudos de regressão múltipla, em que um conjunto de p variáveis independentes estão disponíveis, o modelo de regressão não necessariamente precisa de todas estas variáveis. Deve-se escolher o modelo que melhor explique a variação na variável dependente Y, incluindo apenas variáveis independentes que contribuem para a eficiência do modelo, expressa pela sua capacidade de predizer Y. O modelo mais adequado para explicar as variações na variável dependente é aquele que inclui todas as variáveis relevantes, e associadas a variável dependente, sem incluir variáveis redundantes (ou seja, correlacionadas entre si).

2.4.1. Método da Sonda

O Método da Sonda consiste numa metodologia, baseada no Algoritmo Genético (Holland,1975), para a qual um número N de subamostras (Sondagens) é considerado. A cada "sondagem", uma amostra aleatória de m variáveis explicativas é tomada para

ajustar a variável resposta. Uma vez realizada a análise de regressão, é estabelecido um coeficiente de importância da variável, para a i-ésima sona (i=1,2...k) dado pelo produto entre o coeficiente de regressão, em valor absoluto, e de determinação do modelo. Após todas as sondagens, é estabelecido um índice de importância relativa (IR_{xi}), associado a cada variável, conforme descrito:

$$IR_{x_i} = \frac{1}{k}\sum_{i=1}^{k} \beta_i R_i^2, \text{ para i = 1,2, ..., m}$$

Em que:
k: número de vezes que a variável x_i participou das N subamostras do total de 20000 mil subamostras testado;
β_i: coeficiente da regressão associado à variável x_i incluída no modelo de regressão obtido numa determinada sonda S_i que tenha sido incluída;
R_i^2: coeficiente de determinação obtido pela regressão ajustada na sonda S_i.
m: número total de variáveis estudadas, neste trabalho igual a 1000 variáveis (ou marcadores).

De posse das importâncias relativas de cada variável a ser considerada na equação, uma nova regressão é determinada, por meio da combinação das variáveis de maior desempenho (maior valor de IR_{x_i}). Essa nova combinação constitui a subamostra a ser adotada em posteriores análises de seleção genômica.

2.4.2. Método de Regressão Stepwise

O modelo se inicia com a inclusão da variável que apresente maior correlação simples com a variável dependente de interesse. A partir daí o passo *forward* ("para a frente") – com base na correlação parcial – é determinante para a inclusão de novas variáveis no modelo. O passo *backward* ("para trás"), por sua vez, ocorre logo após uma variável ter sido incluída no modelo.

Na execução do procedimento *Stepwise* são executados os seguintes passos: Calcula-se as correlações entre todas as características explicativas (independentes) e a característica resposta (dependente) que está sendo analisada. Seleciona-se como a primeira variável a ser considerada na análise de regressão aquela de mais alto coeficiente de correlação simples com a variável resposta. Posteriormente, outras variáveis poderão ser incluídas no modelo considerando a correlação parcial entre a variável dependente a

ser incluída e a variáveis resposta, removendo os efeitos daquelas já incluídas no modelo. Com base na estatística F parcial, e dos valores referenciais de probabilidade de entrada e de saída, é determinado se a variável avaliada permanecerá ou será excluída do modelo. Neste trabalho, a propabilidade de entrada e saída era de 0.05, de forma que a variável era mantida no modelo quando o *p-value* era inferior a este nível de significância.

2.4.3 Determinação do Número de Marcas

Neste trabalho, adaptamos três critérios que permitissem avaliar a qualidade do modelo reduzido, uma vez que o resultado da análise do conjunto de dados poderia ser afetado por dois fatores importantes que são a multicolinearidade e a dimensionalidade proporcionada por um número mais elevado de marcadores (variáveis independentes) do que de observações. Assim, além de consideradas as probabilidades de entrada e saída associadas ao teste F parcial, para o caso da metodologia *Stepwise,* e por meio da combinação das variáveis de maior importância relativa no caso da Sonda, a redução da dimensionalidade foi estabelecida observando o comportamento de três medidas estatísticas, sendo a raiz do erro quadrático médio do modelo (REQM), o R^2 ao incluir as marcas selecionadas indicando a porcentagem da variação na variável resposta que os marcadores conseguiam explicar e o número de condição da matriz de correlação (NC) (Figura 2). Quanto aos dois primeiros critérios, o REQM escolhido foi aquele que apresentava o menor valor possível atrelado aos melhores valores possíveis para o R^2 (quanto maior melhor). Já o terceiro critério, o número de condição da matriz de correlação entre as variáveis explicativas, verifica, se o grau de multicolinearidade na matriz de correlação X'X (Montgomery e Peck, 1982). Quando o NC resultante dessa divisão foi menor ou igual a 100, considerou-se haver multicolinearidade fraca entre as variáveis explicativas; para $100<NC<1000$ multicolinearidade moderada a severa e $NC\geq1000$, considerou-se multicolinearidade severa. A Figura 2 ilustra o procedimento do estabelecimento do número adequado de marcadores para predição em variável resposta, na regressão *Stepwise*, sendo monitorados por medidas de qualidade do modelo relativas ao coeficiente de determinação, erro quadrático médio e número de condição.

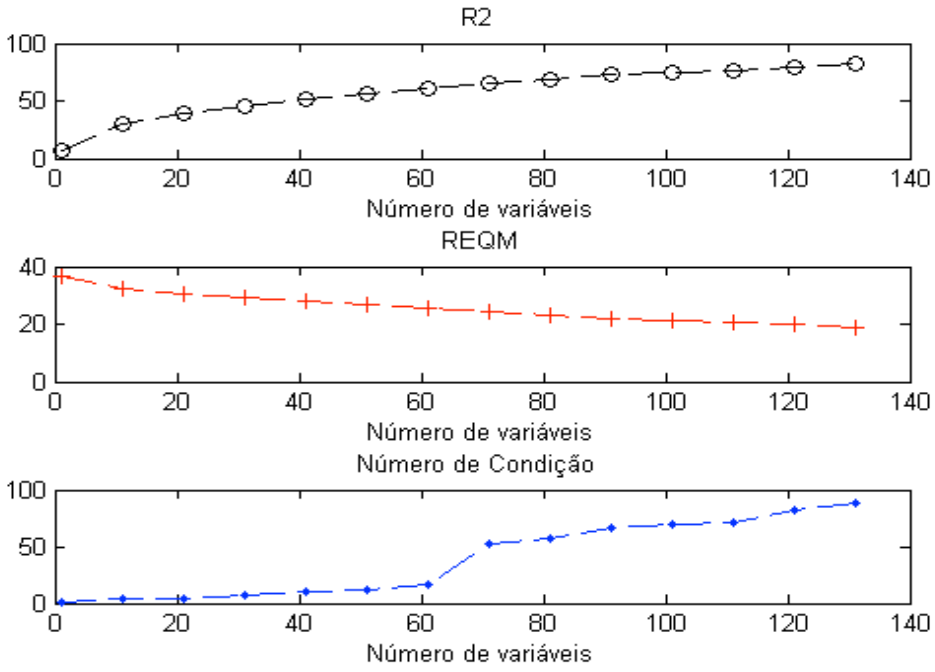


Figura 2: Representação gráfica do número de condição, o R^2 e o REQM para N=100 marcas pela metodologia de *Stepwise*.

3. MÉTODO ESTATÍSTICO EMPREGADO PARA FINS DE PREDIÇÃO GENÔMICA

3.1. Predição do GEBVs (*Genomic estimated Breding Value*)

Três métodos foram utilizados para predição dos valores genômicos: RR-BLUP (Meuwissen et al., 2001), Redes Neurais do tipo de Base Radial (Long et al., 2010) e Rede Neural do tipo Perceptron de Múltiplas camadas (Gianola et al., 2011).

3.1.1. Predição de valores genéticos por meio de RR-BLUP com e sem redução de dimensionalidade do conjunto de dados explicativos

Para estimar os efeitos de marcas e os dos valores genômicos (GEBVs) foi utilizado a metodologia RR-BLUP conforme descrito por Meuwissen et al. (2001) onde:

$$y = Xb + Za + e,$$

em que y é o vetor de observações fenotípicas, b é o vetor de efeitos fixos, a é o vetor dos efeitos aleatórios dos marcadores e e refere-se ao vetor de erros aleatórios. X e Z são as matrizes de incidência para b e a. A estrutura de médias e variâncias no modelo em questão é definida como: $a \sim N(0,G)$, $E(y)=Xb$, $e \sim N(0,R=I)$, $Var(y)=V=ZGZ' + R$.

$G = I\sigma_{\frac{g}{n}}^2$

em que n é o número de marcadores dipostos no genoma, Z_{ij} é a linha da matriz de incidência que aloca o genótipo do i=ésimo marcador para cada indivíduos, 0, 1,-1 para os genótipos AA, Aa, AA, respectivamente, para marcadores bialélicos e codominantes, e $\hat{\alpha}_i$ é o efeito estimado do i-ésimo marcador.

De posse dos efeitos de marcadores foram estimados os efeitos dos indivíduos (GEBVs) por meio do seguinte estimador:

$$G\hat{E}BVs = \hat{y}_j = \sum_i^n Z_{ij}\hat{a}_i$$

As equações de predição apresentadas acima foram modeladas assumindo a priori de que todos os locos explicam quantidades iguais da variação genética e portanto apresentam σ_g^2 comum. Assim, a variação genética explicada por cada loco é dada por (σ_g^2/n),em que σ_g^2 é a variação genética total e *n* é o número de marcadores utilizados considerado como sendo originalmente igual a 1000 e após redução, por regressão Stepwise ou Sondagem, iguais a 100.

3.1.2. Predição de valores genéticos por meio de RNA-RBF após redução de dimensionalidade do conjunto de dados explicativos

Foi adotada uma rede RNA-RBF de arquitetura do tipo *feedforward* que consiste de uma camada de entrada, uma camada oculta, e a camada de saída. O número máximo de neurônios fornecidos foi 200 e critério de parada estabelecido quando o limite mínimo do EQM=0.01 fosse atingido. O tamanho do raio necessário para que esse EQM fosse atingido variou de 5 a 100, sendo em média de 55. As entradas da rede eram as informações dos marcadores utilizados considerado como sendo originalmente igual a 1000 e, após redução por regressão Stepwise ou Sondagem, iguais a 100. A função de ativação é gaussiana para a camada oculta considerada conforme pode ser visto na Figura 3.

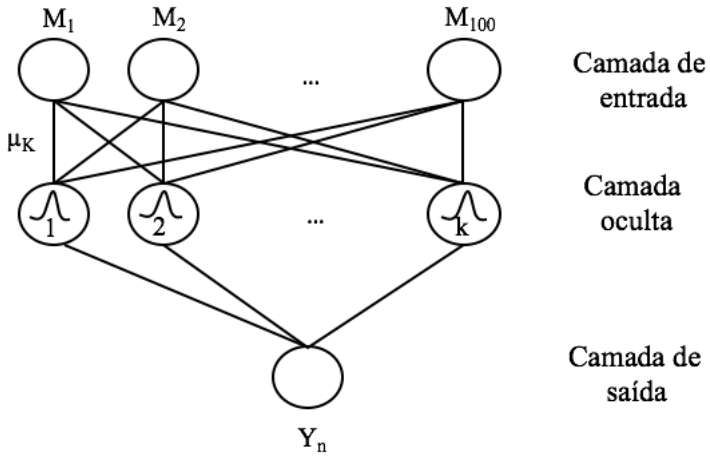


Figura 2. Arquitetura e topologia de uma Rede Funções de Base Radial com número de entradas igual a 100, K neurônios na camada intermediária (variando de 1 a 200) e uma saída (Y_n) que envolvia 400 observações no processo de treinamento e 100 no processo de validação.

O treinamento da rede função de base radial foi realizado utilizando-se o procedimento de validação cruzada fundamentado na reamostragem de um grupo de indivíduos via procedimento *k-fold* (Bengio e Grandvalet, 2004). E seguindo o procedimento de validação cruzada, a saída desejada eram os valores fenotípicos observados pelos indivíduos e as saídas de redes representavam os valores genotípicos preditos.

Para predizer o valor genético foi utilizado uma topologia RNA-RBF que faz com que o valor genético seja aproximadamente uma combinação linear do *m* funções de base radial com todos os marcadores incorporados. Isso significa que $k(x_i, x_j)$ depende da ($\|x_i - x_j\|$) distância euclidiana ($\|x_i - x_j\|$) entre dois vetores.

$$y_i = \sum_{j=1}^{m} k(x_i, x_j)\alpha_j \text{ em que:}$$

μ é a media geral; $k(x_i, x_j)$ é a função de base radial dos genótipos $(x_i \text{e } x_j)$de dois indivíduos (*i* and *j*); Os valores de x_j(j=1,…,m) são temas centrados na função de base radial. Para cada termo centrado em x_j tem um coeficiente de regressão α_j que é desconhecido.

3.1.3. Predição de valores genéticos por meio de Perceptron Múltiplas Camadas-RNA-MLP com redução de dimensionalidade do conjunto de dados explicativos

Neste trabalho foi adotada também a rede *Multilayer Perceptron* (RNA -RNA-MLP) para predição dos valores genéticos em comparação com as abordagens de SG e RNA-RBF. Como apresentado para o modelo de rede de base radial, consideram-se os marcadores como as variáveis explicativas (X_1, X_2 ...X_n) como entradas e Y (valor fenotípico do indivíduo) como a resposta desejada da rede. As RNA-RNA-MLP, quando faz uso de topologias mais complexas, demandam maior tempo computacional para a busca de uma solução ótima e, normalmente, esta solução é encontrada entre uma de muitas opções de análises em que se varia o número de camadas, o número de neurônios por camada, o tipo de função de ativação, o uso de algum tipo de aprendizado específico e o algoritmo de treinamento.

Neste trabalho, as RNA-MLP foram utilizadas com três camadas ocultas de neurônios, com variação de um a três neurônios por camada. O número máximo de épocas foi igual a 3000. O software utilizado para implementação da rede neural foi o Matlab, por meio do modulo integração no portal Genes (Cruz, 2016).

A arquitetura da rede foi estabelecida tal qual a Figura 4 apresentada a seguir:

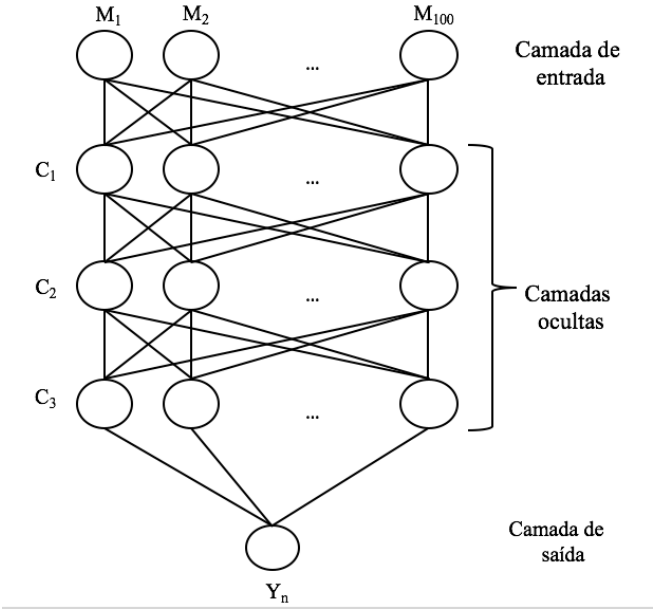


Figura 3. Arquitetura da RNA. Entradas de M_1 a M_{100} na camada entrada estão relacionadas com as informações dos marcadores e são consideradas como entrada (modelo reduzido). As camadas ocultas foram compostas por n_i neurônios (n_i variando de 1 a 3), com funções de ativação *tansig* ou *logsig*. Todas as combinações foram exploradas. Na camada de saída, a RNA retornou o valor genético predito.

Na fase de treinamento da rede, o algoritmo adotado foi o algoritmo de *backpropagation* é realizado mediantes duas etapas principais: fase 1 – *forward* – e fase 2 – *backward.* Na fase 1, as entradas e seus respectivos pesos sinápticos são propagados, camada a camada ao longo da rede, até que a saída de rede seja produzida e comparada à saída desejada. Por fim, calcula-se o erro da rede. A fase 2 de retropropagação (*backward*), por sua vez, começa na camada de saída. O erro é propagado para as camadas anteriores, permitindo que os pesos sinápticos sejam recalculados de acordo com a regra Delta até que se retorne à primeira camada da rede.

$$\Delta w_{(t)} = \alpha \Delta w_{(t-1)} + \eta \delta_{(t)} y_{(t)}$$

Em que:

α: é a constante de *momentum* com $0 < \alpha < 1$;

δ: é o gradiente local;

η: taxa de aprendizagem;

y: a saída desejada.

No caso do modelo de predição de valores genômicos temos a matriz X com N linhas (i=1,2,...N) e P colunas (j=1,2,..,P). Cada elemento da matriz X é dado por Xij com valores de -1(aa),0(Aa) e 1(AA) representando cada um dos indivíduos genotipados. A camada oculta é composta por T neurônios (t=1,2,..,T) e cada um deles gera um vetor de pesos, w_{1j}^{t} e um vetor bias, b_t. Isso resulta em uma combinação linear que é transformada usando uma uma função de ativação linear ou não linear que gera uma resposta, $a_i^{[t]}$ do neurônio t para o individuo i, conforme descrito na equação abaixo:

$$a_i^{[t]} = f(\sum_{j=1}^{P} w_{1j}^{[t]} + b_t),$$

A resposta do neurônio $a_i^{[t]}$dependera da importância da variável de entrada para a variável resposta e será transformada pela função de ativação, que pode ser linear ou não. Na camada de saída o escore obtido pelos indivíduos i no neurônio T é então transformado em uma nova combinação linear pela função de ativação g(.).

$$y_i = g(\sum_{t=1}^{t} w_{2t} a_i^{[t]} + b_t) = g(\sum_{t=1}^{T} w_{2t} f\ (\sum_{j=1}^{P} w_{1j}^{[t]}\ x_{ji} + b_t) + b)$$

Esse resultado é então backpropagado para que os pesos e o viés sejam atualizados de forma a minimizar a diferença entre a saída obtida e a saída desejada. Em nosso trabalho a função de ativação g era umas das funções de ativação adotadas *logsig* ou *tansig* para a primeira, segunda e terceira camada oculta e *purelin* para a camada de saída. Cada neurônio é conectado aos neurônios das camadas anteriores por meio de pesos adaptativos

(Haykin, 2009). A forma como esses pesos são avaliados é crucial. O conjunto de pesos que minimiza alguma função de perda é tomado como solução do problema de aprendizagem para a rede neural.

Para fins de comparação as análises considerando as etapas de treinamento e validação seguiram as mesmas estratégias aplicadas no estudo da SG e RNA-RBF. A apresentação de todos os vetores de treinamento à rede define uma época de treinamento, nesta fase a condição de parada é testada e se não for satisfeita, o conjunto de treinamento é embaralhado e a rede continua seu processamento iterativamente.

3.1.4. Medidas Biométricas associadas a Redes Neurais Artificiais para comparação entre os modelos avaliados

No presente trabalho para avaliar as eficiências do modelos utilizados RR-BLUP, RNA-RBF e RNA-MLP na seleção genômica (SG) foram utilizados as seguintes estatísticas: confiabilidade ou acurácia seletiva do modelo, raiz do erro quadrático médio (REQM) ou acuracia preditiva e o critério de informação de Akaike (AIC) (apenas para RR-BLUP).

a) Acurácia e confiabilidade seletiva

A acurácia da SG depende da proporção da variação genética genômica (VGG) do indivíduo explicada pelos marcadores (Resende et al., 2014). Quando a população de validação é independente da população utilizada para predição de efeitos dos marcadores, a correlação entre o efeito estimado dos marcadores e o valor fenotípico dos indivíduos é predominantemente de natureza genética (Resende et al., 2014). Nesse caso, a correlação passa a ser definida como sendo a capacidade preditiva da SG e é dada pela equação abaixo:

$$CP = cor(y , y_p)$$

Sendo y e y_p os valores observados e os valores predito respectivamente.

.

O quadrado dessa correlação, também conhecido como confiabilidade, é uma medida do quadrado da correlação entre o valor estimado e os valores verdadeiros, ou seja, mede o quanto a estimativa obtida é relacionada com o valor real do parâmetro.

Nas análises realizadas foi obtido o quadrado da correlação entre o valor predito e o valor utilizado no aprendizado supervisionado e o valor fenotípico em analogia ao quadrado da correlação entre a média fenotípica e o valor fenotípico verdadeiro, que expressa a herdabilidade da característica.

$$R^2 = (cor(\ y\ ,\ y_p)\ )^2$$

b) Acurácia preditiva

Outra estatística usada na avaliação da eficiência da SG foi a acurácia preditiva medida pela diferença quadrática entre valores verdadeiros e estimados, ou seja, apresenta pelo mínimo erro quadrático médio (REQM) de predição, dado por:

$$REQM = \frac{\sqrt{\sum(y_p - y)^2}}{n}$$

d) Critério de informação de Akaike (AIC)

O Critério de informação de Akaike (AIC) foi proposto por Akaike (1974), sendo um critério de seleção de modelos que utiliza a Informação de Kullback-Leibler (K-L). Quanto menores valores do AIC melhor o ajuste global no modelo utilizado (Akaike, 1974). Este critério foi utilizado no RR-BLUP, para comparar modelos que envolviam todos marcadores ou apenas um número reduzido de marcadores.

3.2. Recursos Computacionais

A simulação e análise dos dados foram realizadas no Laboratório de Bioinformática da Universidade Federal de Viçosa, localizado no Instituto de Biotecnologia aplicada a Agropecuária (BIOAGRO) utilizando aplicativo computacional Genes (Cruz, 2016).

O Método do RR-BLUP foi implementado no Genes, no módulo integração com o software R (R Core Team, 2017).

Para determinar o numero ideal de marcadores a serem utilizados no modelo reduzido e também para a implementação das redes neurais foi utilizado o programa Genes no modulo de integração como software Matlab(Matlab, 2011),

O computador utilizado apresentava as seguintes características processador Intel(R)Core(TM) it-3770 CPU @3.4GHz 3.9GHz como Memoria RAM de 7.88GB Sistema Operacional Windows 64Bitz. O tempo de processamento dos métodos de redução de dimensionalidade foi menor para a regressão *Stepwise* do que para o procedimento da Sonda.

Quanto aos métodos de predição do valor genético o método do RR-BLUP era o mais rápido, seguido pela RNA-RBF e o RNA-RNA-MLP.

4. RESULTADOS E DISCUSSÃO

4.1. Procedimento de Redução de Dimensionalidade

Antes de fazer a predição dos valores genéticos dos indivíduos estudados, a partir das mensurações fenotípicas agregadas às informações moleculares, optou-se por avaliar a viabilidade de reduzir o conjunto de marcadores, utilizando procedimento estatístico adequado- técnica da regressão Stepwise e do método da Sonda- para orientar quantos e quais os marcadores deveriam ser mantidos para conservar, dentro de um modelo de ajuste linear, uma boa relação entre os valores preditos e os observados, sem que as informações genéticas fossem comprometidas. Como este estudo é feito a partir de dados simulados, já há expectativa prévia da maior, ou menor, dificuldade de predição tendo em vista os efeitos perturbadores do ambiente, da dominância e da epistasia.

O cenário de maior complexidade (V6) - herdabilidade baixa, dominância completa e epistasia - foi utilizado para a determinação do número ótimo de marcadores necessário para de predição do valor genético na população analisada. O resultado, para o conjunto de dados analisado, é apresentado na Figura 5. Assim, após análise gráfica, foi estabelecido que o número adequado de marcadores para fins de predição seria 100, uma vez que proporcionaria valor de R^2 adequado, com estimativa acima de 50%, associado à estatística REQM em níveis baixos (no trabalho, inferior a 40) e, principalmente, número de condição inferior a 100 que indica a ausência de multicolinearidade entre colunas da matriz de valores associados aos marcadores moleculares. Acima de 200 marcas a multicolinearidade existente nos dados dos marcadores moleculares se torna severa.

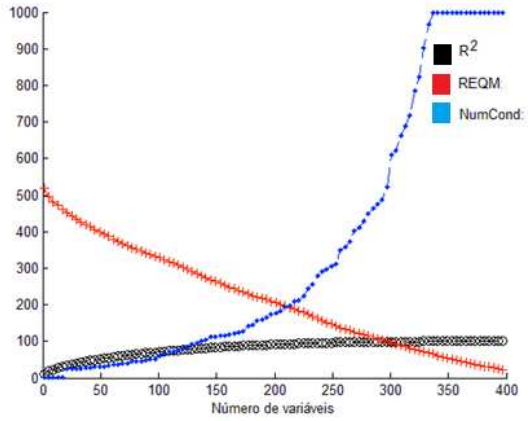


Figura 4: Representação gráfica conjunta dos valores de R^2, *REQM e NC* obtidos pelo método da regressão Stepwise ao incluir de 1 a 400 marcadores moleculares no modelo de regressão múltipla.

Os resultados obtidos pelo método da Sonda (Figura 6) indicam que é importante considerar a inclusão de um número de marcadores entre 100 e 130, uma vez que, para esse método, também foram obtidos NC < 100, valores estabilizados de R^2 e um aceitável valor REQM no intervalo sugerido. Desse modo, optou-se por aplicar os métodos de *Stepwise* e Sonda visando a seleção de 100 marcadores moleculares.

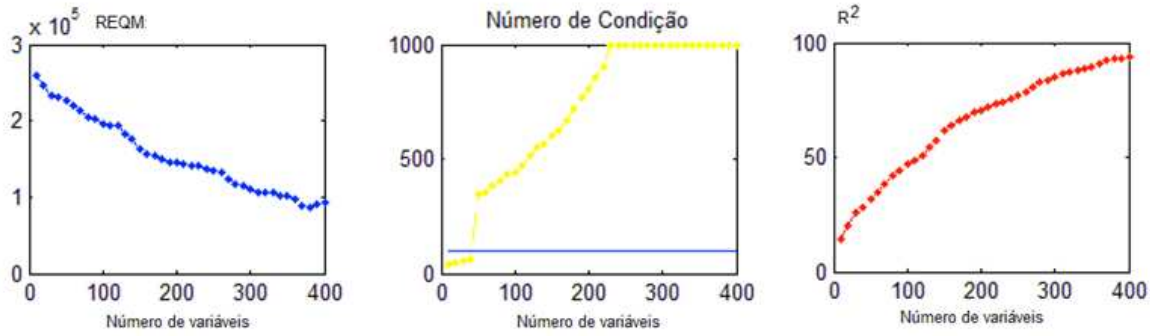


Figura 5:Representação gráfica conjunta dos valores de R^2, *REQM e N*C obtidos pelo método de Sonda ao incluir de 1 a 400 marcadores moleculares no modelo de regressão múltipla.

Embora a incorporação de marcas em todo o genoma representa um avanço no contexto de seleção de genótipos superiores e tenha feito a seleção genômica superar em muitos aspectos a seleção assistida por marcador molecular (Dekker et al., 2002), muitos autores (Long et al. 2007; Habier et al. 2009; Usai et al., 2009; Macciota et al., 2009;Weigel et al., 2009; Vasquez et al., 2010) tem percebido que utilizar um subconjunto SNPs, cuidadosamente escolhido para o caráter de interesse, pode resultar em confiabilidade razoavelmente elevada na predição de valores genômicos. Por exemplo, Weigel et al. (2009) descobriram que, usando um conjunto de 300 SNPs de maiores efeitos estimados do total de 32518 SNPs, obtém a metade da acurácia predição que o modelo completo utilizando a abordagem do Lasso Bayesiano. Já Usai et al. (2009), em suas simulações, relataram acurácia mais elevadas quando a predição foi realizada utilizando 169 marcadores selecionados dos 6000 SNPs usando o Lasso. Em nosso trabalho, como será visto posteriormente, a utilização dos SNPs selecionados melhora a acurácia do processo preditivo, para qualquer que seja o cenário, e simplifica a modelagem ou a arquitetura, no caso de redes neurais, utilizada no processo preditivo.

O número de marcadores escolhido em cada um dos cenários foi 100. Essa escolha se baseou principalmente em evitar os problemas relacionados à multicolinearidade que leva a estimativas superestimadas de R^2. Em cada cenário, diferentes marcadores foram escolhidos com a expectativa de que diferentes genes sejam mais, ou menos, importantes

na determinação do caráter. Essa expectativa foi confirmada nos trabalhos desenvolvidos por (Weigel et al. 2010a, b). Os autores compararam o uso de alguns de marcadores igualmente espaçados no genoma, realizando a imputação dos demais marcadores baseada em uma população referencia com todos os marcadores genotipados, com a utilização de um conjunto de marcadores selecionados devido a seu efeito na caráter de interesse. Os autores concluíram que, quando o número de marcadores selecionados é pequeno, a capacidade preditiva do modelo com os marcadores selecionados de acordo com o efeito é superior a utilização de um conjunto menor de marcadores dispersos no genoma.

4.2. Acurácia da predição de valores genéticos, por RR-BLUP, com e sem redução de dimensionalidade do conjunto de dados explicativos

No geral, os métodos de seleção genômica mais utilizados para predição de valores genéticos são os de regressão penalizada sob os enfoques frequentistas (RR-BLUP, G-BLUP, Lasso) e bayesianos (Bayes A e B e Lasso Bayesiano)(Resende et al.,2014). Entretanto, devido ao grande número de marcadores moleculares encontrados em todo o genoma, várias metodologias foram propostas para que os problemas gerados pela dimensionalidade da matriz de marcadores e a multicolinearidade inerente ao desequilíbrio de fase gamética fossem solucionados. Assim, dentre estas metodologias, destacaram aquelas capazes de reduzir a dimensionalidade e indicar variáveis mais relevantes a serem utilizadas em estudos genômicos por muitos autores. No presente trabalho, a metodologia de Sonda, aqui proposta, e a metodologia *Stepwise* foram escolhidas para selecionar as variáveis que pudessem melhor explicar cada cenário testado. Em nosso trabalho a redução da dimensionalidade foi utilizada com intuito de selecionar os SNPs relacionados às características de interesse com o foco de comparar a posterior performance dos métodos de inteligência computacional.

Para comparar a eficiência da redução da dimensionalidade, por meio do uso dos procedimentos de Sonda e *Stepwise,* para posterior estimação dos parâmetros genéticos pelo RR-BLUP uma comparação entre o R^2, REQM e a informação de Akaike (AIC) foram utilizadas em conjunto de dados de validação. Os resultados obtidos (Tabela 2) indicam que há maior eficácia no uso de RR-BLUP associado à estratégia de redução de dimensionalidades do tipo *Stepwise* para todos os cenários avaliados por meio de dois dos três parâmetros avaliados e, principalmente, pela queda no valor do índice AIC indicando

que o modelo menos parametrizado tem comportamento de predição mais vantajoso que o modelo que inclui os 1000 marcadores originais.

Tabela 2. Resultado da acurácia seletiva (R^2) e preditiva (REQM) de valores genéticos e índice de Akaike (AIC) a partir da abordagem por RR-BLUP sem (1000 marcadores) e com (100 marcadores) redução de dimensionalidade obtidos pela metodologia *Stepwise* (Sw) ou Sonda (S).

CENÁRIOS	R^2			REQM			AIC		
	1000	100 Sw	100 S	1000	100 Sw	100 S	1000	100 Sw	100 S
V1-D0H30_Ad	0,10 ± 0,02	0,57 ± 0,03	0,27 ± 0,05	97,35 ± 2,14	90,28 ± 0,03	37.08 ± 3.98	5952 ± 11	3997 ± 16	4079 ± 9
V2-D0.5H30_Ado	0,12 ± 0,07	0,58 ± 0,05	0,26 ± 0,10	122,65 ± 3,02	138,20 ± 3,66	66.56 ± 3,350	6004 ± 6	4031 ± 7	4149 ± 5
V3-D1H30_Ado	0,01 ± 0,01	0,54 ± 0,06	0,24 ± 0,07	267,59 ± 21,7	291,61 ± 17,7	241,53 ± 11,70	7565 ± 15	5629 ± 9	5691 ± 12
V4-D0H30_Ep	0,01 ± 0,01	0,42 ± 0,05	0,22 ± 0,04	278,61 ± 23,9	442,59 ± 48,1	244,12 ± 27,20	7596 ± 14	5699 ± 16	5731 ± 16
V5-D0.5H30_Ep	0,02 ± 0,02	0,54 ± 0,06	0,30 ± 0,07	366,06 ± 19,7	249,37 ± 18,9	300,23 ± 34,40	7785 ± 15	5838 ± 7	5893 ± 15
V6-D1H30_Ep	0,06 ± 0,05	0,47 ± 0,04	0,25 ± 0,05	575,41 ± 23,6	427,91 ± 40,57	478,29 ± 27,80	8148 ± 9	6240 ± 7	6273 ± 12
V7-D0H60_Ad	0,36 ± 0,07	0,79 ± 0,03	0,55 ± 0,09	85,76 ± 2,92	71,38 ± 0.98	48,27 ± 2,70	5602 ± 9	3527 ± 15	3679 ± 15
V8-D0.5H60_Ado	0,30 ±0,07	0,73 ± 0,03	0,47 ± 0,04	111,50 ± 3,87	107,40 ± 1.06	72,96 ± 4,70	5651 ± 17	3615 ± 16	3756 ± 13
V9-D1H60_Ado	0,19 ± 0,05	0,64 ± 0,01	0,32 ± 0,05	137,09 ± 3,85	145,95 ± 5.18	87,16 ± 10,30	5816 ± 12	3821 ± 11	3962 ± 9
V10-D0H60_Ep	0,03 ± 0,05	0,57 ± 0,04	0,32 ± 0,08	197,38 ± 12,1	320,11 ± 36,40	154,14 ± 5,11	7293 ± 13	5327 ± 9	5387 ± 11
V11-D0.5H60_Ep	0,08 ± 0,07	0,59 ± 0,05	0,34 ± 0,09	274,19 ± 21,9	280,93 ± 28,90	243,35 ± 18,90	7512 ± 17	5547 ± 15	5606 ± 17
V12-D1H60_Ep	0,13 ± 0,07	0,58 ± 0,09	0,38 ± 0,05	434,49 ± 26,4	473,80 ± 22,00	306,43 ± 31,70	7837 ± 18	5885 ± 20	5933 ± 10

Sob aspecto estatístico, a utilização de ambos os métodos de redução de dimensionalidade se mostrou eficaz já que os valores das estatísticas AIC, R^2 e REQM, associados aos modelos com número de marcadores reduzidos, se mostraram mais favoráveis indicando ter melhoria de qualidade do modelo ao se beneficiar da redução da dimensionalidade e também a multicolinearidade existentes nos dados originais. Muitos autores tem utilizado com sucesso metodologias de redução de dimensionalidade como (Long et al., 2011a,b; Azevedo et al, 2013;2014; Akidemir et al., 2017). Entretanto, para esses autores os métodos utilizados não eram propriamente de exclusão de marcadores mas de adotar estratégia de reduzir a dimensionalidade do modelo de regressão fazendo uso de outras variáveis explicativas que representavam as melhores combinações lineares dos marcadores disponíveis. Assim, a redução era feita considerando a regressão por componentes principais (PCR), regressão via componentes independentes (ICR) e a regressão parcial de mínimos quadrados (PLSR). Em outras situações, o pesquisador tem realmente o interesse no desenvolvimento de algoritmos capazes de selecionar SNPs mais relacionados com as características de interesse (Long et al., 2007; 2010 ; Akidemir et al., 2017) pelos seus benefícios tanto em modelos de regressão quanto em arquiteturas diversificadas de inteligência computacional.

O sucesso da abordagem do *Stepwise*, ao se considerar os valores da acurácia seletiva (R^2), em relação a metodologia da Sonda se deve ao fato de como os dois métodos selecionam suas variáveis. No caso da Sonda, os coeficientes de determinação e de regressão são contabilizados de acordo com as amostragens aleatórias das combinações das marcas durante as 20 mil combinações realizadas. No final deste processo, que não envolve um critério global de comparação, espera-se encontrar uma solução ótima; porém este número, não tão elevado de buscas (20000 de um total de $\boldsymbol{C_{1000}^{100}}$), pode não garantir que a melhor combinação linear fosse de fato encontrada, já no método *Stepwise* existe um critério fundamentado em correlação (simples ou parcial) que classifica cada marcador para fins de utilização no modelo. Esta classificação não é estática pois, a cada passo, são consideradas probabilidades de entrada e saída das marcas no modelo e associadas ao teste F parcial, e a um *p- value* < 0.05, e pela contabilização do R^2 do modelo, ao incluir cada marcador, é possível chegar ao final do processo com uma porcentagem da variação na variável resposta que os marcadores conseguem explicar mais elevada. Essa estratégia parece conduzir a melhor combinação de variáveis e, por isso, apresenta grande potencial de aplicação em estudos de seleção genômica quando o interesse é maior valor da acurácia seletiva (R^2). Entretanto, se o interesse é reduzir valores do erro quadrático médio, que resultaria em maior acurácia preditiva, o método da *Sonda* se destaca como de grande potencial por proporcionar resultados de REQM relativamente mais baixos. Por fim, verifica-se, em termos de índice de Akaike, as técnicas de redução da dimensionalidade são igualmente eficientes com pequena vantagem para o uso da metodologia Stepwise em relação à Sonda.

Como a redução da dimensionalidade tem propósito de proporcionar maior acurácia de predição de valores genéticos, torna-se necessário uma apreciação dos resultados da Tabela 2 também no contexto biológico. Assim, sob o aspecto genético podemos deduzir que, além das vantagens estatísticas, a redução da dimensionalidade preservou a natureza genética das variáveis pois, de maneira geral, os modelos mais complexos, que incluíam epistasia e, ou, dominância, tiveram queda no desempenho nas três medidas estatística utilizadas em relação aos cenários mais simples como o estabelecido no modelo aditivo (V1 e V7).

No trabalho realizado já havia a expectativa de que a acurácia (seletiva ou preditiva) de predição seria reduzida com o grau de complexidade do modelo (aditivo, aditivo-dominante e epistático como verificado em V1, V2 e V4 para herdabilidade de 30% ou V7, V8 e V10, para herdabilidade de 60%). Isso se deve ao fato de que o modelo RR-BLUP utilizado não ter sido parametrizado de forma a incluir matrizes de incidência

dos efeitos de dominância e da epistasia. Como verificado, a redução da dimensionalidade deveria melhorar a acurácia, mas preservaria esta diferença de desempenho entre os modelos mais e menos complexos pois, em termos de tipos de qualitativos de efeitos, o modelo reduzido era semelhante ao não reduzido.

Um resultado que ainda merece destaque é aquele referente às variáveis V1 e V7 que representam cenários em que o valor genotípico é determinado apenas pela ação aditiva dos genes. Nesta situação não haveria efeitos interativos a ser captados pelo modelo RR-BLUP e a melhoria da redução da dimensionalidade representa, de fato, maior facilidade estatística de estimação em processos iterativos e que necessitam de convergência de valores para obtenção de valores acurados de predição.

Sendo assim, a análise de regressão Stepwise é promissora para redução e seleção de marcadores em problemas de seleção genômica sem que haja perda de informações relevantes, uma vez que o elevado número de marcadores moleculares utilizados nos modelos preditivos de SG torna inviável a avaliação de todos os modelos possíveis (Schuster e Cruz., 2013).

4.3 Predição de valores genéticos por meio de RNA-RBF com e sem redução de dimensionalidade do conjunto de dados explicativos

Segundo Long et al.(2007) e Habier et al.(2007), a utilização de metodologias de seleção de marcadores é essencial para uma boa predição em características de interesse. Entretanto, como vimos no presente trabalho, a complexidade da característica analisada tem grande influência sobre os resultados obtidos. Não adianta simplesmente reduzir o número de informações se o modelo não contemplar, por exemplo, as interações intra e interalélicas determinante do padrão genotípico e fenotípico da variável. Os efeitos de interação podem ser incluídos de forma explícita, como nos modelos de regressão, ou de forma implícita como nas abordagens de redes neurais por meio das associações entre neurônios nas camadas ocultas. E, certamente, parametrizar uma rede neural com mais camadas ocultas e mais neurônios por camada será tarefa mais simples quando o número de entradas (marcadores moleculares) for reduzido.

Em nosso trabalho a redução da dimensionalidade foi utilizada com intuito de selecionar os SNPs relacionados às características de interesse e melhorar também a performance das Redes de Base Radial como sugerido Gonzáles-Camacho et al., (2012) e realizado nos trabalhos anteriores de (Long et al., 2007; 2010; 2011a; b). Assim, Sonda

e *Stepwise* foram as metodologias empregadas com o intuito de que, no modelo preditivo, apenas os SNPs mais relacionados com as características de interesse fossem incluídos. O grande número de marcadores distribuídos no genoma quando incluídos no modelo de predição representa um enorme desafio computacional e compromete a performance da Rede de Base Radial como pode ser visto no resultado da Tabela 3.

A Tabela 3 apresenta o resumo dos resultados de predição realizada por meio da Rede de Base Radial obtidos pela análise comparativa entre os resultados com e sem a utilização do método de redução de dimensionalidade. Novamente, resultados estatísticos e genéticos favoráveis ficam evidentes. Assim, sob aspecto estatístico, o resultado revela que a redução orientada da dimensionalidade traz grandes benefícios ao ajuste de modelo proporcionando melhoria na acurácia seletiva (maior R^2) e melhoria na acurácia preditiva (menor REQM), como já sugerido por Long et al., (2004). Novamente, destacam os resultados referentes às variáveis V1 e V7 que representam cenários em que os valores genotípicos são determinados apenas pela ação aditiva dos genes. Nesta situação não haveria efeitos interativos a ser captados e a melhoria da redução da dimensionalidade representa a maior facilidade estatística de aprendizado e de convergência em arquiteturas de redes com menor número de entradas.

Tabela 3. Resultado da predição de valores genéticos a partir da abordagem por RNA-RBF com (1000 marcadores) e sem (100 marcadores) redução de dimensionalidade obtidos em conjunto de dados de validação envolvendo procedimentos de validação cruzada.

	R²			REQM		
CENARIOS	1000	100 SW	100 S	1000	100 SW	100 S
V1 - D0H30_Ad	0,03 ± 0,12	0,59 ± 0,02	0,24 ± 0,05	5,89 ± 0,12	4,91 ± 0,14	5,65± 0,20
V2 - D0.5H30_Ado	0,12 ± 0,06	0,59 ± 0,03	0,24 ± 0,08	6,17 ± 0,14	5,05 ± 0,13	5,86 ± 0,10
V3 – D1H30_Ado	0,02 ± 0,01	0,56 ± 0,07	0,22 ± 0,09	16,78 ± 1,19	5,05 ± 0,12	15,31 ± 0,67
V4 - D0H30_Ep	0,03 ± 0,00	0,45 ± 0,05	0,23 ± 0,03	16,85 ± 1,17	13,54 ± 0,12	15,58 ± 0,57
V5 – D0.5H30_Ep	0,05 ± 0,02	0,56 ± 0,05	0,30 ± 0,07	18,51 ± 0,55	13,55 ± 0,34	17,13 ± 0,56
V6 – D1H30_Ep	0,07 ± 0,05	0,50 ± 0,05	0,28 ± 0,08	23,84 ± 1,65	14,19 ± 0,34	21,74 ± 0,87
V7-D0H60_Ad	0,38 ± 0,08	0,79 ± 0,03	0,52 ± 0,08	4,52 ± 0,12	19,95 ± 0,44	4,24 ± 0,11
V8-D0.5H60_Ado	0,34 ± 0,07	0,74 ± 0,03	0,43 ± 0,03	4,67 ± 0,19	3,49 ± 0,06	4,51± 0,11
V9-D1H30_Ado	0,18 ± 0,04	0,64 ± 0,02	0,27 ± 0,02	5,40 ± 0,16	3,71 ± 0,05	5,21 ± 0,18
V10-D0H60_Ep	0,06 ± 0,03	0,58 ± 0,05	0,31 ± 0,06	13,54 ± 0,30	4,52 ± 0,15	12,64 ± 0,30
V11-D0.5H60_Ep	0,10 ± 0,02	0,62 ± 0,04	0,37 ± 0,07	15,48 ± 0,46	10,94 ± 0,29	14,05 ± 0,27
V12-D1H60_Ep	0,13 ± 0,03	0,58 ± 0,08	0,42 ± 0,07	18,83 ± 0,75	15,68 ± 0,55	17,11 ± 0,56

A melhora dos valores de predição de RBF, com a redução da dimensionalidade,

era esperada já que os trabalhos desenvolvidos por alguns autores. Assim, González-Camacho et al. (2012), utilizando que RNA-RBF e espaço de Kernel, sugeriram que é possível melhorar a predição dos modelos não paramétricos utilizando alguma metodologia que não inclua todos os SNPs não relacionados às características de interesse. Long et al. (2010), utilizando simulação de apenas 8 SNPS, introduziram pesos diferenciais entre os SNPs conforme (Long et al., 2007) e melhoraram as estimativas de RBFs em comparação com uma outra RNA-RBF em que todos os SNPs foram utilizados sem a inclusão de pesos. Outros autores relataram que os resultados preditivos favoreceram o RNA-RBF sobre um modelo BayesA (Meuwissen et al., 2001), para uma série de herdabilidades.

Sob aspecto genético, a redução da dimensionalidade também revela que os efeitos importantes sobre o mecanismo de ação dos caracteres, tais como dominância e epistasia, também estão presentes e preservados como resultados na análise dos dados. A presença destes efeitos é captada pela RNA-RBF que extrai as interações entre os SNPs dos próprios indivíduos e, assim, espera-se que capture a complexidade genética do caráter pelas associações entre neurônios em sua camada oculta.

Alguns autores relatam o sucesso da RNA-RBF em capturar as relações não lineares existentes entre os marcadores, como no trabalho desenvolvido por Long et al. (2011). Os autores, utilizando simulação de caracteres quantitativos sob diferentes modos de ação de genes (aditividade, dominância e epistasia), descobriram que a RNA-RBF tinha melhor habilidade de prever o mérito dos indivíduos em futuras gerações na presença de efeitos não aditivos do que um modelo aditivo linear tal como o Lasso bayseano. No caso de efeitos de genes puramente aditivos, a RNA-RBF foi ligeiramente pior do que o Lasso. Ainda neste trabalho, os autores relatam o emprego da metodologia de redução de dimensionalidade, do tipo componentes principais, antes de utilizar RNA-RBF e também mostraram que, com a seleção de marcadores, o desempenho da rede de base radial era maior.

4.4. Predição de valores genéticos por diferentes abordagens (GS, RNA e RBF) com redução de dmensionalidade do conjunto de dados explicativos

A escolha de um método ótimo de predição de valores genéticos deve basear-se no critério de uma inferência mais precisa e realista possível, a qual deve ser avaliada segundo parâmetros estatísticos adequados. Ignorar os efeitos genéticos não aditivos tem várias conseqüências, incluindo estimativas incorretas da variância genética aditiva

(Palucci et al., 2007) e menor acurácia na predição de fenótipos (Lee et al., 2008), dentre muitos outros aspetos. Como apontado primeiramente por Gianola et al. (2006) e, posteriormente, por Long et al. (2010), os modelos não paramétricos não impõem suposições fortes sobre a relação fenótipo-genótipo e eles têm o potencial de capturar interações entre locos pelas interações existentes entre neurônios de diferentes camadas (camada de entrada e ocultas).

A inclusão dos efeitos não lineares e, ou, multiplicativos, podem ser feitos de forma explícita em vários métodos paramétricos como RR-BLUP, Bayes A, B, e BLASSO tornando-os muito complexos pela hiperparametrização ou de forma implícita, como nos métodos semi e não paramétricos como Kernel, máquina de vetor de suporte e redes neurais artificias (RNA).

No nosso trabalho tomou-se por base as estatísticas R^2, que mede a acurácia seletiva, e a estatística REQM, que mede a acurácia preditiva, e obteve-se os resultados relacionados na Tabela 4. A utilização das redes neurais, Redes de Base Radial e Perceptron Multicamadas, se mostrou igualmente eficaz apresentando valores muito semelhantes de R^2 e de REQM entre si, e com resultados mais favoráveis que o obtido para o RR-BLUP.

De maneira geral, uma das vantagens das redes neurais é que não é necessário fazer pressupostos a priori sobre a distribuição dos dados e dos resíduos e tem uma grande flexibilidade para lidar com diferentes tipos de efeitos não aditivos complexos, como a epistasia (Gianola et al., 2011; Perez-Rodriguez et al., 2012; Felipe et al., 2014; Howard et al., 2014). Isso se deve ao fato das redes neurais artificiais (RNA) fornecerem uma alternativa de análise, pois atuam como aproximadores universais de funções complexas e podem capturar relações não-lineares entre preditores e respostas, com a interação entre as variáveis aprendidas de forma adaptativa (Gianola et al., 2011). RNAs são candidatas interessantes para a análise de caracteres afetados por ação genética com efeitos epistáticos.

Destacam-se, na Tabela 4, os valores de erro quadrático médio muito pequenos (REQM <20) para cenários com dominância e epistáticos ao se utilizar as técnicas de inteligência computacional para análise de dados. Em relação a acurácia de predição (REQM) obtida pelo RR-BLUP os valores variam de 90 a 428 sendo maiores conforme a complexidade do cenário aumenta.

Observa-se (Tabela 4) que, para o cenário que apresenta dominância completa e epistasia (cenário mais complexo), a eficácia do processo medida pela acurácia seletiva R^2 ainda é de 50% tendo uma redução de apenas 10% em relação ao cenário em que a

expressão genotípica é dada unicamente pelo modelo aditivo (mais simples). Em relação ao RR-BLUP observamos uma discrepância mais acentuada de 15% na eficácia de predição do cenário mais complexo em relação ao mais simples, mesmo assim para todos os cenários a herdabilidade da característica é superada pela estimativa de R^2 obtido.

Tabela 4. Resultado da predição de valores genéticos a partir das abordagens por GS, RNA-RBF e RNA-MLP com (100 marcadores) redução de dimensionalidade. (*Stepwise*).

	R²			REQM		
CENARIOS	RBF	PML	RRBLUP	RBF	RNA-MLP	RRBLUP
V1 - D0H30_Ad	0,59 ± 0,02	0,57 ± 0,03	0,57 ± 0,03	4,79 ± 0,13	4,91 ± 0,03	90,23 ± 0,03
V2 - D0.5H30_Ado	0,59 ± 0,03	0,59 ± 0,04	0,58 ± 0,05	5,03 ± 0,11	5,05 ± 0,12	138,20± 3,66
V3 – D1H30_Ado	0,56 ± 0,07	0,52 ± 0,07	0,54 ± 0,06	13,17± 0,23	13,55 ± 0,34	291,61 ± 17,73
V4 - D0H30_Ep	0,45 ± 0,05	0,47 ± 0,07	0,42 ± 0,05	14,21 ± 0,44	14,19 ± 0,55	442,59 ± 48,10
V5 – D0.5H30_Ep	0,56 ± 0,05	0,54 ± 0,04	0,54 ± 0,06	15,19 ± 0,20	15,47 ± 0,22	249,37 ± 18,94
V6 – D1H30_Ep	0,50 ± 0,05	0,44 ± 0,05	0,41 ± 0,04	19,96 ± 0,45	20,73 ± 0,29	427,91 ± 40,57
V7-D0H60_Ad	0,79 ± 0,03	0,78 ± 0,03	0,79 ± 0,03	3,44 ± 0,13	3,49 ± 0,06	71,38 ± 0,98
V8-D0.5H60_Ado	0,74 ± 0,04	0,74 ± 0,03	0,73 ± 0,03	3,71 ± 0,11	3,71 ± 0,05	107,40 ± 1,06
V9-D1H30_Ado	0,64 ± 0,02	0,59 ± 0,04	0,64 ± 0,01	4,37± 0,07	4,52 ± 0,15	145,95 ± 5,18
V10-D0H60_Ep	0,58 ± 0,05	0,59 ± 0,03	0,57 ± 0,04	11,14 ± 0,28	10,94 ± 0,29	320,11± 36,36
V11-D0.5H60_Ep	0,62 ± 0,04	0,60 ± 0,06	0,59 ± 0,05	12,42 ± 0,45	12,52 ± 0,55	280,93 ± 28,96
V12-D1H60_Ep	0,58 ± 0,08	0,59 ± 0,08	0,58 ± 0,09	15,74 ± 0,72	15,67 ± 0,55	473,80 ± 22,04

O fato de as estimativas de R^2 terem sido semelhantes entre o RR-BLUP e as redes neurais em cenários aditivos também foi observado por muitos autores utilizando RR-BLUP ou métodos baysianos. Por exemplo, Gloria et al. (2016), utilizando simulação com duas característica quantitativa de natureza aditiva, avaliaram o desempenho de seis redes neurais baysianas, RR-BLUP e BLASSO. Como esperado, os desempenhos foram semelhantes. A rede neural mais simples superou outras redes mais parametrizadas, em termos de capacidade preditiva, e forneceu as melhores estimativas de h^2.

Tusell et al. (2013) compararam o espaço de Kernel e duas redes neurais diferentes com alguns modelos de regressão linear (RR-BLUP, LASSO, G-BLUP), encontrando uma habilidade preditiva igual ou melhor para os métodos de aprendizado da máquina.

Pérez-Rodríguez, et al. (2012), utilizando 306 linhagens elite de trigo do genotipado com 1717 marcadores DArT e dois caracteres, dias até antese/florescimento e rendimento de grãos, medidos em cada um dos 12 ambientes no CIMMYT*(International Maize and Wheat Improvement Center),* compararam os modelos paramétricos(LASSO, RR-BLUP Bayseano (BRR), Bayes A e Bayes B) e não-paramétricos (espaço de Kernel e redes neurais regularizadas bayesiana e redes de base

radial. Verificaram que os três modelos não lineares apresentaram uma superioridade consistente.

Gianola et al.(2011), utilizando dados de trigo representavam 599 linhagens genotipada com 1.279 marcadores, verificaram que as redes neurais superaram os modelos lineares na capacidade preditiva e os resultados obtidos foram comparados com diversos resultados de trabalhos anteriores realizado por seu grupo de pesquisa e os levaram a concluir que as redes neurais baysianas para trigo são, pelo menos, tão boas quanto os métodos baysianos na análise dos mesmos dados em outros dois estudos incluindo predições realizadas.

Por fim, destaca-se o trabalho de Howard et al. (2014) que comparam 14 métodos paramétricos e não paramétricos que foram desenvolvidos para fins de predição de fenótipos. Estes autores analisaram os resultados obtidos por métodos paramétricos, incluindo método de quadrados mínimos, RR-BLUP, BRR, LASSO, BLASSO, BLUP, Bayes A, Bayes B, Bayes C e Bayes Cpi. Também analisamos métodos não paramétricos, incluindo o estimador de Nadaraya-Watson, reproduzindo o espaço de Kernel, máquina vetor de suporte e as redes neurais usando arquiteturas genéticas simuladas aditivas e epistáticas em uma população F_2. Os autores concluíram que os métodos paramétricos não conseguiram predizer valores fenotípicos quando a arquitetura genética se baseava inteiramente na epistasia. Os métodos paramétricos foram ligeiramente melhores que os métodos não paramétricos para arquiteturas genéticas aditivas. Os autores destacaram também que a herdabilidade teve o segundo maior impacto nas estimativas de acurácia e erro quadrático médio.

Dessa forma podemos considerar a utilização de inteligência artificial no melhoramento genético como uma estratégia alternativa eficaz e confiável. Para espécies em que o valor genético do indivíduo é muito importante como é o caso de clones, animais e espécies de propagação vegetativa esse baixo erro de predição torna-se ainda mais atrativo. Embora as redes neurais do tipo RNA-MLP e RNA-RBF sejam menos utilizadas elas se mostraram igualmente importante e capazes de predições acuradas em cenários de dominância ou epistasia.

5.CONCLUSÕES

Os resultados das análises evidenciam a possibilidade que o pesquisador tem para a tarefa de reduzir o número de variáveis explicativas e, também, garantir que em estudos por outras abordagens não enfrente problemas de multicolinearidade e de dimensionalidade, sem que haja perda de informações genéticas relevantes tal como a influência da dominância e da epistasia. Acredita-se que, com a utilização de procedimento de seleção de variáveis, as técnicas fundamentadas em inteligência computacional possam ser facilmente empregadas sem demandar recursos computacionais sofisticados.

Além disso, pelos resultados obtidos as redes neurais do tipo perceptron de múltiplas camadas (RNA-MLP) ou rede de bases radial (RBF-RNA) são igualmente recomendáveis para a predição do valor genético principalmente em caracteres muito influenciados por efeitos da epistasia e, ou, da dominância.

6.REFERÊNCIAS BIBLIOGRÁFICAS

Akaike, H. 1974.A new look at the statistical model identification **IEEE Transaction on Automatic Control.** v.19, p.716-723.

Akdemir, D., Jannink, J.L. and Isidro-Sánchez, J., 2017. Locally epistatic models for genome-wide prediction and association by importance sampling. **Genetics Selection Evolution**, *49*(1), p.74.

Azevedo, C.F., de Resende, M.D.V., Fonseca, F., Lopes, P.S. and Guimarães, S.E.F., 2013. Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. **Pesquisa Agropecuária Brasileira**, *48*(6), pp.619-626.

Azevedo, C.F., Silva, F.F., Resende, M.D., Lopes, M.S., Duijvesteijn, N., Guimarães, S.E.F., Lopes, P.S., Kelly, M.J., Viana, J.M.S. and Knol, E.F., 2014. Supervised independent component analysis as an alternative method for genomic selection in pigs. **Journal of Animal Breeding and Genetics**, *131*(6), pp.452-461.

Bengio, Y. and Grandvalet, Y., 2004. No unbiased estimator of the variance of k-fold cross-validation. **Journal of machine learning research**, *5*(Sep), pp.1089-1105.

Bishop, C.M., 2006. **Pattern Recognition and Machine Learning**. Springer, New York.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J.M., 2009. Predicting quantitative traits with regression models for dense molecular markers. **Genetics** 182, 375–385.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., Camacho-González, J.M., Pérez-Elizalde, S., Beyene, Y. and Dreisigacker, S., 2017. Genomic selection in plant breeding: Methods, models, and perspectives. **Trends in plant science.**

Cruz, C.D. 2016.Genes Software – extended and integrated with the R, Matlab and Selegen. **Acta Scientiarum**. Agronomy. Maringá, v. 38, n. 4, p. 547-552, Oct.-Dec.,.

Dreisigacker, S., 2017. Genomic selection in plant breeding: Methods, models, and perspectives. **Trends in plant science**.

Dekkers, J.C., 2002. Multifactorial genetics: The use of molecular genetics in the improvement of agricultural populations. **Nature Reviews Genetics**, 3(1), p.22.

Felipe, V.P., Okut, H., Gianola, D., Silva, M.A., Rosa, G.J., 2014. Effect of genotype imputation on genome-enabled prediction of complex traits: an empirical study with mice data. **BMC Genetics**. 15, 149.

Fisher, R.A., 1918. 009: **The Correlation Between Relatives on the Supposition of Mendelian Inheritance.**

Gianola, D., Fernando, R.L. and Stella, A., 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. **Genetics**, 173(3), pp.1761-1776.

Gianola, D. and De Los Campos, G., 2008. Inferring genetic values for quantitative traits non-parametrically. **Genetics Research**, *90*(6), pp.525-540.

Gianola, D., Okut, H., Weigel, K.A. and Rosa, G.J., 2011. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. **BMC genetics**, 12(1), p.87.

Glória, L.S., Cruz, C.D., Vieira, R.A.M., de Resende, M.D.V., Lopes, P.S., de Siqueira, O.H.D. and e Silva, F.F., 2016. Accessing marker effects and heritability estimates from genome prediction by Bayesian regularized neural networks. **Livestock Science**, 191, pp.91-96.

González-Camacho, J.M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J.E., Mahuku, G., Babu, R. and Crossa, J., 2012. Genome-enabled prediction of genetic values using radial basis function neural networks. **Theoretical and Applied Genetics**, *125*(4), pp.759-771.

González-Recio, O., Rosa, G.J. and Gianola, D., 2014. Machine learning methods and predictive ability metrics for genome-wide prediction of complex traits. **Livestock Science**, *166*, pp.217-231.

González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, L. and Gianola, D., 2016. Genome-enabled prediction using probabilistic neural network classifiers. **BMC genomics**, *17*(1), p.208.

Habier, D., Fernando, R.L. and Dekkers, J.C.M., 2007. The impact of genetic relationship information on genome-assisted breeding values. **Genetics**, *177*(4), pp.2389-2397.

Habier, D., Fernando, R.L. and Dekkers, J.C., 2009. Genomic selection using low-density marker panels. **Genetics**.

Horne, B.D. and Camp, N.J., 2004. Principal component analysis for selection of optimal SNP-sets that capture intragenic genetic variation. **Genetic epidemiology**, 26(1), pp.11-21.

Holland J., 1975. **Adaption in Natural and Artificial Systems**. MIT Press, 1975.

Howard R., Carriquiry A. L., Beavis W. D. 2014.Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. **G3: Genes, Genomes, Genetics,** 2014, 4(6): 1027–1046

James, G., Witten, D., Hastie, T. and Tibshirani, R., 2013. **An introduction to statistical learning** (Vol. 112). New York: springer.

Lee, S.H., van der Werf, J.H., Hayes, B.J., Goddard, M.E. and Visscher, P.M., 2008. Predicting unobserved phenotypes for complex traits from whole-genome SNP data. **PLoS genetics**, *4*(10), p.e1000231.

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A. and Avendano, S., 2007. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. **Journal of animal breeding and genetics**, 124(6), pp.377-389.

Long, N., Gianola, D., Rosa, G.J., Weigel, K.A., Kranis, A. and Gonzalez-Recio, O., 2010. Radial basis function regression methods for predicting quantitative traits using SNP markers. **Genetics research**, 92(3), pp.209-225.

Long N, Gianola D, Rosa GJ, Weigel KA. Marker-assisted prediction of non-additive genetic values. **Genetica**. 2011a. Jul 1;139(7):843-54.

Long, N., Gianola, D., Rosa, G.J.M. and Weigel, K.A., 2011b. Dimension reduction and variable selection for genomic selection: application to predicting milk yield in Holsteins. **Journal of animal breeding and genetics**, *128*(4), pp.247-257.

Mackay, T.F., Stone, E.A. and Ayroles, J.F., 2009. The genetics of quantitative traits: challenges and prospects. **Nature Reviews Genetics**, *10*(8), p.565.

Macciotta N., Gaspa G., Steri R., Pieramati C., Carnier P., Dimauro C. , 2009. Preselection of most significant SNPs for the estimation of genomic breeding values. **BMC proceedings**, 3(Suppl 1), S14.

Matlab (2010). Matlab Version 7.10.0. Natick, Massachusetts: The Math Works Inc.

Meuwissen T. H. E., Hayes B. J., Goddard M. E., 2001. Prediction of total genetic value using genome-wide dense marker maps. **Genetics**, 157(4): 1819–1829.

Montgomery, D.C.; Peck, E.A. **Introduction to linear regression analysis**. New York: J. Wiley, 1982. 504p.

Moser, G., B. Tier, R. E. Crump, J. Soelkner, K. R. Zenger et al., 2007. **Estimation of molecular breeding values in genome wide selection using supervised dimension reduction based on partial least squares**, Toulouse, France, edited by A. Legarra.

Ornella, L., Perez, P., Tapia, E., Gonzalez-Camacho, J.M., Burgueno, J., Zhang, X., Singh, S., Vicente, F.S., Bonnett, D., Dreisigacker, S., Singh, R., Long, N., Crossa, J., 2014. Genomic-enabled prediction with classification algorithms. **Heredity** 112, 616–626.

Palucci, V., Schaeffer, L.R., Miglior, F. and Osborne, V., 2007. Non-additive genetic effects for fertility traits in Canadian Holstein cattle (Open Access publication). **Genetics Selection Evolution**, 39(2), p.181.

Piyasatian, N., Fernando, R.L. and Dekkers, J.C.M., 2007. Genomic selection for marker-assisted improvement in line crosses. **Theoretical and Applied Genetics,** *115*(5), pp.665-674.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J.M., Crossa, J., Manès, Y. and Dreisigacker, S., 2012. Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. **G3:Genes, Genomes, Genetics**, *2*(12), pp.1595-1605.

Pimentel, E.C., König, S., Schenkel, F.S. and Simianer, H., 2009. Comparison of statistical procedures for estimating polygenic effects using dense genome-wide marker data. In **BMC proceedings,** 3: 1, p. S12). BioMed Central.

R CORE TEAM. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2017. Available at: <https://www.R-project.org/>.

Resende, M.D.V.; Silva, F.F.; Azevedo, C.F. **Estatística matemática, biométrica e computacional: Modelos Mistos, Multivariados, Categóricos e Generalizados (REML/BLUP), Inferência Bayesiana, Regressão Aleatória, Seleção Genômica, QTL-GWAS, Estatística Espacial e Temporal, Competição Sobrevivência.** Viçosa: Suprema, 881p. 2014.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. **Nature** 405, 847–856.

Schuster, I.; Cruz, C.D., **2004. Estatística genômica aplicada a populações derivadas de cruzamentos controlados.** Viçosa: UFV. 568p.

Tier, B., J. Cavanagh, R. Crump, M. Khatkar, G. Moser et al., 2007 **Genome wide selection: experiences from the Australian Dairy** Industry. The 3rd International Conference on Quantitative Genetics, Hangzhou, China.

Tusell, L., Pérez-Rodríguez, P., Forni, S., Wu, X.L., Gianola, D., 2013. **Genome-enabled methods for predicting litter size in pigs: a comparison.** Animal 7, 1739–1749.

Usai M.G., Goddard M.E., Hayes B.J., 2009. LASSO with cross-validation for genomic selection. **Genetics Research,** 91, 427–436.

Vazquez A.I., Rosa G.J.M., Weigel K.A., de los Campos G., Gianola D., Allison D.B., 2010. Predictive ability of subsets of single nucleotide polymorphisms with and without parent average in US Holsteins. **Journal of Dairy Science**, 93, 5942–5949.

Yamamoto, A., Zwarts, L., Callaerts, P., Norga, K., Mackay, T.F.C. e Anholt, R.R.H., 2008. **Neurogenetic networks for startle-induced locomotion in Drosophila melanogaster.** Proceedings of the National Academy of Sciences of the USA 105, 12393–12398.

Woolaston, A. F., B. Tier and R. D. Murison, 2007. **Principal components regression of SNP data to predict genetic merit.** Papers and Abstracts From the Workshop on QTL and Marker-Assisted Selection, Toulouse, France, edited by A. Legarra.

Weigel K., de los Campos G., Vazquez A., Rosa G., Gianola D., Tassell C.V., 2010a Accuracy of direct genomic values derived from imputed single nucleotidepolymorphism genotypes in Jersey cattle. **Journal of Dairy Science**, 93, 5423–5435.

Weigel K., Tassell C. V., O'Connell J., VanRaden P., Wiggans G. 2010b., Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms**. Journal of Dairy Science**,93, 2229–2238.

Weigel, K.A., Van Tassell, C.P., O'Connell, J.R., VanRaden, P.M. and Wiggans, G.R., 2010. Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. **Journal of Dairy Science,** *93*(5), pp.2229-2238.

CONCLUSÕES GERAIS

Muitos são os desafios existentes nos programas de melhoramento de plantas de caracteres quantitativos. Com o advento da genética molécular, o principal benefício do melhoramento de plantas é a possibilidade de utilização direta das informações de DNA que tem contribuído para que a seleção de genótipos superiores possa ser realizada de forma mais acurada e eficiente. Entretanto, para isso, é preciso que essas informações sejam utilizadas da forma mais completa e realística possível levando em consideração as peculiaridades da arquitetura dos caracteres complexos que muitas vezes exige uma modelagem estatística avançada e computacionalmente intensiva.

Desta forma, o objetivo deste trabalho no primeiro capítulo foi avaliar a eficiência da seleção genômica (SG) e das redes neurais artificias do tipo de base radial (RNA-RBF) na predição do valor genético em população natural com desequilíbrio gamético e os resultados obtidos mostram que, na presença de uma complexa relação genótipo-fenótipo (isto é, não linearidade e não aditividade), os modelos RNA-RBF superaram um modelo aditivo linear, RR-BLUP, na predição de valores genéticos totais de caracteres quantitativos usando marcadores SNP.

Ao lidar com um número grande número de marcadores, a demanda computacional da RNA-RBF foi intensiva sugerindo a utilização de uma seleção de variáveis como uma solução interessante para melhoria do processo preditivo a ser abordados. Para tanto, no segundo capitulo, testamos se a utilização de procedimento de seleção de variáveis poderia facilitar a predição do valor genéticos de forma que os resultados obtidos após a redução do número de marcadores preservasse as mesmas informações genéticas. Para tanto os métodos de seleção genômica RR-BLUP e as neurais artificias do tipo de base radial (RNA-RBF) e Perceptron de Múltiplas camadas (RNA-MLP) foram utilizados na predição do valor genético após a redução do número de marcadores. Os resultados das análises evidenciam a possibilidade que o pesquisador tem para a tarefa de reduzir o número de variáveis explicativas e, também, garantir que em estudos por outras abordagens não enfrente problemas de multicolinearidade e de dimensionalidade, sem que haja perda de informações genéticas relevantes tal como a influência da dominância e da epistasia. Acredita-se que, com a utilização de procedimento de seleção de variáveis, as técnicas fundamentadas em inteligência computacional possam ser facilmente empregadas sem demandar recursos computacionais sofisticados sendo a RNA-RBF igualmente eficiente a RNA-MLP.

Anexos

Resultado da predição de valores genéticos a partir das abordagens por GS, RNA-RBF e RNA-MLP com (100 marcadores) redução de dimensionalidade usando Sonda.

	R^2			REQM		
CENÁRIOS	RBF	RNA	RRBLUP	RBF	RNA	RRBLUP
V1 - D0H30_Ad	0,24 ± 0,05	0,18 ± 0,04	0,27 ± 0,05	5,65 ± 0,20	5,75 ± 0.14	37,08 ± 3,98
V2 - D0.5H30_Ad	0,24 ± 0,08	0,20 ± 0,05	0,26 ± 0,10	5,86 ± 0,10	5,99 ± 0.05	66,56 ± 3,35
V3 – D1H30_Ad	0,22 ± 0,09	0,23 ± 0,08	0,23 ± 0, 07	15,31 ± 0,67	15,48 ± 0.76	241,53 ± 11,65
V4 - D0H30_Ep	0,23 ± 0,03	0,21 ± 0,04	0,22 ± 0,04	15,59 ± 0,57	16,05 ± 0.82	244,12 ± 27,19
V5 – D0.5H30_Ep	0,30 ± 0,07	0,56 ± 0,05	0,30 ± 0,07	17,13 ± 0,56	15,19 ± 0.20	300,23 ± 34,4
V6 – D1H30_Ep	0,28 ± 0,08	0,24 ± 0,06	0,25 ± 0,05	21,74 ± -0,87	22,00 ± 0.81	478,29 ± 27,75
V7-D0H60_Ad	0,52 ± 0,08	0,52 ± 0,07	0,55 ± 0,09	4,25 ± 0,11	4,27 ± 0.11	48,27 ± 2,74
V8-D0.5H60_Ad	0,44 ± 0,03	0,44 ± 0,05	0,47 ± 0,04	4,51± 0,11	4,52 ± 0.12	72,96 ± 4,71
V9-D1H30_Ad	0,28 ± 0,02	0,28 ± 0,05	0,32 ± 0,05	5,21 ± 0,18	5,23 ± 0.22	87,16 ± 10,27
V10-D0H60_Ep	0,310 ± 0,06	0,30 ± 0,05	0,32 ± 0,08	12,64 ± 0,30	12,53 ± 0.20	154,14 ± 5,11
V11-D0.5H60_Ep	0,365 ± 0,07	0,31 ± 0,09	0,34 ± 0,09	14,05 ± 0,27	14,46 ± 0.35	243,35 ± 18,86
V12-D1H60_Ep	0,419 ± 0,07	0,32 ± 0,05	0,38 ± 0,05	17,11 ± 0,56	17,87 ± 0.65	306,43 ± 31.74