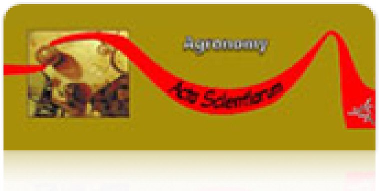

Acta Scientiarum. Agronomy
ISSN: 1679-9275
ISSN: 1807-8621
Editora da Universidade Estadual de Maringá - EDUEM

Sant’Anna, Isabela de Castro; Silva, Gabi Nunes; Nascimento, Moysés; Cruz, Cosme Damião

Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks

Acta Scientiarum. Agronomy, vol. 43, e46307, 2021, January-December
Editora da Universidade Estadual de Maringá - EDUEM

DOI: https://doi.org/10.4025/actasciagron.v43i1.46307

Acta Scientiarum

http://periodicos.uem.br/ojs
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v43i1.46307

GENETICS AND PLANT BREEDING

Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks

Isabela de Castro Sant'Anna[1*], Gabi Nunes Silva[2], Moysés Nascimento[1] and Cosme Damião Cruz[1,3]

[1]Departamento de Estatística, Laboratório de Bioinformática, Universidade Federal Viçosa, Avenida P.H. Rolfs, s/n, 36570-900, Viçosa, Minas Gerais, Brazil. [2]Departamento de Matemática e Estatística, Universidade Federal de Rondônia, Ji-Paraná, Rondônia, Brazil. [3]Departamento de Biologia Geral, Laboratório de Bioinformática, Universidade Federal Viçosa, Viçosa, Minas Gerais, Brazil. *Author for correspondence. E-mail: isabelacsantanna@gmail.com. Orcid-id: https://orcid.org/0000-0002-6944-5361

ABSTRACT. This paper aimed to evaluate the effectiveness of subset selection of markers for genome-enabled prediction of genetic values using radial basis function neural networks (RBFNN). To this end, an F1 population derived from the hybridization of divergent parents with 500 individuals genotyped with 1000 SNP-type markers was simulated. Phenotypic traits were determined by adopting three different gene action models – additive, additive-dominant, and epistatic, representing two dominance situations: partial and complete with quantitative traits having a heritability (h^2) of 30 and 60%; traits were controlled by 50 loci, considering two alleles per locus. Twelve different scenarios were represented in the simulation. The stepwise regression was used before the prediction methods. The reliability and the root mean square error were used for estimation using a fivefold cross-validation scheme. Overall, dimensionality reduction improved the reliability values for all scenarios, specifically with h^2 =30 the reliability value from 0.03 to 0.59 using RBFNN and from 0.10 to 0.57 with RR-BLUP in the scenario with additive effects. In the additive dominant scenario, the reliability values changed from 0.12 to 0.59 using RBFNN and from 0.12 to 0.58 with RR-BLUP, and in the epistasis scenarios, the reliability values changed from 0.07 to 0.50 using RBFNN and from 0.06 to 0.47 with RR-BLUP. The results showed that the use of stepwise regression before the use of these techniques led to an improvement in the accuracy of prediction of the genetic value and, mainly, to a large reduction of the root mean square error in addition to facilitating processing and analysis time due to a reduction in dimensionality.

Keywords: neural networks; genomic prediction; stepwise regression.

Received on January 21, 2019.
Accepted on September 30, 2019.

Introduction

One of the major challenges facing genetic breeding today is understanding the genetic variation of quantitative trait loci (QTL), which are conditioned by a large number of genes with small effects whose interactions often result in nonlinearity in relations between phenotypes and genotypes (Long et al., 2010; Mackay, Stone, & Ayroles, 2009).

With the advent of genomic selection (GS) (Meuwissen, Hayes, & Goddard, 2001), it became possible to estimate the genomic value of individuals (GEBV) without the need for phenotyping, which led to an increase in genetics gain by reducing time and money. Therefore, for many traits of agronomic importance, genetic values are determined by multiple genes of small effects, and their phenotypic expression is strongly affected by genetic interactions between their additive, dominant and epistatic effects. However, most applications of GS include only the additive portion of the genetic value; therefore, a more realistic representation of the genetic architecture of quantitative traits should include dominance and epistatic interactions, since these effects are crucial factors to increase the accuracy of prediction (Akidemir, Jannink, & Isidro-Sanchez, 2017).

The inclusion of these interactions is computationally challenging and leads to the superparametrization of the models that are already in high dimensionality because of the large number of markers in the genome and the smallest number of individuals (Long et al., 2010). Before fitting the model, it is necessary to define the model effects to be estimated. In this context, artificial neural networks (ANNs) have strong potential because they can capture nonlinear relationships between markers from the data themselves (without a previous model definition), which most of the models commonly used in the GS cannot do (Long et al., 2010; Long, Gianola, Rosa, & Weigel, 2011a; Howard, Carriquiry, & Beavis, 2014).

Radial basis function neural networks (RBFNN) are a particular class of artificial neural networks (ANNs) that have properties that make them attractive to GS applications. According to Gianola, Okut, Weigel, and Rosa (2011), RBFNN have the ability to learn from the data used in their training and provide a unique solution and are faster than standard ANNs (Gonzales-Camacho et al., 2012). RBFNN have been used in GS applications for many authors (Gianola et al., 2011; González-Camacho et al., 2012; Pérez-Rodríguez et al., 2012; Long et al., 2010; Long, Gianola, Rosa, & Weigel, 2011ab; Cruz & Nascimento, 2018; Sant'Anna et al., 2019), suggesting that they have a good ability to deal with spatial interactions in comparison with semiparametic and linear regressions.

However, the inclusion of all markers in the RBFNN prediction model increases the chances of a high correlation between the markers (Crossa et al., 2017). The number of markers represents a considerable challenge that leads to less precision and a great computational demand for ANN training. This occurs because RBFNN use a good portion of their resources to represent irrelevant portions of the search space and compromise the learning process because there are thousands of markers available in the genome (Long et al., 2011b). Thus, a more realistic model should include only SNPs related to the traits of interest (Long et al., 2011a).

For this reason, a subset of SNPs can be used for training, since by reducing the search space, RBFNN improve the learning process and increase the predictive power of the model. This procedure was realized by Long et al. (2010), who used two types of RBFNN models: one considering a common weight parameter for each SNP and another in which each SNP has specific parameters of importance. The results showed that an RBFNN trained with specific parameters of importance improved the prediction ability.

However, due to the importance of ANNs for the improvement of the prediction of quantitative traits, there is still a need to test different dimensionality reduction methods and prediction models for polygenic traits. Therefore, this paper aimed to evaluate the efficiency of subset selection of markers to apply genome-enabled prediction by radial basis function neural networks in quantitative traits. The results were compared with those obtained by one of the standard GS models: Ridge Regression BLUP (RR-BLUP).

Material and methods

Origin of populations

To assess the reliability of GS prediction, data were simulated by considering a diploid species with 2n = 20 chromosomes as a reference, and the total length of the genome was set to 1.000 cM. Genomes were generated with a saturation level of 101 molecular markers spaced by 1 cM per linkage group, totaling 1010 markers. Divergent parental line genomes were simulated in the Hardy-Weinberg equilibrium population, as well as genomes from their cross F_1 with 500 individuals. The effective size of the base population used in this study F_1 is the size of F_1 itself, since F_1 was derived from two contrasting homozygous parents.

Simulation of quantitative traits

Quantitative traits were simulated by considering three degrees of dominance (d/a = 0, 0.5, and 1), two broad sense heritability levels (h^2 = 30 and 60), representing three gene actions: additive, dominance and epistatic, thereby totaling twelve scenarios (Table 1). Each trait was controlled by 50 randomly chosen loci with 2 alleles per locus.

Table 1. Simulated scenarios composed of a combination of traits and an action genetic model, heritability and dominance degree.

Scenarios	Heritability (%)	Model	dominance
V1-D0H30_Ad	30	additive	0
V2-D0.5H30_Ado	30	additive-dominant	0.5
V3-D1H30_Ado	30	additive-dominant	1
V4 - D0H30Ep	30	epistatic	0
V5 - D0.5H30Ad	30	epistatic	0.5
V6 - D1H30Ep	30	epistatic	1
V7 - D0H60Ad	60	additive	0
V8 - D0.5H60Ado	60	additive-dominant	0.5
V9 - D10H30Ado	60	additive-dominant	1
V10 - D0H60Ad	60	epistatic	0
V11 - D1H60Ado	60	epistatic	0.5
V12 - D1H60Ep	60	epistatic	1

The genotypic value for the monogenic model is defined by u + a, u + d, u - a for the genotypes AA, Aa e aa, respectively. In a polygenic model, the total genotypic value expressed by a given individual belonging to the population was the sum of each additive effect of the individual locus estimated by the following expression:

$$\mathrm{G}_i = \mu + \mathrm{a}_i + \mathrm{d}_i \qquad (1)$$

where the additive effect (a) of each locus is one-half the difference in mean phenotype between the two homozygous genotypes (for each individual i). The dominance effect (d) is the difference between the mean phenotype of the heterozygous genotype and the average phenotype of the two homozygous genotypes.

The phenotypic value of the ith individual was obtained according to the additive model as follows:

$$Y_i = \mu + \sum_{j=1}^{50} p_j \alpha_j + E_i \qquad (2)$$

where: α_j is the effect of the favorable allele at locus *j*, with codes 1, 0 or -1 for the genotypic classes AA, Aa and aa, respectively, and p_j is the contribution of locus *j* to the manifestation of the trait under consideration. In this study, the contribution of each locus was established as being equivalent to the probability of the set generated by the binomial distribution X~ b $(a+b)^s$, where a = b = 0.5 and s = (50). The value of d_i was defined according to the average degree of dominance expressed in each trait. E_i is the environmental effect, generated according to a normal distribution with means equal to zero and variance given by the equation below:

$$\sigma_e^2 = \frac{\sigma_g^2(1 - h^2)}{h^2} \qquad (3)$$

where: σ_e^2 is the variance given by the environmental values, σ_g^2 is the variance of the genetic values, and h^2 is the heritability defined for the trait. The genetic variance is defined from the information of the genetic control and the importance of each locus in the polygenic model.

$$\sigma_{g=}^2 \ \frac{1}{2}\bar{a}^2 + \frac{1}{4}\bar{d}^2, \qquad (4)$$

where: $\bar{a}^2$, $\bar{d}^2$ were defined by the mean values of the effects associated with the homozygote and heterozygous genotypes for each of the 50 loci, respectively.

For the epistatic model, the phenotypic value of the ith individual was obtained according to the following equation:

$$Y_i = \mu + \sum_{j=1}^{50} p_j \alpha_j + \sum_{j=1}^{49} p_j \alpha_j \, \alpha_{j+1} + E_i \qquad (5)$$

In the above equation, the first summation of the expression refers to the contribution of the individual locus through its additive and dominant effects, and the second summation represents the multiplicative effects corresponding to the epistatic interactions between pairs of loci.α_j is the multiplicative effect of the favorable allele in locus *j*, and *j+1* and p_j is the contribution of locus *j* to the manifestation of the trait under consideration.

RR-BLUP

The RR-BLUP model used to obtain the genomic estimated breeding values (GEBV) (Meuwissen, Hayes, & Goddard, 2001):

$$y = Xb + Za + e, \qquad (6)$$

where: *y* is the vector of phenotypic observations, *b* is the vector of fixed effects, *a* is the vector of random marker effects, and *e* refers to the vector of random errors, N(0, σ_e^2); *X* and *Z* are matrices of incidence for *b* and *a,* respectively. Individual GEBVs were estimated by the following equation:

$$GEBVs = \hat{y}_i = \sum_j^n Z_{ij} \hat{a}_j \, , \qquad (7)$$

where: *n* is the number of markers arranged in the genome, Z_{ij} is the line of the incidence matrix that allocates the genotype of the jth marker for each individual (i), 1, 0, -1 for genotypes AA, Aa, aa, respectively, for biallelic and codominant markers, and $\hat{a}_j$ is the effect of the jth marker estimated by RR-BLUP. In this model, the incidence matrix associated with the effects of dominance was not included. However, it should be remembered that the population has a probable allele frequency p that is different from q; therefore, the additive effects estimated through matrix Z capture dominance effects.

Radial Basis Function Neural Network (RBFNN)

The RBFNN in the present study is a three-layered feed-forward neural network that consists of an input layer, which connects the ANN to its environment (groups the input data into clusters); a hidden layer that applies a nonlinear transformation of the space using Gaussian radial base activation functions (Figure 1); and the output layer, which applies a linear transformation in the hidden space providing an output to the network (Braga, Carvalho, & Ludermir, 2011). Radial functions represent a special class of functions whose value decreases or increases in relation to the distance from a central point (Braga et al., 2011).

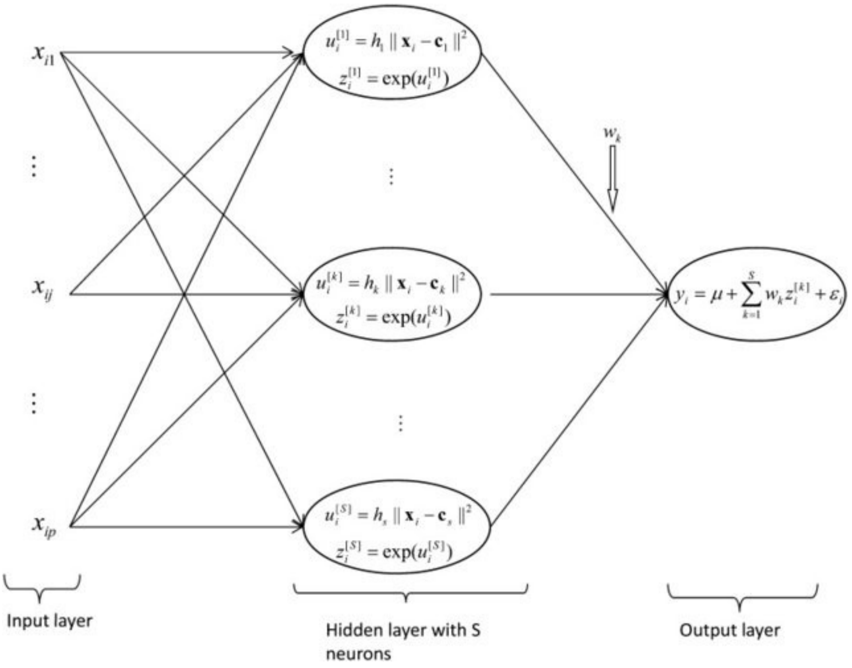
x_{i1}
x_{ij}
x_{ip}
$u_i^{[1]} = h_1 \|\mathbf{x}_i - \mathbf{c}_1\|^2$
$z_i^{[1]} = \exp(u_i^{[1]})$
$u_i^{[k]} = h_k \|\mathbf{x}_i - \mathbf{c}_k\|^2$
$z_i^{[k]} = \exp(u_i^{[k]})$
$u_i^{[S]} = h_s \|\mathbf{x}_i - \mathbf{c}_s\|^2$
$z_i^{[S]} = \exp(u_i^{[S]})$
w_k
$y_i = \mu + \sum_{k=1}^{S} w_k z_i^{[k]} + \varepsilon_i$
Input layer
Hidden layer with S neurons
Output layer

Figure 1. Structure of a radial basis function neural network (RBFNN).

The training of RBFNN optimization includes the weights between the hidden layer and the output layer, the activation function, the center of activation functions, the distribution of center of activation functions, and the number of hidden neurons (Cruz & Nascimento, 2018). During the training process, only the weights between the hidden layer and the output layer are modified. The vector of weights $\omega=\{w_{1,\dots,}w_S\}$ of the linear output layer is obtained using the ordinary least-squares fit that minimizes the mean squared differences between $\hat{y}_i$ (from RBFNN) and the $\hat{y}_i$ observed in the training set, provided that the Gaussian RBFs for centers c_k and h_k of the hidden layer are defined.

The radial basis function selected is usually a Gaussian kernel selected using the K-means clustering algorithm. The centers are selected using the orthogonalization least-squares learning algorithm, as described by Chen, Crowan, and Grant (1991), and are implemented in MATLAB (2010). The centers are added iteratively such that each new selected center is orthogonal to the others. The selected centers maximize the decrease in the mean squared error of the RBFNN, and the algorithm stops when the number of centers (neurons) added to the RBFNN attains a desired precision (goal error) or when the number of centers is equal to the number of input vectors, that is, when S = n.

To select the best RBFNN, a grid for training the net was generated, containing different spread values and different precision values (goal error). The spread value ranged from 5 to 100, and an initial value of 0.01 for the goal error was considered.

SNP subset selection

To determine the number of markers, stepwise regression was used in the scenario with epistatic effects, dominance and low heritability. In this procedure, the maximum number of markers was determined in conjunction with measures representative of the data as the mean square error root of the model (MSER), the determination coefficient (R^2) obtained by inclusion of the selected markers, and the condition number (CN) of the correlation matrix. For the first two criteria, the MSER chosen was the one that presented the lowest possible value tied to the best possible values for R^2 (the higher the better). The third criterion was used to avoid multicollinearity problems. The condition number of the correlation matrix between the explanatory variables verifies the degree of multicollinearity in the correlation matrix X'X (Montgomery, Peck, & Vining, 1982). When the CN resulting from this division was lower than or equal to 100, there was weak multicollinearity between the explanatory variables; for 100 < CN < 1000, there was moderate to severe multicollinearity, and for CN ⩾ 1000, severe multicollinearity was considered. Therefore, based on a graphical analysis, the number was determined by the graphical point with the best R^2, and the lowest REQM was observed when 100 < CN.

$$\mathrm{NC} = \frac{\lambda_n}{\lambda_1} \qquad (8)$$

where: λ_n is the eigenvalue of the largest absolute value and λ_1 is that of the smallest.

Computational applications for data analysis

The methodologies were compared using the reliability (R^2) defined as the squared correlation between the predicted GEBVs of the individuals with no phenotypic traits and the mean squared error root (MSER) using predicted and realized values. A fivefold cross-validation scheme was used to determine the reliability of the genomic prediction of a selected subset of SNPs in the population. The individuals (500) were randomly split into five equal-size groups, and each group with approximately 100 individuals (20% of the population) was, in turn, assigned phenotypic values and used as the validation set. The reliability reported in the study was the average of the reliability of genomic prediction from fivefold groups. For comparison purposes, the reliability of genomic prediction from all the SNPs (1000) was also calculated in addition to 100 SNPs selected to be even. The simulations were implemented with software GENES (Cruz, 2016) and the statistical analyses were performed with software R, with the RR-BLUP package (R core team, 2018), and the RBFNN and the stepwise regression were implemented using Genes software in integration with MATLAB (MATLAB, 2010). Before the dimensional reduction, we used a server DELL 12° generation, Intel Xeon E5-26 processor 3,30 GHz, RAM with 64 GB and Hard drive with 1024 GB, and after the dimensionality reduction, we used PC Processor Intel i7-3770 CPU 3.40 GHz memory RAM de 8GB e operational system of 64Bits.

Results

Dimensionality reduction was performed using a graphical procedure that considers the determination coefficient (R^2) obtained by including the selected markers, the error associated and the condition number (CN) of the correlation matrix. The number of markers was determined by the graphical point that presented the larger (R^2 and the lowest MSER when 100 < CN (Figure 2). After defining the optimal number of markers, stepwise regression was used to select from all markers the ones used in the fit.

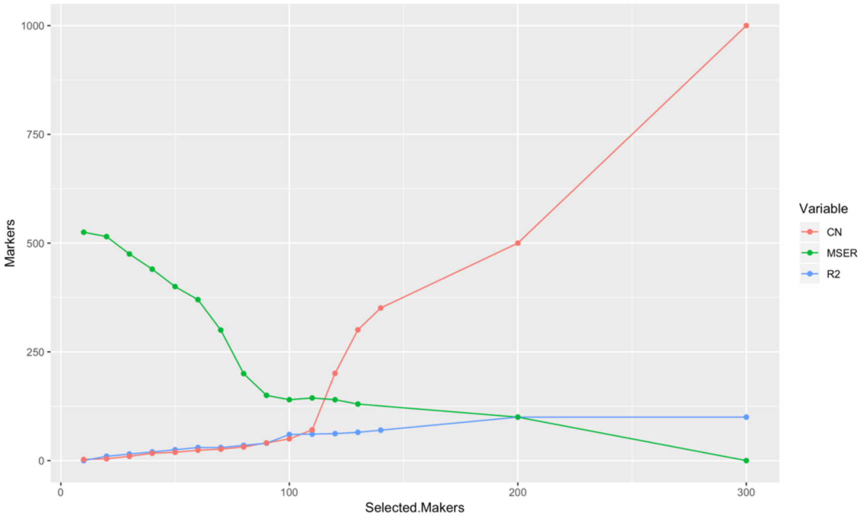

Figure 2. Graphical representation of the values of determination coefficient (R^2) in blue, mean squared error root (MSER) in green and the condition number (CN) in red obtained by the stepwise regression method by including 1 to 300 molecular markers (from the total of 1,000) in the stepwise regression model.

Twelve different scenarios considering different levels of heritability, dominance and epistatic effects were evaluated (Tables 2 and 3). Fivefold cross-validation was used to access the reliability (R^2) of fit models (RBFNN and RR-BLUP), considering or not considering dimensionality reduction. The dimensionality reduction allowed the time for the development of the analysis to decrease from an average of 20h using a normal computer and approximately 12h in a super computer for less than 1h after the reduction of dimensionality. The dimensionality reduction procedure was also delayed for approximately 20h to determine the ideal number of markers.

Overall, dimensionality reduction improved the reliability values for all scenarios, specifically with h^2 = 30 the reliability value from 0.03 to 0.59 using RBFNN and from 0.10 to 0.57 with RR-BLUP in the scenario with additive effects. In the additive dominant scenario, the reliability values changed from 0.12 to 0.59 using RBFNN and from 0.12 to 0.58 with RR-BLUP, and in the epistasis scenarios, the reliability values changed from 0.07 to 0.50 using RBFNN and from 0.06 to 0.47 with RR-BLUP (Table 2).

In the scenarios with h^2 = 60, the reliability value improved from 0.38 to 0.79 using RBFNN and from 0.36 to 0.79 with RR-BLUP in the scenario with additive effects. In the scenario with additive dominance, the values changed from 0.34 to 0.79 using RBFNN and from 0.30 to 0.73 with RR-BLUP, and in the epistatic scenarios, the average of reliability values changed from 0.10 to 0.60 using RBFNN and from 0.08 to 0.58 with RR-BLUP (Table 2).

Table 2. Reliability values of selection obtained from RBFNN and RR-BLUP through all markers (1,000) or selected markers (100) by stepwise regression in a set of validation data involving cross-validation procedures.

	Reliability values			
Scenarios	1000 RBFNN	1000 RRBLUP	100 RBFNN	100 RR-BLUP
V_1-D0H30_Ad	0.11 ± 0.12	0.10 ± 0.02	0.59 ± 0.02	0.57 ± 0.03
V_2-D0.5H30_Ado	0.12 ± 0.06	0.12 ± 0.07	0.59 ± 0.03	0.58 ± 0.05
V_3-D1H30_Ado	0.02 ± 0.01	0.01 ± 0.01	0.56 ± 0.07	0.54 ± 0.06
V_{4-}D0H30_Ep	0.03 ± 0.00	0.01 ± 0.01	0.45 ± 0.05	0.42 ± 0.05
V_{5-}D0.5H30_Ep	0.05 ± 0.02	0.02 ± 0.02	0.56 ± 0.05	0.54 ± 0.06
V_6-D1H30_Ep	0.07 ± 0.05	0.06 ± 0.05	0.50 ± 0.05	0.47 ± 0.04
V_7-D0H60_Ad	0.38 ± 0.08	0.36 ± 0.07	0.79 ± 0.03	0.79 ± 0.03
V_8D0.5H60_Ado	0.34 ± 0.07	0.30 ± 0.07	0.74 ± 0.03	0.73 ± 0.03
V_9-D1H60_Ado	0.18 ± 0.04	0.19 ± 0.05	0.64 ± 0.02	0.64 ± 0.01
V_{10-}D0H60Ep	0.06 ± 0.03	0.03 ± 0.05	0.58 ± 0.05	0.79 ± 0.05
V_{11-}D0.5H60_Ep	0.10 ± 0.02	0.08 ± 0.07	0.62 ± 0.04	0.59 ± 0.09
V_{12}-D1H60_Ep	0.13 ± 0.03	0.13 ± 0.07	0.58 ± 0.08	0.58 ± 0.09

Table 3 shows the range of values for the accuracy of prediction (MSER = mean squared err root). For all scenarios with and without dimensionality reduction, RBFNN outperformed RRBLUP. In addition, dimensionality reduction also improved the accuracy of RBFNN and RRBLUP. The MSER value ranged from 3.5 to 23.8 for RBFNN and from 71.4 to 575.4 for RR-BLUP. Specifically, with h^2 = 30, the MSER value ranged from 5.9 to 4.9 using RBFNN and from 33.7 to 23.4 with RR-BLUP in the scenario with additive effects. In the additive-dominance scenario, the average of MSER values changed from 11.5 to 9.3 using RBFNN and from 47.1 to 29.4 with RR-BLUP, and in the epistasis scenarios, the average of MSER values changed from 19.73 to 13.76 using RBFNN and from 380.7 to 2,773.9 with RR-BLUP.

Table 3. MSER obtained from RBFNN and RR-BLUP through all markers (1,000) or selected markers (100) by stepwise regression in a set of validation data involving cross-validation procedures.

	MSER –Mean squared error root			
Scenarios	1,000 RBF	1,000 RR-BLUP	100 RBF	100 RR-BLUP
V_1-D0H30_Ad	5.9 ± 0.1	33.7 ± 2	4.9 ± 0.1	23.4 ± 0.0
V_{2-}D0.5H30_Ado	6.2 ± 0.1	36.0 ± 1.3	5.1 ± 0.1	24.9 ± 3.7
V_3-D1H30_Ado	16.8 ± 1.2	58.2 ± 5.7	13.5 ± 0.1	34.9 ± 17.7
V_4-D0H30_Ep	16.9 ± 1.2	267.9 ± 10.5	14.2 ± 0.1	205.9 ± 48.1
V_5-D0.5H30_Ep	18.5 ±0.6	338.1 ± 17.7	15.5 ± 0.3	232.1 ± 18.9
V_6-D1H30_Ep	23.8 ± 1.7	534.6 ± 24	20.8 ± 0.3	395.9 ± 40.6
V_7-D0H60_Ad	4.5 ± 0.1	20.7 ± 1	3.4 ± 0.4	11.99 ± 0.9
V_8-D0.5H60_Ado	4.7 ± 0.2	22.4 ± 2	3.7 ± 0.1	107.4 ± 1.1
V_9-D1H60_Ado	5.4 ± 0.2	28.23 ± 2	4.3 ± 0.1	145.9 ± 5.2
V_{10}-D0H60Ep	13.5 ± 0.3	181.5 ± 12	11.7 ± 0.2	320.1 ± 36.4
V_{11}-D0.5H60_Ep	15.5 ± 0.5	236.8 ± 19	12.4 ± 0.3	280.9 ± 28.9
V_{12}-D1H60_Ep	18.8 ± 0.8	355.5 ± 26	15.7 ± 0.6	473.8 ± 22.0

In the scenarios with $h^2 = 60$, the MSER value improved from 4.5 to 3.4 using RBFNN and from 85.8 to 71.4 with RR-BLUP in the scenario with additive effects. In the scenario with additive dominance, the average of values changed from 5.0 to 4.0 using RBFNN and from 23.76 to 88.43 with RR-BLUP, and in the epistasis scenarios, the average of reliability values changed from 15.9 to 13.2 using RBFNN and from 257.9 to 358.3 with RR-BLUP.

Discussion

The dimensionality reduction for the model fit is a recurring theme in several studies aimed at genomic prediction of genetic values (Long et al., 2011ab; Azevedo, de Resende, Fonseca, Lopes, & Guimarães, 2013; Azevedo et al., 2014; Akidemir et al., 2017). However, it is worth noting that there is a difference between two approaches that are usually considered dimensionality reduction approaches. The first approach uses such methods as main and independent components to obtain latent variables that will be used to fit the models. With that strategy, the main goal is not to exclude markers but to use the latent variables, which are linear combinations of all available markers, to fit the model. In the second approach, the researcher has an interest in selecting the markers most related to the traits of interest and uses them in fitting the models for their benefits both in regression models and in diversified architectures of computational intelligence (Long, Gianola, Rosa, Weigel, & Avendano, 2007; Long et al., 2010; Akidemir et al., 2017). The present study considers the second approach.

In general, in terms of reliability, dimensionality reduction positively impacted all the scenarios evaluated, which represented different genetic architectures (Table 1). Better performance was not observed regarding the use of neural networks compared with the results obtained with RR-BLUP. These results suggest that the degree of simulated epistasis, in which only dual interactions between subsequent markers are considered, was not a determining factor in differentiating the fit of regression models and neural networks. In terms of dominance, as already reported in the literature (Long et al., 2007; Denis & Bouvet, 2011; Almeida Filho et al., 2016; Santos et al., 2016; Xu et al., 2018), that factor is not regarded as a problem in genomic prediction studies. Therefore, even if nonparametric models (such as artificial neural networks) do not need to impose strong assumptions upon the phenotype-genotype relationship, presenting the potential to capture interactions between loci by the interactions between neurons of different layers (Gianola, Fernando, & Stella, 2006; Long et al.; 2010), a substantial improvement in the prediction process depends on the level of epistasis present. In terms of reliability, similar results were observed in the studies carried out by (Long et al., 2010; Long, 2011a), which were based on complete genome simulation with 2000 markers in a random mating population of bulls and heifers in three scenarios: additive, dominant and epistatic. In these researchers' study, two RBFNN models were used; in the first, there were specific weights for each SNP, while in the second, all SNPs had the same importance. In most cases, the model with specific weights was better than that with a common weight for each SNP.

Weigel et al. (2010a) and Weigel, Van Tassell, O'Connell, VanRaden, and Wiggans (2010b) compared the use of some equally spaced markers in the genome and imputed other markers based on a reference population with all of the genotyped markers using a set of markers selected according to their effect on the character of interest. The above authors concluded that when the number of selected markers is small, the predictive capacity of the model with markers selected according to the effect is higher than the use of a smaller set of markers scattered throughout the genome.

Conversely, considering the results within the two approaches evaluated (RBFNN and RR-BLUP), dimensionality reduction also caused a reduction in the MSER values. These results were similar to those obtained by the authors in González-Camacho et al. (2012), who observed that it is possible to improve prediction, both in terms of R^2 and MSER, predicting genetic values by means of nonparametric models when the selection includes markers that are not related to the traits of interest. When the methods were compared, a gain was observed in terms of MSER when the fitting was performed by means of neural networks.

In the case of RR-BLUP, the effects of dominance and epistasis contributed to the increase of the error by increasing the difference between the expected and the observed values. In this way, when the interest is to select only a small number of individuals, the best 20% for example may not be the same. Similar results were observed in the study developed by the authors in Long et al. (2011a), who used simulation of quantitative characters under different modes of gene action (additive, dominant, and epistatic) and found that RBFNN had a better ability to predict the merit of individuals in future generations in the presence of non-additive effects than by using an additive linear model, such as the Bayesian Lasso. In the case of purely additive gene effects, RBFNN was slightly worse than Lasso. In the above study, the authors reported the use of the dimensionality reduction method – of the main component type – before using RBFNN and showed that with the selection of markers, the performance of the radial base network was better.

In nonparametric models, no assumption is made regarding the form of the genotype–phenotype relationship. Instead, this relationship is described by a smoothing function and driven primarily by the data. Therefore, RBFNN should be flexible with respect to the type of input data and mode of gene action, such as epistasis (Gianola et al., 2011; Perez-Rodriguez et al., 2012; Felipe, Okut, Gianola, Silva, & Rosa, 2014; Howard et al., 2014). This finding is observed because artificial neural networks (ANNs) can capture nonlinear relationships between predictors and responses and learn about functional forms in an adaptive manner because they act as universal approximators of complex functions (Gianola et al., 2011). ANNs are interesting candidates for the analysis of characters affected by genetic action with epistatic effects.

Due to the importance of epistasis in studies of quantitative traits in plants (Holland, 2006; Dudley, 2008; Zheng, Li, & Wang, 2011; Dudley & Johnson, 2009; Denis & Bouvet, 2011; Viana & Piepho, 2017), explicit (in the model) or implicit (in hidden layers) inclusion of epistatic interactions may increase the accuracy of prediction (Lee, van der Werf, Hayes, Goddard, & Visscher, 2008). Furthermore, the frequency variation of the epistatic allele between populations may cause the gene-of-interest effect to be significant in one population but not in another, and the effect may even be inverse on the character in different environments (Long et al., 2011a), which reinforces the importance of using computational intelligence methods that easily incorporate interactions between linear effects through their hidden layers.

Conclusion

The use of a variable selection procedure is an effective strategy to improve the prediction accuracy of computational intelligence techniques that successfully allow incorporating interaction effects, which in the present study represent biological epistatic interactions.

Acknowledgements

CAPES – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior –, FAPEMIG – Fundação de Amparo à Pesquisa do Estado de Minas Gerais – and CNPq – Conselho Nacional de Desenvolvimento Científico e Tecnológico, for the financial support and the grant conceded.

References

Akidemir, D., Jannink, J. L., & Isidro-Sánchez, J. (2017). Locally epistatic models for genome-wide prediction and association by importance sampling. *Genetics Selection Evolution*, *49*(1), 49-74. DOI: 10.1186/S1271101703488

Almeida-Filho, J. E., Guimarães, J. F. R., Silva, F. F., de Resende, M. D. V., Muñoz, P., Kirst, M., & Resende Jr., M. F. R. (2016). The contribution of dominance to phenotype prediction in a pine breeding and simulated population. *Heredity*, *117*(1), 33-41. DOI: 10.1111/1468-0009.12357

Azevedo, C. F., de Resende, M. D. V., Fonseca, F., Lopes, P. S., & Guimarães, S. E. F. (2013). Regressão via componentes independentes aplicada à seleção genômica para características de carcaça em suínos. *Pesquisa Agropecuária Brasileira*, *48*(6), 619-626. DOI: 10.1590/S0100-204X2013000600007

Azevedo, C. F., Silva, F. F., de Resende, M. D. V., Lopes, M. S., Duijvesteijn, N., Guimarães, S. E. F., ... 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), 452-461. DOI: 10.1111jbg12104

Braga, A.P., Carvalho, A. P. L. F., & Ludermir, T. B. (2011). *Redes neurais artificiais - teoria e aplicações* (2a. ed.). Rio de Janeiro, RJ: LTV.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., ... Dreisigacker, S.(2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961-975. DOI: 10.1016/j.tplants.2017.08.011

Chen, S., Cowan, C. F., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, *2*(2), 302-309. DOI: 10.11097280341

Cruz, C. D., & Nascimento, M. (2018). *Inteligência computacional aplicada ao melhoramento genético*. Vicosa, MG: Editora UFV.

Cruz, C. D. (2016) Genes Software-extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum. Agronomy*, *38*(4), 547-552. DOI: 10.4025/actasciagron.v38i4.32629

Denis, M., & Bouvet, J. M. (2011). Genomic selection in tree breeding: testing accuracy of prediction models including dominance effect. *BMC Proceedings*, *5*(7), 1-2. DOI: 10.1186/175365615S7O13

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. DOI: 10.2135/cropsci2007.04.0242

Dudley, J. W., & Johnson, G. R. (2009). Epistatic models improve prediction of performance in corn. *Crop Science*, *49*(3), 763-770. DOI: 10.2135/cropsci2008.08.0491

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*(1), 1-10. DOI: 10.1186/s12863-014-0149-9

Gianola, D., Fernando, R. L., & Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, *173*(3), 1761-1776. DOI: 101534genetics105049510

Gianola, D., Okut, H., Weigel, K. A., & Rosa, G. J. (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genetics*, *12*(1), 1-14. DOI: 10.1186/1471-2156-12-87

González-Camacho, J. M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J. E., Mahuku, G., ... Crossa, J. (2012). Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, *125*(4):759-771. DOI: 10.1007s0012201218689

Holland, J.B. (2006). Theoretical and biological foundations of plant breeding. In K. R. Lamkey, & M. Lee (Ed.), *Plant breeding: the Arnel R Hallauer International Symposium* (p. 127-140). Ames, IA: Blackwell Publishing. DOI: 10.1002/9780470752708.ch9

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*, *4*(6), 1027-1046. DOI: 101534g3114010298

Lee, S. H., van der Werf, J. H., Hayes, B. J., Goddard, M. E., & Visscher, P. M. (2008). Predicting unobserved phenotypes for complex traits from whole-genome SNP data. *PLoS Genetics*, *4*(10), 1-11. DOI: 10.1371journalpgen1000231

Long, N., Gianola, D., Rosa, G. J., & Weigel, K. A. (2011a). Marker-assisted prediction of non-additive genetic values. *Genetica*, *139*(7), 843-854. DOI: 10.1007s1070901195887

Long, N., Gianola, D., Rosa, G. J. M., & 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), 247-257. DOI: 10.1111j14390388201100917x

Long, N., Gianola, D., Rosa, G. J., Weigel, K. A., Kranis, A., & Gonzalez-Recio, O. (2010). Radial basis function regression methods for predicting quantitative traits using SNP markers. *Genetics Research*, *92*(3), 209-225. DOI: 10.1017S0016672310000157

Long, N., Gianola, D., Rosa, G. J., Weigel, K. A., & 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), 377-389. DOI: 101159000317279

Mackay, T. F., Stone, E. A., & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, *10*(8), 565. DOI: 101111j14390388200700694x

MATLAB. (2010). *Matlab Version 7.10.0*. Natick, MA: 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., & Vining, G. G. (1982). *Introduction to linear regression analysis*. New York, US: John Wiley and Sons.

Pérez-Rodríguez, P., Gianola, D., González-Camacho, J. M., Crossa, J., Manès, Y., & Dreisigacker, S. (2012). Comparison between linear and non-parametric regression models for genome-enabled prediction in wheat. *G3: Genes, Genomes, Genetics*, *2*(12), 1595-1605. DOI: 101534g3112003665

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, AU: R Foundation for Statistical Computing. Retrieved on Dec. 20, 2018 from https://www.R-project.org.

Santos, V. S., Martins Filho, S., Resende, M. D. V., Azevedo, C. F., Lopes, P. S., Guimarães, S. E. F., & Silva, F. F. (2016). Genomic prediction for additive and dominance effects of censored traits in pigs. *Genetics and Molecular Research*, *15*(4), 1-16. DOI: 10.4238/gmr15048764

Sant'Anna, I. C., Nascimento, M., Silva, G. N., Cruz, C. D., Azevedo, C. F., Silva, F. F., & Gloria, L. S. (2019). Genome-enabled prediction of genetic values for using radial basis function neural networks. *Functional Plant Breeding Journal*, *1*, 29-40. DOI:10.35418/2526-4117/v1n2a1

Viana, J. M. S., & Piepho, H. P. (2017). Quantitative genetics theory for genomic selection and efficiency of genotypic value prediction in open-pollinated populations. *Scientia Agricola*, *74*(1), 41-50. DOI: 10.1590/0103-9016-2014-0383

Weigel, K. A., de Los Campos, G., Vazquez, A. I., Rosa, G. J. M., Gianola, D., & Van Tassell, C. P. (2010a). Accuracy of direct genomic values derived from imputed single nucleotide polymorphism genotypes in Jersey cattle. *Journal of Dairy Science*, *93*(11), 5423-5435. DOI: 103168jds20103149

Weigel, K. A., Van Tassell, C. P., O'Connell, J. R., VanRaden, P. M., & Wiggans, G. R. (2010b). Prediction of unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels and population-based imputation algorithms. *Journal of Dairy Science*, *93*(5), 2229-2238. DOI: 10.3168jds20092849

Xu, Y., Wang, X., Ding, X., Zheng, X., Yang, Z., Xu, C., & Hu, Z. (2018). Genomic selection of agronomic traits in hybrid rice using an NCII population. *Rice*, *11*(1), 1-10. DOI: 10.1186s1228401802234

Zheng, S. J., Li, Z. Q., & Wang, H. T. (2011). A genetic fuzzy radial basis function neural network for structural health monitoring of composite laminated beams. *Expert Systems with Applications*, *38*(9), 11837-11842. DOI: 101016jeswa201103072