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

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], Cosme Damião Cruz[3]

[1] Department of Statistics, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

[2] Department of Statistics, Federal University of Rondônia,Ji-Paraná, Rondônia, Brazil

[3] Department of General Biology, Federal University of Viçosa, Viçosa, Minas Gerais, Brazil

*Corresponding author

E-mail: isabelacsantanna@gmail.com

Abstract

This paper aimed to evaluate the efficiency of subset selection of markers for genome-enabled prediction of genetic values using radial basis function neural networks (RBFNN). For this purpose, an F1 population from hybridization of divergent parents with 500 individuals genotyped with 1,000 SNP-type markers was simulated. Phenotypic traits were determined by adopting three different gene action models – additive, additive-dominant, and epistasic , complying with two dominance situations: partial and complete with quantitative traits admitting heritability (h^2) equal to 30 and 60%, each one controlled by 50 loci, considering two alleles per locus, 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 methods were used: dimensionality reduction and stepwise regression. The square of the correlation between the predicted genomic estimated

breeding value (GEBV) and the phenotype value was used to measure predictive reliability. For $h^2 = 0.3$ in the additive scenario, the R^2 values were 59% for neural network (RBFNN) and 57% for RR-BLUP, and in the epistatic scenario, R^2 values were 50% and 41%, respectively. Additionally, when analyzing the mean-squared error root, the difference in performance between the techniques is even greater. For the additive scenario, the estimates were 91 for RR-BLUP and 5 for neural networks and, in the most critical scenario, they were 427 for RR-BLUP and 20 for neural network. The results showed that the use of neural networks and variable selection techniques allows capturing epistasis interactions, leading to an improvement in the accuracy of prediction of the genetic value and, mainly, to a large reduction of the mean square error, which indicates greater genomic value.

Keywords: simulation; genomics; linkage disequilibrium; breeding.

Introduction

One of the major challenges of genetic breeding today is understanding the genetic variation of quantitative traits, QTL (Quantitative trait loci), which are conditioned by a large number of genes with small effects [1] whose interaction often results in non-linearity in relations between phenotypes and genotypes [2,3].

With the advent of Genomic Selection (GS) [4], it became possible to estimate the genomic value of individuals (GEBV) without the need of 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 interaction between their additive, dominant and epistatic effects. However, most applications of GS include only the additive portion of the genetic value, so 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 [5].

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 [2]. Besides that, before fitting the model it is necessary to define the model effects to be estimated. In this context, Artificial Neural Networks (ANNs) have a great potential because they can capture non-linear relationships between markers from the data themselves (without a previous model definition), which most of the models commonly used in the GS cannot do [2,6,7].

Radial Basis Function Neural Networks (RBFNN) are a particular class of Neural Networks (NN) that have properties that make it attractive to GS applications. According to Gianola et al. [8], RBFNNs have the ability to learn from the data used in their training, have universal approximation properties [9], give a unique solution and are faster than standard ANNs [10].

However, the inclusion of all markers in the prediction RBFNN model increases the chances of a high correlation between the markers [11]. The number of markers represents a huge challenge that leads to less precision and a great computational demand for NN training. This happens because NNs use good part of their resources to represent irrelevant portions of the search space and compromising the learning process because there are thousands of markers available in the genome [12].Thus, a more realistic model should include only SNPs related to the traits of interest [6].

For this reason, a subset of SNPs can be used for training, since, by reducing the search space, RBFNNs improve the learning process and increase the predictive power of the model, as realized by [2]. These authors used two types of RBFNN models: one considering a common weight parameter for each SNP, and the other in which each SNP has specific parameters of

77 importance. However, due to the importance of NNs for the improvement of prediction of
78 quantitative traits, there is still a need to test different dimensionality reduction methods and
79 prediction models for polygenic traits.

80 In view of the above, this paper aimed to evaluate the efficiency of genome-enabled
81 prediction by radial basis function neural networks (RBFNN) in the prediction of genetic values
82 by considering a subset of markers in simulated data set with different gene actions (as
83 dominance and epistasis) and degrees of heritability. The results were compared with those
84 obtained by one of the standard GS model: RR-BLUP.

85

86 Material and methods

87 Origin of populations

88 In order to assess the reliability of GS predictions, data were simulated by considering
89 a diploid species with 2n = 2x = 20 chromosomes as the reference, and the total length of the
90 genome was stipulated in 1.000 cM. Genomes were generated with a saturation level of 101
91 molecular markers spaced by 1cM per linkage group, totaling 1010 markers. Divergent parental
92 line genomes were simulated, as well as genomes from the base population (F_1). Since the base
93 population was derived from two contrasting homozygous parents, the effective size of the base
94 population is the size of F_1 itself.

95 Simulation of quantitative traits

96 Quantitative traits were simulated in three scenarios by considering three degrees of
97 dominance (d/a = 0, 0.5 and 1) and two broad sense heritability (h^2 = 0.30 and 0.60), considering

two gene actions: additive and epistatic, thus totaling six scenarios. Each trait was controlled by 50 randomly chosen loci, with 2 alleles per locus.

The phenotypic values of the i^{th} individuals were obtained according to the additive model as follows:

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

where α_j is the effect of the favorable allele in locus *j*, considered equal to 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 bellow:

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

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.

For the epistatic model, the phenotypic values of the i^{th} individuals were obtained according to the following equation:

$$Y_i = \mu + \sum_{j=1}^{50} p_j \alpha_j + \sum_{j=1}^{49} p_j \alpha_j \alpha_{j+1} + E_i \quad (3)$$

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

122 loci.α_j is the multiplicative effect of the favorable allele in locus j, and $j+1$ and p_j is the
123 contribution of locus j to the manifestation of the trait under consideration.

124 **Table 1.** Scenarios composed by combination of traits, action genenic model.
125 heritabilty and dominance degree.

Traits	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

126
127
128

129 # RR-BLUP

130 The RR-BLUP model was used to obtain the genomic estimated breeding values
131 (GEBV) [4]:

132 $$y = Xb + Za + e, \quad (4)$$

133 where y is the vector of phenotypic observations, b is the vector of fixed effects, a is the vector
134 of random marker effects, and e refers to the vector of random errors, N(0, σ_e^2); X and Z are
135 matrices of incidence for a and b. The effects of the individuals (GEBVs) were estimated by
136 the equation below:

137 $$GEBVs = \hat{y}_i = \sum_j^n Z_{ij}\hat{\alpha}_j\,, \quad (5)$$

where *n* is the number of markers arranged in the genome, Z_{ij} is the line of the matrix of incidence that allocates the genotype of the j[th] marker for each individual (i), 1, 0, -1 for genotypes A_1A_1, A_1A_2, A_2A_2, respectively, for biallelic and codominant markers, and $\hat{\alpha}_j$ is the effect of the j[th] 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 probably allele frequency being *p* different from *q* and therefore the additive effects estimated through matrix *Z* capture dominance effects.

Radial Basis Function Neural Network (RBFNN)

A RBFNN is an artificial neural network that uses radial basis functions as activation functions. The RBFNN in the present study is a three layered feed-forward neural network, where the first layer is linear and only distributes the input signal, while the next layer is nonlinear and uses Gaussian functions (Fig 1).

Fig. 1 Structure of a radial basis function neural network (RBFNN). In the hidden layer, each input vector ($\mathbf{x_{i1}},\ldots,\mathbf{x_{ip}}$)is summarized by the Euclidean distance between the input vectors x_i and the centers c_m,(m = 1, ..., M) neurons, i.e., $h_m||x_i - c_m||$, where h_m is a bandwidth parameter. Then distances are transformed by the Gaussian kernel $\mathbf{exp}(-(h_m||x_i - c_m||)^2)$ for obtaining the response, $y_i = \mathbf{w}0 + \sum_{m=1}^{M} + \mathbf{w}m\mathbf{z}mi + \varepsilon_i$ (extracted from Gonzalez-Camacho et al., 2012).

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 [13]. During the training process, only the weights between the hidden layer and the output layer are modified. The vector of weights $\omega=\{w_{1,\ldots,}w_s\}$ of the linear output layer is obtained using the ordinary least-squares fit that minimizes the mean squared differences between y_i (from RBFNN) and

163 the observed y_i observed in the training set, provided that the Gaussian RBFs for centers c_k
164 and h_k of the hidden layer are defined.

165 The radial basis function selected is usually a Gaussian kernel selected using K-means
166 clustering algorithm. The centers are selected using the orthogonalization least-squares
167 learning algorithm, as described by [14] and implemented in [15]. The centers are added
168 iteratively such that each new selected center is orthogonal to the others. The selected centers
169 maximize the decrease in the mean squared error of the RBFNN, and the algorithm stops
170 when the number of centers (neurons) added to the RBFNN attains a desired precision (goal
171 error) or when the number of centers is equal to the number of input vectors, that is, when
172 S=n.

173 To select the best RBFNN, a grid for training the net was generated, containing
174 different spread values and different precision values (goal error). The spread value ranging
175 from 5 to 100 and an initial value of 0.01 for the goal error was considered. The spread
176 parameter allows adjusting the form of the Gaussian RBFNN such that it is sufficiently large
177 to respond to overlapping regions of the input space but not so big that it could induce the
178 Gaussian RBFNN to have a similar response [16].

179 SNP subset selection

180 To determine the number of markers, stepwise regression was used in the scenario with epistatic
181 effects, dominance and low heritability. In this procedure, the maximum number of markers
182 was determined in conjunction with measures representative of the data as the mean square
183 error root of the model (MSER), determination coefficient (R^2) obtained by inclusion of the
184 selected markers, and the condition number (CN) of the correlation matrix. As for the first two
185 criteria, the MSER chosen was the one that presented the lowest possible value tied to the best
186 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 [17]. When the CN resulting from this division was lower than or equal to 100, it was considered that there was weak multicollinearity between the explanatory variables; for 100 < CN <1000 moderate to severe multicollinearity, and for CN≥1000 severe multicollinearity was considered. So, based on a graphical analysis, the number was determined by the graphical point with the best R^2 , the lowest REQM when 100 <CN.

$$NC=\frac{\lambda_n}{\lambda_1} \quad (6)$$

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

Computational applications for data analysis

The models 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 root mean squared error (RMSE) using predicted and realized values. A five-fold cross-validation scheme was used to determine the reliability of 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 about 100 individuals (20% of the population) was in turn assigned with phenotypic values and used as the validation set. The reliability of genomic prediction was calculated as the squared correlation between the predicted GEBVs of the individuals with no phenotypic traits. The reliability reported in the study was the average of the reliability of genomic prediction from 5-fold 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 [18] and the statistical analyses were performed with software R, with the RR-BLUP package [19].

Results

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

Figure 2: A graphical representation of the values of determination coefficient (R^2) in black, Mean squared error root (MSER) in red and the condition number(CN) in blue obtained by the stepwise regression method by including 1 to 400 molecular markers (from the total of 1000) in the stepwise regression model.

Twelve different scenarios considering different levels of heritability, dominance and epistatic effects were evaluated (Table 2 and 3). Five cross-validation folds were used to access the reliability (R^2) of fit models (RBFNN and RR-BLUP), considering or not dimensionality reduction.

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 (Radial Basic Neural Network) and RR-BLUP through all markers (1000) or selected markers (100) by Stepwise Regression(SWR) 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. Besides that, 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 2773.9 with RR-BLUP .

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 257.9.0 to 358.3 with RR-BLUP.

Table 3. Mean squared error root obtained from RBFNN and RR-BLUP through all markers (1000) or selected markers (100) by Stepwise Regression(SR)in a set of validation data involving cross-validation procedures.

MSER –**mean squared error root**

Scenarios	1000 RBF	1000 RR-BLUP	100 RBF	100-RRBLUP
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

Discussion

The dimensionality reduction for the model fit is a recurring theme in several studies aimed at genomic prediction of genetic values [6,12,20,21]. However, it is worth noting that there is a difference between two approaches usually considered as dimensionality reduction ones. The first approach uses methods such 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 whether they are regression ones or diversified architectures of computational intelligence [22,2,5] for their benefits both in regression models and in diversified architectures of computational intelligence. 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 when compared with the results obtained with RR-BLUP. These results suggest that the degree of simulated epistasis, in

283 which only dual interactions between subsequent markers are considered, was not a determining
284 factor in differentiating the fit of regression models and neural networks. In terms of dominance,
285 as already reported in the literature [22,23,24,25,26], that is not regarded as a problem in
286 genomic prediction studies. Therefore, even if non-parametric models such as artificial neural
287 networks do not need to impose strong assumptions upon the phenotype-genotype relationship
288 presenting the potential to capture interactions between loci by the interactions between neurons
289 of different layers [27,2], a substantial improvement in the prediction process depends on the
290 level of epistasis present. In terms of reliability, similar results were observed in the studies
291 carried out by [2,6], which were based on complete genome simulation with 2000 markers in a
292 random mating population of bulls and heifers in three scenarios: additive, dominant and
293 epistatic. In the present study, two RBFNN models were used, and in the first one there were
294 specific weights for each SNP; while in the second one, all SNPs had the same importance. In
295 most cases, the model with specific weights was better than that with a common weight for
296 each SNP.

297 Weigel et al. [28,29] compared the use of some equally spaced markers in the genome
298 and imputed other markers based on a reference population with all the genotyped markers
299 using a set of markers selected according to their effect on the character of interest. The above
300 authors concluded that when the number of selected markers is small, the predictive capacity
301 of the model with markers selected according to the effect is higher than the use of a smaller
302 set of markers scattered throughout the genome.

303 On the other hand, considering the results within the two approaches evaluated (RBFNN
304 and RR-BLUP), dimensionality reduction also caused a reduction in the RMSE values. These
305 results were similar to those obtained by the authors in [10], who observed that it is possible to
306 improve prediction, both in terms of R^2 and RMSE, predicting genetic values by means of non-
307 parametric models when the selection includes markers that are not related to the traits of
308 interest. When the methods were compared, a gain was observed in terms of RMSE when the
309 fitting was performed by means of Neural Networks.

310 In the case of RR-BLUP, the effects of dominance and epistasis contributed to the
311 increase of the error by increasing the difference between the expected and the observed values.
312 In this way, when the interest is to select only a few individuals, the best 20% for example may
313 not be the same. Similar results were observed in the study developed by the authors in [6],
314 who used simulation of quantitative characters under different modes of gene action (additive,
315 dominant, and epistatic) and found that RBFNN had a better ability to predict the merit of
316 individuals in future generations in the presence of non-additive effects than by using an

317 additive linear model, such as the Bayesian Lasso one. In the case of purely additive gene effects,
318 RBFNN was slightly worse than Lasso. Still in the above study, the authors reported the use of
319 the dimensionality reduction method – of the main component type – before using RBFNN and
320 also showed that with the selection of markers the performance of the radial base network was
321 better.

322 In non-parametric models, no assumption is made regarding the form of the genotype–
323 phenotype relationship. Rather, this relationship is described by a smoothing function and
324 driven primarily by the data . Because of that, RBFNN should be flexible with respect to type
325 of input data and mode of gene action, such as epistasis [8,30,31,7]. This is due to the fact that
326 artificial neural networks (ANNs) can capture non-linear relationships between predictors and
327 responses and learn about functional forms in an adaptive manner, since they act as universal
328 approximators of complex functions [8]. ANNs are interesting candidates for the analysis of
329 characters affected by genetic action with epistatic effects.

330 Due to the importance of epistasis in studies of quantitative traits in plants
331 [32,33,34,35,36,37], explicit (in the model) or implicit (in hidden layers) inclusion of epistatic
332 interactions may increase the accuracy of prediction [38]. Furthermore, the frequency variation
333 of the epistatic allele between populations may cause the gene-of-interest effect to be significant
334 in one population but not in another, and the effect may even be inverse on the character in
335 different environments [5], which reinforces the importance of using computational intelligence
336 methods that easily incorporate interactions between linear effects through their hidden layers.

337 # Conclusion

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

342

343 # References

344

345 1.Risch NJ. Searching for genetic determinants in the new millennium. Nature. 2000;405

346 (6788):847-856.

2. Long N, Gianola D, Rosa GJ, Weigel KA, Kranis A, Gonzalez-Recio O. Radial basis function regression methods for predicting quantitative traits using SNP markers. Genet Res. 2010; 92(3):209-225.

3. Mackay, TF, Stone, EA and Ayroles, JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10(8), 565-577.

4. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157(4):1819-18295. Akidemir D, Jannink JL, Isidro-Sánchez J. Locally epistatic models for genome-wide prediction and association by importance sampling. Genet Sel Evol. 2017; 49(1):74/s12711-017-0348-8

6. Long N, Gianola D, Rosa GJ, Weigel KA. Marker-assisted prediction of non-additive genetic values. Genetica. 2011;139(7):843-854.

7. Howard R, Carriquiry AL, Beavis WD. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. G3. 2014; 4 (6):1027-1046.

8. Gianola D, Okut H, Weigel KA, Rosa GJ. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 2011.12(1):87. *(doi:10.1186/1471-2156-12-87)*

9.Asvadi A, Karami M, Baleghi, Y. Efficient object tracking using optimized K-means segmentation and radial basis function neural networks. Int J Res Eng Technol. 2011;4(1):29-39.

10. González-Camacho JM, de Los Campos G, Pérez P, Gianola D, Cairns JE, Mahuku G, Babu R, Crossa J. Genome-enabled prediction of genetic values using radial basis function neural networks. Theor Appl Genet. 2012;125(4):759-771.

370 11. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos
371 G, Burgueño J, Camacho-González JM, Pérez-Elizalde S, Beyene Y, Dreisigacker S.
372 Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci.
373 2017;22(11):961-975.

374 12. Long N, Gianola D, Rosa GJ, Weigel KA. Dimension reduction and variable selection for
375 genomic selection: application to predicting milk yield in Holsteins. J Anim Breed Genet .
376 2011;128(4):247-257.

377 13. Cruz CD, Nascimento M. Inteligência Computacional aplicada ao melhoramento
378 genético. 1st ed. Vicosa: Editora UFV; 2018.14. Chen S, Cowan CF and Grant PM. Orthogonal
379 least squares learning algorithm for radial basis function networks. IEEE Trans Neural Netw
380 Learn Syst. 1991; 2(2):302-309.
381

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

383 16. Asvadi A, Karami M, Baleghi M. Efficient Object Tracking Using Optimized K-means
384 Segmentation and Radial Basis Function Neural Networks. Int J Eng Res Appl. 2011;4:(1),
385 29-39.

386 17. Montgomery DC, Peck EA, Vining GG. Introduction to linear regression analysis. New
387 York: John Wiley and Sons. New York: J. Wiley, 1982. 504p.

388 18. Cruz CD. Genes Software-extended and integrated with the R, Matlab and Selegen Acta
389 Sci Agron. 2016;38(4):547-52.

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

392 20. Azevedo CF, de Resende MD, Fonseca F, Lopes PS, Guimarães SE. Regressão via
393 componentes independentes aplicada à seleção genômica para características de carcaça em
394 suínos. PAB. 2013;48(6):619-626.

395 21. Azevedo CF, Silva FF, de Resende MD, Lopes MS, Duijvesteijn N, Guimarães SE, Lopes
396 PS, Kelly MJ, Viana JM, Knol EF. Supervised independent component analysis as an
397 alternative method for genomic selection in pigs. J Anim Breed Genet. 2014;131(6):452-61.

398 22. Long N, Gianola D, Rosa GJ, Weigel KA, Avendano S. Machine learning classification
399 procedure for selecting SNPs in genomic selection: application to early mortality in broilers.
400 . J Anim Breed Genet. 2007;124(6):377-89.

401 23. Denis M, Bouvet JM. Genomic selection in tree breeding: testing accuracy of prediction
402 models including dominance effect. BMC Proc.2011; 5(7): O13.

403 24. Almeida Filho JE, Guimarães JF, e Silva FF, de Resende MD, Muñoz P, Kirst M,Resende
404 Jr MF. The contribution of dominance to phenotype prediction in a pinebreeding and simulated
405 population. Heredity. 2016;117(1):33-41.

406 25. Santos VS, Martins Filho S, de RESENDE MD, Azevedo CF, Lopes PS, Guimarães SE,
407 Silva FF. Genomic prediction for additive and dominance effects of censored traits in pigs.
408 Genet Mol. Res. 2016, 15(4)/gmr15048764.

409 26. Xu Y, Wang X, Ding X, Zheng X, Yang Z, Xu C, Hu Z. Genomic selection of agronomic
410 traits in hybrid rice using an NCII population. Rice. 2018;11(1):32/ s12284-018-0223-427.

411 27. Gianola D, Fernando RL, Stella A. Genomic-assisted prediction of genetic values with
412 semi-parametric procedures. Genetics. 2006; 173:1761–1776.

413 28. Weigel KA, de Los Campos G, Vazquez AI, Rosa GJ, Gianola D, Van Tassell CP.
414 Accuracy of direct genomic values derived from imputed single nucleotide polymorphism
415 genotypes in Jersey cattle. J Dairy Sci. 2010;93(11):5423-5435.

416 29. Weigel KA, Van Tassell CP, O'Connell JR, VanRaden PM, Wiggans GR. Prediction of
417 unobserved single nucleotide polymorphism genotypes of Jersey cattle using reference panels
418 and population-based imputation algorithms. J Dairy Sci. 2010;93(5):2229-2238.
419

420 30. Pérez-Rodríguez P, Gianola D, González-Camacho JM, Crossa J, Manès Y, Dreisigacker
421 S. Comparison between linear and non-parametric regression models for genome-enabled
422 prediction in wheat. G3. 2012;2(12):1595-1605.
423

424 31. Felipe, V.P., Okut, H., Gianola, D., Silva, M.A. and Rosa, G.J., 2014. Effect of genotype
425 imputation on genome-enabled prediction of complex traits: an empirical study with mice data.
426 BMC genet., 15(1):149.
427

428 32. Holland JB. Theoretical and biological foundations of plant breeding. In: Lamkey KR, Lee
429 M,editors. Plant breeding: the Arnel R Hallauer International Symposium. Blackwell
430 Publishing; 2006. pp. 127-140
431

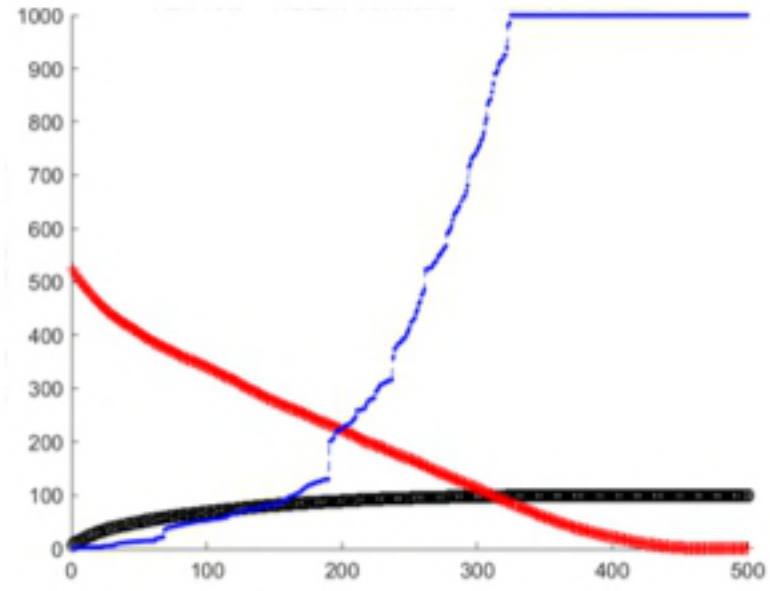
432 33. Dudley JW. Epistatic interactions in crosses of Illinois high oil 9 Illinois low oil and of
433 Illinois high protein 9 Illinois low protein. Crop Sci. 2008; 48:59–68.
434

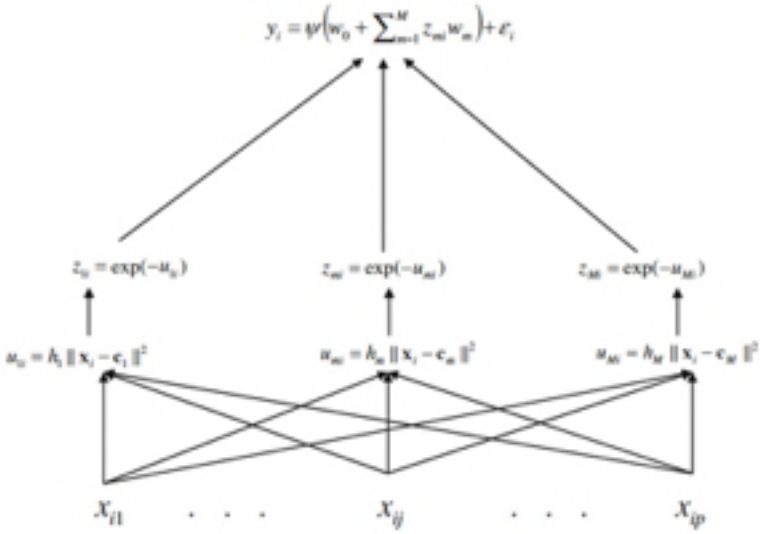
435 34. Zheng S, Li Z, Wang H A genetic fuzzy radial basis function neural network for structural
436 health monitoring of composite laminated beams. Expert Syst Appl. 2011; 38:11837–11842.
437

438 35. Dudley JW, Johnson GR .Epistatic models improve between years prediction and prediction
439 of testcross performance in corn. Crop Sci. 2010;50:763–769.
440
441 36. Denis M, Bouvet JM. Genomic selection in tree breeding: testing accuracy of prediction
442 models including dominance effect. BMC Proc. 2011; 5(Suppl7): O13.
443
444 37. Viana JM, Piepho HP. Quantitative genetics theory for genomic selection and efficiency of
445 genotypic value prediction in open-pollinated populations. Sci Agric. 2017;74(1):41-50.
446
447 38. Lee SH, van der Werf JH, Hayes BJ, Goddard ME, Visscher PM. Predicting unobserved
448 phenotypes for complex traits from whole-genome SNP data. PLoS genet.
449 2008;4(10):e1000231.

450

451


1000
900
800
700
600
500
400
300
200
100
0
0
100
200
300
400
500


$y_i = \psi\left(w_0 + \sum_{m=1}^{M} z_{mi} w_m\right) + \varepsilon_i$
$z_{1i} = \exp(-u_{1i})$
$z_{mi} = \exp(-u_{mi})$
$z_{Mi} = \exp(-u_{Mi})$
$u_{1i} = h_1 \| \mathbf{x}_i - \mathbf{c}_1 \|^2$
$u_{mi} = h_m \| \mathbf{x}_i - \mathbf{c}_m \|^2$
$u_{Mi} = h_M \| \mathbf{x}_i - \mathbf{c}_M \|^2$
X_{i1}
X_{ij}
X_{ip}