Optimizing mate selection: a genetic algorithm approach

D. de Carvalho Neves da Fontoura[1,2], S. da Silva Camargo[1], R. Augusto de Almeida Torres Junior[3], H. Gomes de Carvalho[4] and F. Flores Cardoso[1,4]

[1]Graduate Program on Applied Computing - Federal University of Pampa and Embrapa South Livestock, Bagé, RS, Brazil
Corresponding Author: fernando.cardoso@embrapa.br
[2]Federal Institute of Education, Science and Technology, Bagé, RS, Brazil
[3]Embrapa Beef Cattle, Brazilian Agricultural Research Corporation, Campo Grande, MS, Brazil
[4]Embrapa South Livestock - Brazilian Agricultural Research Corporation, Bagé, RS, Brazil

Abstract

Background: Genetic Improvement Programs (GIP) aim to enhance production efficiency of beef cattle. The main way to guide this enhancement is by choosing the best mates among sires and cows, in order to maximize the offspring Genetic Qualification Index (QGI), which is measured by an index defined by the GIP and computed for each animal of the herd. This paper describes a genetic algorithm, which can recommend an optimal set of matings among sires and cows, in order to maximize the QGI of the herd. Breeders can define constraints regarding level of problems, which must be avoided, and they also can alter the traits relative importance considered in QGI, according their particular interests. This algorithm was applied to a herd of a Brazilian breeder, which participates of a GIP, and it found optimal matings in order to increase QGI value. We have simulated different scenarios considering variations on fitness functions, which combine QGI and level of problems, in order to find the optimal matings. Proposed approach was successfully used to recommend optimal mating decisions by Brazilian Hereford and Braford cattle breeders Association leading to an improvement of offspring QGI.

Keywords: Genetic Improvement, Beef Cattle, Artificial Intelligence, Evolutionary Computing.

Background

In recent decades, genetic improvement has been used as an approach for enhancing production efficiency of beef cattle. In order to measure and guide this enhancement, recording and evaluation programs have been collecting phenotypic, genetic and pedigree relationship data about the animals (Miller, 2010). Data typically involves values of economic importance traits, which should be optimized. The main way to guide this optimization provided by a genetic improvement program is by choosing the best mates among sires and cows.

Some examples can be found in literature regarding mate selection for genetic improvement using artificial intelligence techniques, like evolutionary computing or genetic algorithms (GA). Carvalheiro; Queiroz and Kinghorn (2010) developed a solution, based on the Differential Evolution (DE) technique, which searches for the

optimal genetic contribution during the selection of reproduction candidates. The fitness function used Expected Progeny Difference (EPD) data and penalty restriction for inbred matings. The results show the program was computationally efficient and feasible to be applied in practical situations. Expected consequences of its application, when compared to empirical inbreeding control procedures and/or selection based only on the expected genetic value, would be the improvement of future genetic response and more effective limitation of inbreeding rate. Authors concluded that it is possible to use differential evolution as an optimization method to make the optimal selection of genetic contribution. Kinghorn (2011) described a mating selection algorithm, which has an extension allowing the application of restrictions in certain matings, according the groups to which sires and cows belong. As a result, this algorithm is faster than the DE from its previous publication and it can penalize constraint-breaking solutions. This way, the performance of this new algorithm extended the use of partner selection and allowed implementation in relatively large genetic improvement programs. Barreto Neto (2014) has applied the genetic contribution theory to optimize the next generation genetic value of 12 selective matings nuclei, which contained 500 Santa Inês ewes. The author used GA to find the optimal genetic contribution to the next generation of these animals, which have a structured pedigree and EPDs of economic importance traits estimated through Best Linear Unbiased Predictors (BLUP). The results showed the effectiveness of the use of animal selection using BLUP-EPD, as well as the efficiency of the GA in the process. Results also shown the increasing exchange of genetic material between nuclei is highly recommended to increase genetic gain and Artificial Intelligence may be a method to achieve this goal, if an inbreeding control measure is also being used.

In this context, this paper aims to describe development and testing of an evolutionary computing tool to optimize mating decisions by beef cattle breeders. This solution was successfully implemented in a Genetic Improvement Program. Among specific objectives, can be cited:

1. Use a customizable index for herd specific breeding objectives;
2. Use a penalty for offspring inbreeding;
3. Use a definable minimum and maximum number of offspring per parent; and
4. Use a penalty for low performance on independent culling traits.

The remainder of this paper was divided into three sections. In Material and Methods, we present requirements and decisions taken during the process of solution implementation, the Results and Discussion section approaches the experiments done and a critical analysis of results obtained, and, finally, in Conclusions we summarize some findings and limitations of the solution.

Material and methods

Approval of Animal care and use committee was not needed due to the usage of existing datasets historically collected by the animal breeding program. Experiments were done using R version 3.5.2 (R Core Team, 2018), GA package version 3.2 (Scrucca, 2017), and Rcpp package version 1.0.2 (*https://cran.r-project.org/web/packages/Rcpp/index.html*).

Data source

Data source was provided by the PampaPlus Hereford and Braford GIP conducted by the Brazilian Hereford and Braford Association and Embrapa Pecuária Sul, both institutions located in Bagé, Rio Grande do Sul, Brazil (-31.33; -54.10) (Cardoso *et al.*,

2016). PampaPlus controls performance of herds located in several states of Brazil, and in Uruguay and Paraguay. Sires and cows data, including their EPD traits data, were selected directly from the database of PampaPlus GIP. Main parameters for the optimization include maximum and minimum utilization for each sire and maximum inbreeding value for the calves. Fourteen EPDs were available in our dataset: Birth Weight (BW), Weaning Weight (WW), WW Maternal - Milk (WWm), Total Maternal (TM), Yearling Weight (YW), Post Weaning Gain (PWG), Mature Cow Weight (MCW), Scrotal Circumference (SC), Muscling Score (MSC), Height Scores (HSC), Body Capacity Score (BCS), Cow Body Score (CBC), Navel Size Score (NSC), and Eye Pigmentation (EP).

Proposed solution

Developed genetic algorithm followed the canonical model presented by Goldberg (1989). Each chromosome represents a possible solution for the matings, defining a set of mates among sires and cows. Chromosomes are composed by genes, which amount is equals to the number of cows in a simulation. The content of each gene is the identifier of a sire, which will mate with the cow. The fitness function, which evaluates the quality of each calf generated by the mating, can be computed through any combination of trait values. In our simulations, we have used the traits their respective weights as defined in the PampaPlus GIP Index

(QGI = 30%TM + 15%PWG + 15%YW + 12.5%MSC + 12.5%HSC + 15%SC)

However, each breeder in each simulation can alternatively set traits and weights according to a specific breeding objective. The quality of each solution, or chromosome, is computed by the averages of each fitness values of each gene. As result, the approach will search for the best solutions, or chromosomes, which maximize the fitness function.

Among the fourteen EPD traits measured in the PampaPlus GIP, three of them that are included in the Index should be minimized, namely BW, MCW and NSC. That is, the lower the value is, the better is the calf. This way, the weight of these EPD are negative in fitness function that is, the decreasing of them leads to a increasing of the fitness value.

Generation of new chromosomes must satisfy some constraints, which leads to a penalization of non-valid chromosomes. These restrictions are:

1. User can set minimum and maximum amount of matings to each bull.
2. The maximum calf inbreeding for each mating must be respected.

The first step in a GA execution involves generating an initial set of random chromosomes, which size must be defined. We have used the chromosome population size as twice the amount of cows in the simulation (Carvalheiro; Queiroz and Kinghorn, 2010; Kinghorn, 2011). Random chromosomes which does not obey the constraints are penalized in 50% of their fitness.

We have search for ways to speedup the finding of the best chromosomes. At first, we have used the addicted roulette selection, which gives to individuals with higher fitness value greater odds to be selected for the crossover. Because invalid chromosomes are penalized by the fitness function, valid chromosomes are more likely to be selected. During the tests, convergence was slow. Afterwards, we have combined addicted roulette and tournament. These hybrid technique led to a most efficient selection of valid chromosomes for reproduction. This way, the final implementation of the selection works as follows: 2 chromosomes are chosen through the addicted roulette and a

tournament is held between these two chosen chromosomes, selecting the one with highest fitness value. The process is repeated to select the second parent (chromosome).

After parent selection to the reproduction process, selected parents are combined through the crossover process. Crossover is performed based on a random choice of a position of the chromosomes. Two new chromosomes are generated, combining the old ones. Another concept used was the mutation rate, which is an adjustable probability from 0 to 100%. In order to mutate a gene, a randomly chosen sire replaces the original sire previously defined to mate a cow on that gene.

The stopping condition must be defined to a GA. In our experiments, depending on the parameters, we have verified a convergence between 400 and 800 generations, which is the name given to each cycle where crossover occurs in all chromosomes. This way, we have defined 1000 generations as stopping criteria. When reaching the stopping criteria, the best valid chromosome is indicated as the optimal mating combination.

Results and discussion

We have done four simulations, testing different values for Genetic Algorithm parameters in order to evaluate and optimize the results obtained by our approach. We also have used an actual database, provided by Brazilian Hereford and Braford Cattle Breeders Association. We have selected a single breeder, which represents a typical case in terms of amount of animals owned. In the experiments, we had used 568 cows and 37 sires. Table 1 represents the summary of the results obtained in these four simulations.

The first simulation reported here aims to verify if the GA can find a solution by selecting the best subset of available sires. We have defined 3% as the constraint for maximum inbreeding and 30 as the maximum utilization of each sire. No constraint for minimum utilization was defined. The evaluation fitness of calves generated by each mating was computed by using the PampaPlus QGI index. Thus, sires with higher QGIs were supposed to be chosen and matched with cows in order to maximize the value of the mating fitness. However, Table 1 shows that 48% of proposed matings (275 out of 568) have some level of problem regarding poor performance for undesirable traits, namely in this experiment BW, NSC, and EP. For these traits, a value of one standard deviation below average of the active animals of the PampaPlus GIP was considered the level of problem (LP) for the future progeny performance. The training process for the GA in this simulation is presented in Figure 1. The convergence happened around generation 500. The initial population average fitness was around 50 and in the final generation an average fitness above 250 was reached. The best valid solution, the best solution (which can violate some restriction), and the average of all solutions, or chromosomes, were presented in Figure 1.

While in the simulation 1, the fitness function for genetic algorithm considered only the QGI, in simulation 2 the metric was composed by a combination of 90% QGI and 10% LP. This experiment aimed to evaluate how the impact of LP could be reduced without compromising the QGI. As a result, the final QGI from matings recomended by our solution decreased around 1%, while reduction of LP was near 45%. Consequently, results shows that including LP in fitness function can generate a representative impact in reducing the level of problems, producing a small effect on herd average QGI. The training process for the GA in this simulation is presented in Figure 2. The convergence also happened around generation 500. The initial population average fitness was around

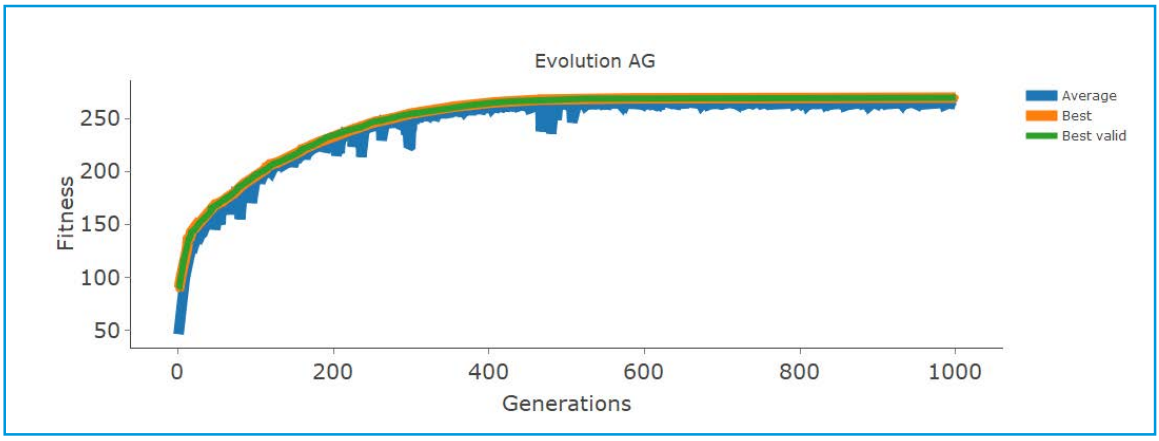

Figure 1. Evolution of fitness values using 100% QGI.

Table 1. Simulation parameters.

Parameter	Value
Amount of Sires	37
Amount of Cows	568
Population size	1136
Inbreeding	<= 3%
Mutation	10%
Stopping generation	1000
Sire maximum utilization	30
Comment	Any sire can mate with any cow up to the its maximum limit of each sire, with no minimum defined

45 and in the final generation an average fitness around 240 was reached. The best valid solution, the best solution (which can violate some restriction), and the average of all solutions, or chromosomes, were presented.

Due to positive results obtained in simulation 2, it was done a new experiment where the weight of LP was incremented to 20% and the weight of QGI was reduced to 80% in the fitness function. In this third simulation, results show a QGI decrease around 4% and 60% in LP. This training process is presented in Figure 3. The convergence happened around generation 700. The initial population average fitness was around 40 and increased to near 200.

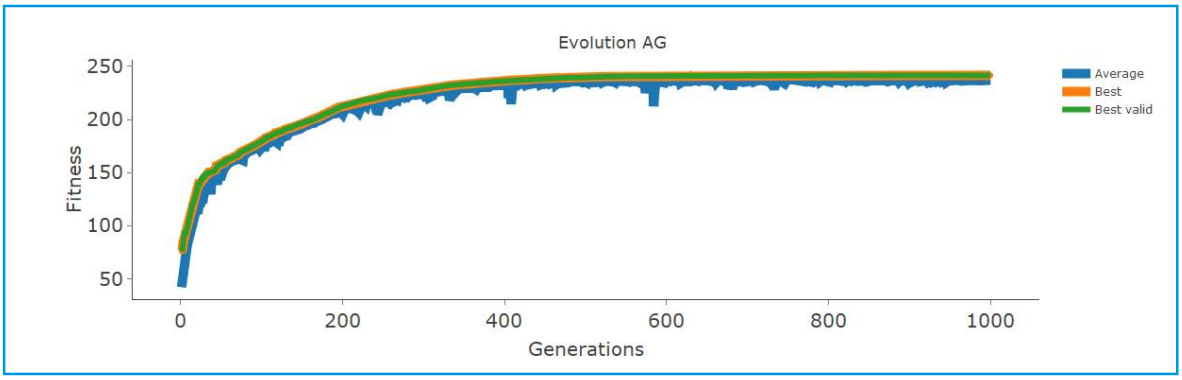

Figure 2. Evolution of fitness values using 90% QGI and 10% LP.

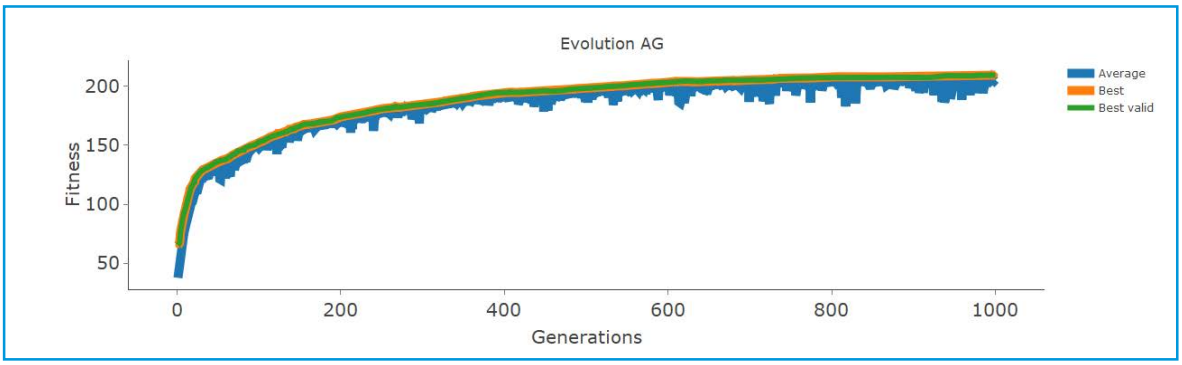

Figure 3. Evolution of fitness values using 80% QGI and 20% LP.

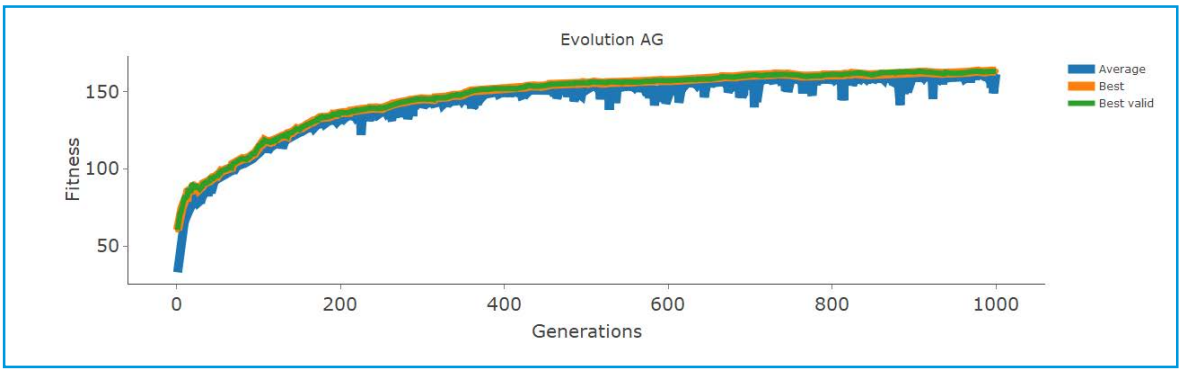

Figure 4. Evolution of fitness values using 70% QGI and 30% LP.

In simulation 4, the QGI represents 70% of the fitness function value while the LP represents 30%. As result, the mating QGI decreased around 14%. Moreover, the LP reduction was near to 87%. Figure 4 presents the evolution of GA of this simulation. The convergence happened around generation 800 and the initial population average fitness was around 35 and increased to near 150.

Table 2 summarizes the results of our simulations combining different weights for QGI and LP in the fitness function. Values of QGI and LP for the best solution are presented, and the amount of undesirable matings (p) as well.

Figure 5 presents the results of our four simulations with different combinations weights for QGI and LP in the fitness function. Considering 568 matings in our simulations, the amount of undesirable matings was divided by 568, in order to be transformed to a proportion. As can be seen in simulations, there is a slight reduction of the QGI index as the weight of QGI in fitness function also decreases. On the other hand, the decreasing of matings with some problem is more significant when the weight of LP increases on the fitness function. In this sense, when LP represents 30% of the fitness function, the amount of undesirable matings was reduced to around 5%.

Conclusions

In this paper we have presented a Genetic Algorithm based approach for optimizing mate selection in Genetic Improvement Programs of beef cattle. The approach uses Expected Progeny Difference and pedigree relationship data in order to evaluate the matings recomended by the algorithm. Different scenarios were tested, combining QGI and LP weights in fitness function. Results showed a slight reduction of QGI of the herd while reaching a significant reduction of the level of problem of calves as the weight of LP is increased in the fitness function. In this sense, our experiments shows

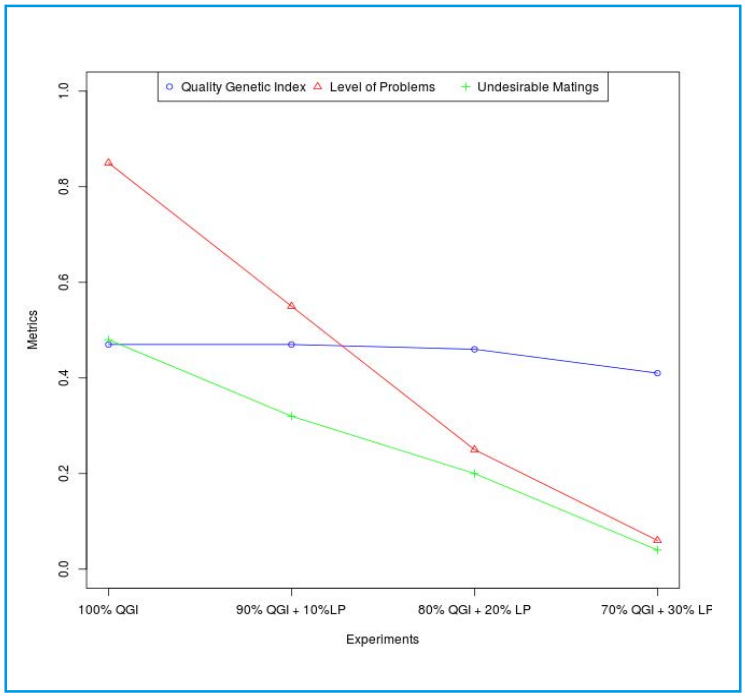

Figure 5. Results of QGI, Level of Problems and Undesirable matings in tested simulations

Table 2. Average genetic qualification index (QGI), level of problems (LP) and proportion (p) of undesirable matings according to different combinations of QGI and LP i the fitness function.

Simulation	Fitness Function	Best Solution		
		QGI	LP	p
1	100% QGI	0,474815	0,845915	275
2	90% QGI + 10% LP	0,472351	0,545775	179
3	80% QGI + 20% LP	0,459677	0,248239	112
4	70% QGI + 30% LP	0,409432	0,063380	24

evolutionary computing was successfully used to optimize mating decisions by Brazilian Hereford and Braford cattle breeders, combining index, independent level culling traits, inbreeding and offspring size.

As future works, this approach will be integrated in the Pampaplus mating tool to guide matings and increase genetic gain. Moreover, relative importance of QGI index and level of problems, in the fitness function, need to be tested in a broader range of scenarios.

List of references

Barreto Neto, A. D. Estrutura populacional e otimização de esquemas de acasalamento em ovinos com uso de algoritmos evolucionários.76 f. Dissertação (Mestrado em Zootecnia) - Universidade Federal de Sergipe, São Cristóvão, Sergipe, Brazil 2014.

Cardoso, F. F.; Lopa, T. M. B. P. and Teixeira, B. B. 2016. PampaPlus: Avaliação Genética Hereford e Braford. Embrapa Pecuária Sul, Bagé.

Carvalheiro, R., Queiroz, S. A., Kinghorn, B. Optimum contribution selection using differential evolution. Revista Brasileira de Zootecnia, v. 39, n. 7, p. 1429-1436, 2010.

Fontoura, D. C. N. Uma Solução De Recomendações De Acasalamentos Baseada Em Algoritmos Genéticos. 95 f. Dissertação (Mestrado em Computação Aplicada) - Universidade Federal do Pampa, Bagé, Rio Grande do Sul, Brazil 2014.

Goldberg, D. E. (1989). Genetic algorithms in search, optimization, and machine learning. Reading, MA: Addison-Wesley.

Kinghorn, B. P. An algorithm for efficient constrained mate selection. Genetics Selection Evolution 2011, 43:4.

Miller, S. Genetic improvement of beef cattle through opportunities in genomics. R. Bras. Zootec., v.39, p.247-255, 2010.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL *https://www.R-project.org/.*

Scrucca, L. (2017) On some extensions to GA package: hybrid optimisation, parallelisation andislands evolution.The R Journal, 9/1, 187-206.