GMR Genetics and Molecular Research
Online Journal ISSN:1676-5680

# Superiority of artificial neural networks for a genetic classification procedure

**I.C. Sant’Anna[1,4], R.S. Tomaz[3], G.N. Silva[2,4], M. Nascimento[2,4], L.L. Bhering[1] and C.D. Cruz[1,2,4]**

[1]Programa de Pós-Graduação em Genética e Melhoramento,
Universidade Federal de Viçosa, Viçosa, MG, Brasil
[2]Programa de Pós-Graduação Estatística Aplicada e Biometria,
Universidade Federal de Viçosa, Viçosa, MG, Brasil
[3]Universidade Estadual Paulista “Júlio de Mesquita Filho”,
Dracena, SP, Brasil
[4]Laboratório de Bioinformática Universidade Federal de Viçosa,
Viçosa, MG, Brasil

Corresponding author: I.C. Sant’Anna
E-mail: isabela.santanna@ufv.br

Genet. Mol. Res. 14 (3): 9898-9906 (2015)
Received February 13, 2015
Accepted May 8, 2015
Published August 19, 2015
DOI http://dx.doi.org/10.4238/2015.August.19.24

**ABSTRACT.** The correct classification of individuals is extremely important for the preservation of genetic variability and for maximization of yield in breeding programs using phenotypic traits and genetic markers. The Fisher and Anderson discriminant functions are commonly used multivariate statistical techniques for these situations, which allow for the allocation of an initially unknown individual to predefined groups. However, for higher levels of similarity, such as those found in backcrossed populations, these methods have proven to be inefficient. Recently, much research has been devoted to developing a new paradigm of computing known as artificial neural networks (ANNs), which can be used to solve many statistical problems, including classification problems. The aim of this study was to evaluate the feasibility of ANNs as an evaluation technique of genetic diversity by comparing their performance with that of traditional methods. The discriminant functions were equally ineffective in discriminating the populations, with error rates of 23-82%, thereby preventing the correct discrimination of individuals between populations. The ANN was effective in classifying populations with low and high differentiation, such as those derived from a genetic design established from backcrosses, even in cases of low differentiation of the data sets. The ANN appears to be a promising technique to solve classification problems, since the number of individuals classified incorrectly by the ANN was always lower than that of the discriminant functions. We envisage the potential relevant application of this improved procedure in the genomic classification of markers to distinguish between breeds and accessions.

**Key words:** Artificial Intelligence; Discrimination; Similarity; Statistics

## INTRODUCTION

Plant breeding is the most useful strategy to sustainably increase productivity while preserving genetic diversity (Barbosa et al., 2011). Genetic diversity in germplasm collections can be studied in many ways, either with the use of genetic markers or based on qualitative/quantitative agronomical traits. In both cases, several statistical techniques can be used to predict the level of diversity. Studies of genetic diversity provide information on the possibility of preliminary selection of superior accessions and the successful use of these genotypes in breeding programs (Cruz et al., 2012), and can also facilitate genebank management, thereby saving time and resources.

To achieve these goals, multivariate statistics have been widely applied to measure genetic diversity. This approach is used to characterize the genetic structure of populations as a selective indicator of promising genotypes, as well as for germplasm conservation (Viana et al., 2006; Oliveira et al., 2007; Barbosa et al., 2011). However, as in most fields of biological sciences, the results of statistical analysis based on traditional algorithms tend to be unsatisfactory, particularly for classification analysis. According to Reby et al. (1997), these methods are often rather inefficient when the data are non-linearly distributed, even after variable transformation.

In this sense, artificial neural networks (ANNs) represent a particularly interesting tool to deal with these problems, since these learning machines can act as universal approximators of complex functions (Gianola et al., 2011). These ANNs have been applied in the context of agriculture in different ways, e.g., to identify the early stages of pest or disease development, classify satellite images (Barbosa et al., 2011), and to establish evaluation categories of growth and yield models (Castro et al., 2013), among others. Furthermore, performing analyses using computational methods that are capable of “learning” represent a breakthrough for studies involving statistical procedures in genetics (Bishop, 2007).

In this context, the purpose of this study was to ratify the efficiency of ANN in studies on genetic diversity or population discrimination, by conducting the first comparison of different techniques at higher levels of difficulty of discrimination. This study is applicable to cases in which there is prior knowledge of the topological structure of the populations tested, since the information used was extracted from simulated data derived from a backcrossed design, which is widely used in plant and animal breeding. The approach of using backcrossed populations was based on the fact that breeders generally consider the recovery of a recurrent parent after 4 to 5 backcrossed generations, making the different lines indistinguishable except with regard to the trait under selection. Thus, overcoming this in distinguish hability threshold would provide evidence of the potential of the technique and its viability for extrapolation to genetic studies of this order of complexity, which would increase the success of mutagenic, transgenic, and genomic analyses.

In this study, the feasibility of the ANN technique for evaluation of genetic diversity was investigated. A multilayer perceptron ANN was implemented that can suggest a classification and the formation of divergent groups, using simulated data in low-differentiation scenarios.

## MATERIAL AND METHODS

Genotypic data of 10 populations in Hardy-Weinberg equilibrium, with 100 plants each, were simulated. Information of 50 codominant markers was generated and used to calculate the dissimilarity matrix by Nei's genetic distance (Nei, 1972). With this index, the pair of the most divergent populations was selected to simulate genotypes in a hierarchical mating system consisting of 10 populations, each with 100 plants, obtained from five backcrossing (Bc) generations (Figure 1).

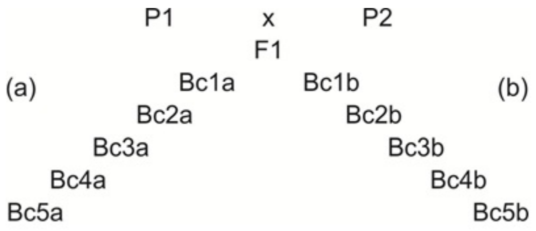

**Figure 1.** Structured diagram of backcrossing between the pair of the most divergent populations (P1 and P2) and their backcrosses (**a**) and (**b**).

The genotypic information of 13 populations ($\pi_i$ with $j = 1$ for P1; $j = 2$ for P2; $j = 3$ for F1; $j = 4, 5, \ldots, 8$ for BcXa; and $j = 9, 10, \ldots, 13$ for BcXb) was used to simulate the phenotypic values of 13 quantitative traits. Each trait was assumed to be controlled by 20 random loci, with differential additive effects determined by weights given by a binomial distribution, representing the importance of the locus in the total genotypic variability of the trait, and the mean degree of dominance equal to zero. We used $Y_{ij} = \mu + G_i + \varepsilon_{ij}$, where: $Y_{ij}$ corresponds to the phenotypic value; $\mu$ is the overall trait mean; $G_i$ the genetic effect associated with the *i*th individual of the *j*th population, given by the weighted sum of the effects of each explanatory marker of the trait; $\varepsilon_{ij}$ is the random error, and $\varepsilon_{ij} \sim N(0, \sigma^2)$; assuming heritability ($h^2$) values of 20, 25, 30, ..., 80%, and numerical mean values equal to the heritability.

The simulated characteristics were separated into two categories: Category 1 for low-heritability traits ($h^2$ = 20, 25, ..., 50%), and Category 2 for high-heritability traits ($h^2$ = 55, 60, ... 80%). The population set for each category of simulated characteristics was considered in six scenarios (1 to 6) with distinct differentiation degrees determined by the similarity degree of the populations involved. The distinctiveness scenarios (DS) to be analyzed by the equations of the discriminant functions and the ANN are presented in Table 1. The processes of genotypic, phenotypic, and breeding simulations were performed using the simulation module of the GENES software (Cruz, 2013).

**Table 1.** Constitution of differentiation scenarios to be analyzed by the discriminant functions and artificial neural networks.

| DS | Genetic design | Sample size |
|---|---|---|
| 1 | P1, P2, F1 | 300 |
| 2 | P1, P2, F1, Bc1 | 500 |
| 3 | P1, P2, F1, Bc1, Bc2 | 700 |
| 4 | P1, P2, F1, Bc1, Bc2, Bc3 | 900 |
| 5 | P1, P2, F1, Bc1, Bc2, Bc3, Bc4 | 1100 |
| 6 | P1, P2, F1, Bc1, Bc2, Bc3, Bc4, Bc5 | 1300 |

## Discriminant analysis

This procedure was adopted under the assumption that the group to which the accessions belong to is known information. Thus, the consistency of the grouping was verified using Fisher's (1936) and Anderson's (1958) discriminant analysis methods, as described by Cruz et al. (2014). Using the discriminant functions and data of the proper populations $\pi_j$, the apparent error rate (AER) was estimated, which measures the efficiency of these functions to classify the accessions correctly in the previously established populations.

The AER was determined by the ratio between the number of erroneous classifications and the total number of classifications (Cruz et al., 2014), according to: $AER(\%) = \frac{1}{N}\sum_{j=1}^{13} m_j$, where $m_j$ is the number of observations of population $\pi_j$ that were classified into another population by the discriminant functions, $\pi_{j\prime}$, where $j' = j$ and $j$ $=$ 1, 2, ..., 13 populations; considering: $N = \sum_{j=1}^{100} n_j$, where $n_j$ is the number of observations related to population $\pi_j$.

## Simulated amplification of experimental data for network training

For training purposes of the ANNs, data were amplified in a process described by Silva et al. (2014), in the case of data obtained from experiments in a randomized complete block design. Information of 2600 genotypes was generated for the ANN training set. The process of data amplification has the unique feature of preserving the data characteristics of the original populations and, as shown in this study, can be a viable alternative in common practical situations in which only a small amount of data are available.

## ANN construction and evaluation

Feed-forward back propagation multilayer perceptron networks were created using the Matlab software version 7.10.0. (The MathWorks Inc., Natick, MA, USA) and the integration module in the program GENES (Cruz, 2013). The training algorithm *trainbr* was used, along with a network architecture consisting of three hidden layers, tansig or log sig activation functions, with the number of neurons varying from 6 to 15 in the first layer, 10 to 40 in the second layer, and 10 to 40 in the third layer. Literature reviews showed that many discriminatory studies have used a lower number of neurons, although preliminary analyses indicated the need to expand this number, which reflects the degree of complexity and scope of a study. In certain situations, the study objective is the discrimination of a parent and the fifth backcross generation in which the similarity degree, determined by quantitative genetics, exceeds 98%.

The maximum number of iterations (or epochs) was 2000. All combinations of neuron numbers and activation functions in the hidden layers were checked.

Two different ANN architectures, one for each category of characteristics, were considered. Six and seven entries (corresponding to the different characteristics evaluated) for Category 1 and 2, respectively, were considered. The output layer consisted of one neuron and the output was represented by the $\pi_j$ value estimated by the ANN. An example of the ANN architecture is provided in Figure 2.

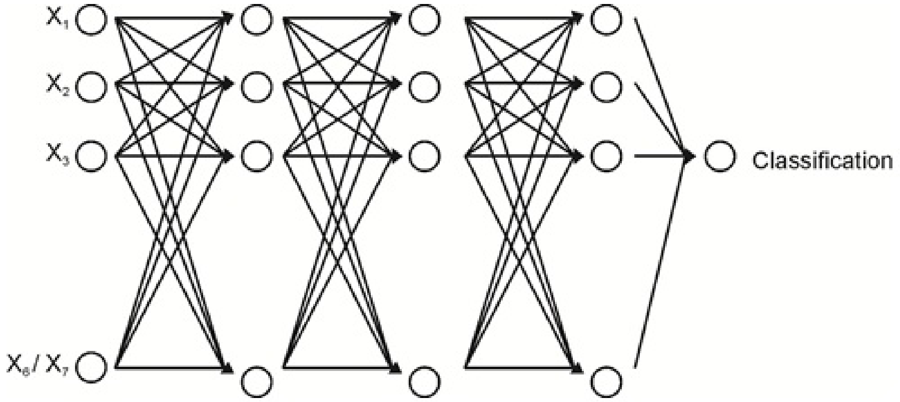

**Figure 2**. Architecture of the artificial neural network (ANN). Entries (X1) to (X 6/X 7) in the input layer are related with the simulated characteristics and are considered input. The hidden layers consisted of $n_i$ (*i* ranging from 1 to 15 or 40 nodes), with tansig or logsig activation functions. All combinations were explored. In the output layer, the RNAs were returned to classifiy the individual to their population.

## RESULTS AND DISCUSSION

### Fisher and Anderson discriminant analyses

The Fisher and Anderson discriminant analyses were equally ineffective in discriminating the populations, with error rates of 23-82%, thereby preventing the correct discrimination of individuals among different populations (Table 2).

For the 1st DS, the AER ranged from 23 to 27% considering Categories 1 and 2, which is considered unsatisfactory. For the 2nd to 6th DS, difficulty of discrimination was expected owing to advancement of the backcross populations. Assuming equitable gametic contribution to the advancement of generations, the recovery of the recurrent parent in generation *x* occurs at a ratio of $[(2^{x+1} - 1) / (2^{x+1})]$. Therefore, the similarity between $Bc_5$ and the recurrent genitor is approximately 98%. This degree of similarity makes the discrimination among different backcross populations and their recurrent parents rather difficult. Thus, the AER measured in the DS containing backcrossed populations was higher than 50%, exceeding 80% in the 6th DS. These results demonstrate that the procedures based on multivariate discriminant functions are unsatisfactory for use in the discrimination of populations derived from controlled crosses.

**Table 2.** Apparent error rate (%) estimated by Fisher's discriminant functions (FIS) and Anderson's discriminant function (AND) for the discrimination of populations in Categories 1 and 2 in the distinctiveness scenarios (DS).

| DS | Category 1 | | Category 2 | |
|---|---|---|---|---|
| | FIS | AND | FIS | AND |
| 1 | 27 | 27 | 23 | 23 |
| 2 | 53 | 53 | 55 | 55 |
| 3 | 66 | 66 | 68 | 68 |
| 4 | 75 | 75 | 74 | 74 |
| 5 | 78 | 78 | 78 | 78 |
| 6 | 81 | 82 | 80 | 80 |

In this context, it is worth emphasizing that discriminant analysis techniques have been used successfully in many breeding projects, and are thus applicable for different purposes in some simple scenarios. However, with the advent of molecular genetics, in which a relatively large amount of information is available and the genotypic differences are very subtle, both in effect and expression, alternative techniques of discrimination have become necessary. In practical experiments, it is rather difficult to establish slightly or very different population pairs, due as much to experimental difficulties as to theoretical aspects, since the effects of factors such as selection, gene flow, genetic drift, and the mating system are difficult to quantify, and thus hamper the acquisition of prior knowledge of work scenarios.

Table 3 shows the results obtained with the ANN. The AER in the training stage was zero in the first four DS in both trait categories, indicating the efficiency of the ANN in unequivocally differentiating populations up to the 3rd backcross generation. Even when considering the 5th DS, which involves $Bc_4$ populations, the discriminating power of the ANN was large, with AER ≤ 2% in both categories. When assessing the 6th DS, AER values lower than 20% were observed in the training stage as well as in the validation stage, considering both trait types. These results are superior to those obtained by the discriminant functions, indicating the great potential of ANN for population discrimination.

**Table 3.** Apparent error rate (%) estimated by the artificial neural network (ANN), in the training (tr.) and validation (val.) stages for the discrimination of populations in Categories 1 and 2 in the distinctiveness scenarios (DS).

| DS | Category 1 | | Category 2 | |
|---|---|---|---|---|
| | ANN tr. | ANN val. | ANN tr. | ANN val. |
| 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 |
| 5 | 1 | 0 | 1 | 0 |
| 6 | 19 | 15 | 19 | 13 |

Results related to the differentiation of populations are highly important, both from the aspect of breeding as well as in studies on the adaptation and evolution of population groups (Cruz et al., 2011). Studies on population or genotype diversity and differentiation are essential for classical breeding programs aimed at establishing heterotic groups and for identifying hybrid combinations with greater vigor, and are equally valuable in genome analyses with the aim to establish a minimum difference between genotypes, at the nucleotide level, in relation to gene expression levels. Our results demonstrate the unfeasibility of using biometric techniques for discriminant analysis to differentiate populations with a medium to high similarity degree, such as backcross derivatives, due to the high misclassification rate. In this sense, the ANN proved to be superior, providing very satisfactory results.

ANNs have been increasingly used in agriculture to solve problems of subjectivity related to the classification of genotypes. Oliveira et al. (2013) used ANNs for the pre-selection of polyploid banana. The authors demonstrated that the ANNs correctly classified 10 of the 11 samples used for validation and that the implemented ANNs were effective for pre-selecting desirable polyploid banana. Espinhosa and Galo (2004) used an ANN to classify water and aquatic plants. The ANN allowed for a clear separation of the different occurrences of geophytes and variations in the water. Nascimento et al. (2013) used ANNs to classify alfalfa genotypes, confirming their superiority over commonly used methodologies.

Attention should also be paid to the fact that the successful application of ANNs depends on factors such as the size and quality of the data set used for training; in particular, the data in the training phase must be representative and sufficiently large in order to ensure the reliability of estimates in the validation stage (Kavzoglu, 2009). In this study, the option of using an expanded database that could preserve the data structure, such as the mean vector and variance-covariance matrix, proven to be a viable alternative for the purpose of training the ANN. It should also be noted that, according to Braga et al. (2011) and Haykin (2001), an excessive increase in the number of neurons can lead to a loss of generalization of ANNs.

Information about the architecture of the ANN to be used is also very important, given the lack of this type of information in the scientific literature. Table 4 shows the number of neurons and their activation functions that provided the most accurate ANNs.

**Table 4.** Architecture of the most accurate artificial neural networks (ANNs). The number of neurons and activation function are listed for each hidden layer ($L_1$, $L_2$, and $L_3$) in the six distinctiveness scenarios (DS).

| DS | Number of nodes | | | Activation function* | | |
|---|---|---|---|---|---|---|
| | $L_1$ | $L_2$ | $L_3$ | $L_1$ | $L_2$ | $L_3$ |
| 1 | 6 | 10 | 18 | *tansig* | *tansig* | *logsig* |
| 2 | 6 | 30 | 30 | *tansig* | *tansig* | *tansig* |
| 3 | 15 | 30 | 35 | *tansig* | *logsig* | *logsig* |
| 4 | 15 | 35 | 40 | *tansig* | *tansig* | *logsig* |
| 5 | 15 | 40 | 40 | *logsig* | *tansig* | *logsig* |
| 6 | 15 | 40 | 40 | *tansig* | *tansig* | *logsig* |

*logsig = logistic sigmoid activation function used by the MATLAB software; tansig = hyperbolic tangent activation function used by the MATLAB software.

Configurations with 6 to 15 neurons were evaluated in the first layer, 15 to 40 in the second and third hidden layers, and their combinations of hyperbolic tangent activation and logistic sigmoid functions were determined. The tested configurations were appropriate for solving the problem. In the first and second hidden layers, a hyperbolic tangent activation function predominated, whereas logistic sigmoid activation functions predominated in the third hidden layer. Silva et al. (2010) reported that the number of neurons in the hidden layer used in classification problems and linear filter patterns is given by ($2i + 1$), where $i$ is the number of input variables. However, the same authors stated that this number may be insufficient in the case of problems of a very complex nature, such as that considered in the present study. Although previous reports have suggested that the second hidden layer should contain fewer neurons than the third (Silva et al., 2010), this result was not consistent with that inferred from our data, since more accurate architectures were obtained.

It should also be emphasized that the establishment of the ANN was computationally intensive, since it took 7-8 weeks to adjust the weights of the 5th and 6th DS. This long period was required for the search of the best network, and involved determining the number of neurons in hidden layers and the appropriate activation function. For this reason, it is very important to obtain information about the ANN architecture required to solve classification problems of this or a similar nature. For the 6th DS, other configurations with a higher number of hidden layers could be investigated, although according to Ardö et al. (1997), the accuracy and number of these ANN are not related.

Examples of the efficiency of ANN in studies focused on classical breeding have been reported in the literature. Barbosa et al. (2011) compared the genetic diversity in papaya (*Carica papaya* L.) accessions considering eight characteristics of a quantitative nature. They compared the performance of ANNs with Anderson's discriminant analysis for the classification of accessions, and concluded that the ANN could efficiently rank the accessions to study genetic diversity. A similar result was found by Cho et al. (2002), in a discrimination study of weed and radish plants. The authors concluded that ANNs could replace the discriminant functions when classifying plants into previously known groups, provided that a suitable ANN is established. Our results, based on a study involving populations with an *a priori* known degree of similarity, can be considered solid and consistent in contrast with the examples from the literature, since, according to Braga et al. (2011), the performance of ANNs is better than that of conventional methods. In view of the possibility of using unsupervised learning algorithms, ANNs have the advantage of not requiring prior knowledge about the number of classification groups, which is a requirement of discriminant functions (Cruz et al., 2011), thereby enabling their use in contexts beyond controlled crosses.

In summary, ANNs were effective in classifying populations with low and high differentiation, such as those obtained via a genetic design established with populations derived from backcrosses, even in cases of low differentiation of the data sets.

The network structure with three hidden layers with 6 to 15 neurons in the first layer, 15 to 40 neurons in the second layer, and 15 to 40 neurons in the third layer allowed for classification of the studied backcross populations.

The tested network settings were also suitable to establish ANN with a zero error for most scenarios evaluated.

In classification analyses, the approach using ANNs performed far better than the conventional discriminant analysis methods.

## ACKNOWLEDGMENTS

Research supported by Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Fundação de Amparo à Pesquisa do Estado de Minas Gerais, and Conselho Nacional de Desenvolvimento Científico e Tecnológico.

## REFERENCES

Anderson TW (1958). An introduction to multivariate statistical analysis. John Wiley & Sons, New York.

Ardö J, Pilesjö P and Skidmore A (1997). Neural networks, multitemporal Lands at Thematic Mapper data and topographic data to classify forest damages in the Czech Republic. *Can. J. Remote Sensing* 23: 217-229.

Barbosa CD, Viana AP, Quintal SSR and Pereira MG (2011). Artificial neural network analysis of genetic diversity in *Carica papaya* L. *Crop Breed. Appl. Biotechnol.* 11: 224-231.

Bishop CM (2007). Pattern recognition and machine learning. Springer, Singapore.

Braga AP, Carvalho APLF and Ludermir TB (2011). Redes neurais artificiais - Teoria e aplicações. 2nd edn. LTV, Rio de Janeiro.

Castro RVO, Soares CPB, Martins FB and Leite HG (2013). Crescimento e produção de plantios comerciais de eucalipto estimados por duas categorias de modelos. *Pesq. Agropec. Bras*. 48: 287-295.

Cho S, Lee DS and Jeong JY (2002). AE - Automation and emerging technologies: weed-plant discrimination by machine vision and artificial neural network. *Biosyst. Eng*. 83: 275-280.

Cruz CD (2013). GENES - a software package for analysis in experimental statistics and quantitative genetics. *Acta Sci.* 35: 271-276.

Cruz CD, Ferreira FM and Pessoni LA (2011). Biometria aplicada ao estudo da diversidade genética. Suprema, Visconde do Rio Branco.

Cruz CD, Regazzi A and Carneiro PCS (2012). Modelos biométricos aplicados ao melhoramento genético. Vol. 1. 4th edn. UFV, Viçosa.

Cruz CD, Carneiro PCS and Regazzi A (2014). Modelos biométricos aplicados ao melhoramento genético. Vol. 2. 3rd. edn. UFV, Viçosa.

Espinhosa MC and Galo MDLBT (2004). O uso de redes neurais artificiais na análise da ambiguidade entre classes de água e plantas aquáticas. *Bol. Cienc. Geodésicas* 10: 193-213.

Fisher RA (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7: 179-188.

Gianola D, Okut H, Weigel KA and Rosa GJM (2011). Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. *BMC Genet*. 12: 87.

Haykin SS (2001). Redes Neurais: princípios e práticas. Trad Paulo Martins Engel., 2nd edn. BOOKMAN, Porto Alegre.

Kavzoglu T (2009). Increasing the accuracy of neural network classification using refined training data*. Environ. Modell. Sof*. 24: 850-858.

Nascimento M, Peternelli LA, Cruz CD, Nascimento ACC, et al. (2013). Artificial neural networks for adaptability and stability evaluation in alfalfa genotypes*. Crop Breed. Appl. Biotechnol*. 13: 152-156.

Nei M (1972). Genetic distance between populations. *Am. Nat*. 106: 238-292.

Oliveira MSP, Ferreira DF and Santos JB (2007). Divergência genética entre acessos de açaizeiro fundamentada em descritores morfoagronômicos. *Pesq. Agropec. Bras.* 42: 501-506.

Oliveira ACL, Pasqual M, Pio LAS, Lacerda WS, et al. (2013). Utilização da modelagem matemática (redes neurais artificiais) na classificação de autotetraploides de bananeira (*Musa acuminata Colla*). *Biosci. J*. 29: 617-622.

Reby D, Lek S, Dimopoulos I, Joachim J, et al. (1997). Artificial neural networks as a classification method in the behavioural sciences. *Behav. Proc.* 40: 35-43.

Silva ID, Spatti DH and Flauzino RA (2010). Redes neurais artificiais para engenharia e ciências aplicadas. Artliber, São Paulo.

Silva GN, Tomaz RS, Sant'Anna IC, Nascimento M, et al. (2014). The use of neural networks for predicting breeding values and genetic gains. *Sci. Agricola* 71: 494-498.

Viana AP, Pereira TNS, Pereira MG, Souza MM, et al. (2006). Genetic diversity in yellow passion fruit populations. *Crop Breed. Appl. Biotechnol*. 6: 87-94.