

Use of a novel evolutionary algorithm for genomic selection

Julie Hamon, Gaël Even, Romain Dassonneville, Julien Jacques, Clarisse Dhaenens

► **To cite this version:**

Julie Hamon, Gaël Even, Romain Dassonneville, Julien Jacques, Clarisse Dhaenens. Use of a novel evolutionary algorithm for genomic selection. 2015. <hal-01100660>

HAL Id: hal-01100660

<https://hal.inria.fr/hal-01100660>

Submitted on 6 Jan 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

RESEARCH

Use of a novel evolutionary algorithm for genomic selection

Julie Hamon^{1,2*}, Gaël Even², Romain Dassonneville², Julien Jacques^{1,3} and Clarisse Dhaenens^{1,4}

*Correspondence:

julie.hamon1@gmail.com

¹Inria Lille - Nord Europe, Lille, France

²Gènes Diffusion, 3595 rte de Tournai, Douai, France

Full list of author information is available at the end of the article

†Equal contributor

Abstract

Background: In the context of genomic selection in animal breeding, an important objective is to look for explicative markers for a phenotype under study. The challenge of this study was to propose a model, based on a small number of markers, to predict a quantitative trait. To deal with a high number of markers, we propose using combinatorial optimization to perform variable selection, associated with a multiple regression model in a first approach and a mixed model in a second, to predict the phenotype.

Results: The efficiency of our two approaches, the first assuming that animals are independent and the second integrating familial relationships, was evaluated on real datasets. This reveals the importance of taking familial relationships into account as the performances of the second approach were better. For example, on PIC data the correlation is around 0.15 higher using our approach taking familial relationships into account than with the Lasso bounded to 96 selected markers. We also studied the importance of familial relationships on phenotypes with different heritabilities. Finally, we compared our approaches with classic approaches and obtained comparable results, sometimes better.

Conclusion: This study shows the relevance of combining combinatorial optimization with a regression model to propose a predictive model based on a reasonable number of markers. Although this implies more parameters to be estimated and, therefore, takes longer to execute, it seems interesting to use a mixed model in order to take familial relationships between animals into account.

Keywords: genomic selection; combinatorial optimization; regression

1

2

3 1 INTRODUCTION

4 Genomic selection of animal breeding deals with a genetic evaluation of animals
5 from their DNA (extracted using biological samples such as blood or hair, or
6 biopsy), based on markers covering the whole genome. Important insight in this
7 domain is gained by establishing predictive models using genomic information. High-
8 throughput genotyping data are analyzed in this study and an important feature of
9 these data is the huge number of markers (p) compared to the number of subjects
10 (n). So, in order to predict a quantitative trait using these data, the classic statis-
11 tical problem of high dimensional regression ($n < p$) has to be solved.

12 Various methods have been proposed, including approaches based on best linear un-
13 biased prediction (BLUP), Bayesian approaches or shrinkage regression methods.

14 The choice of which method to use usually depends on the genetic architecture of
15 the trait studied [1]. Indeed, for a given trait, if the distribution of effects is known
16 to be normal, it is preferable to use a method such as G-BLUP while if the trait
17 depends on areas of the genome with large effects, Bayesian methods are preferred.
18 The challenge of this study was to find a predictive model based on a small number
19 of markers allowing the selection of the best animals for a given phenotype, in order
20 to produce small size chips for the phenotype under study. Indeed, low density chips
21 are cheaper and it can be interesting, for example, to genotype a large amount of
22 animals with this type of chip and genotype only the best one with a high density
23 chip.

24 The problem of variable selection among a huge amount of variables can be seen as
25 a combinatorial problem [2]. We therefore proposed dealing with this problem by
26 using a combinatorial optimization approach. Modeling this problem as a combina-
27 torial optimization problem is interesting as it allows efficient methods which have
28 been developed for this kind of problem to be adopted. Here, the size of the problem
29 is very large (it depends on the number of markers), hence a complete enumeration
30 will not be possible. In this context, heuristic optimization approaches will be used.
31 Such methods have been applied for variable selection in various domains, especially
32 on microarray data or SNPs data in classification contexts. However, they can be
33 adapted to a regression problem to deal with quantitative traits such as milk pro-
34 duction or meat quality.

35 Among combinatorial optimization methods, metaheuristics are approximate algo-

36 rithms that can efficiently explore a very large search space in order to obtain a
37 satisfactory solution [3]. In this study we adopted evolutionary algorithms, which
38 are population based metaheuristics, based on Darwin's theory of evolution [4].
39 For this study, we suggested addressing the problem of variable selection in a high di-
40 mensional regression context by combining a combinatorial optimization approach
41 for selecting subsets of variables and a statistical model to evaluate this subset. An
42 interesting outcome is that the proposed algorithm affords the possibility of includ-
43 ing familial relationships. Hence, to carry out experiments, real datasets from beef
44 cattle and pigs were used to compare the proposed method with classic approaches
45 for traits with various heritabilities.

46 2 MATERIEL AND METHODS

47 2.1 Data

48 In this study, cattle and pig data are used. Cattle data come from the Qualvigène
49 project [5] in which Gènes Diffusion (www.genesdiffusion.com) is involved.

50 This program includes Charolais bulls and young bulls with 48 sires and 1,114
51 bulls. The trait studied was the carcass yields with high heritability ($h^2 = 0.54$).
52 Following pre-treatment on available animal data (including the removal of non-
53 phenotyped animals for that trait), we finally obtained 1,107 animals (48 sires)
54 genotyped in 54K. Following quality control of the genotyping data, 43,896 SNPs
55 were retained for the study. We obtained an SNP data matrix size of $1,107 \times$
56 $43,896$, associated with a vector of size equal to 1,107 for carcass yield. Values of
57 the trait studied here were corrected for environmental effects and form the dere-
58 gressed proofs [6]. To complete this data, pedigree information on 4,741 animals
59 was known.

60 The second dataset used is a pig dataset that PIC (a Genus company) has made
61 available [7]. The dataset consisted of 3,534 animals genotyped on the Porci-
62 neSNP60 chip (64,233 markers). These genotypes were filtered for a Minor Allele
63 Frequency (MAF) > 0.001 and a proportion of missing genotypes by SNPs $> 10\%$.
64 Markers on the X and Y chromosomes were excluded, yielding 52,842 SNPs. Pedi-
65 gree information was also available, including parents and grandparents of the geno-
66 typed animals ($n = 6,473$). Genotyped animals had phenotypes for five traits, with
67 heritability ranging from 0.07 to 0.62. The authors state that, "Each phenotype was

68 either corrected for environmental factors (e.g. year of birth or farm) and rescaled
 69 by correcting for the overall mean (traits 3, 4 and 5) or was a rescaled, weighted
 70 mean of corrected progeny phenotypes (traits 1 and 2), for which many animals
 71 have no individual performance data” [7].

72 2.2 Model

The objective was to predict a quantitative trait from a subset of quantitative variables. This can be modeled as a multiple linear regression, which we propose formulating as follows:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j \gamma_j x_{ij} + e_i, \quad i = 1, \dots, n, \quad (1)$$

$$\mathbf{y} = X(\boldsymbol{\beta} \cdot \boldsymbol{\gamma}) + \mathbf{e},$$

73 where

- 74 • \mathbf{y} is a vector of dimension n (number of animals) representing the quantitative
 75 trait of interest,
- 76 • X are the fixed effects with $X = (\mathbf{1}, \mathbf{x}_1, \dots, \mathbf{x}_p)$ a $n \times (p + 1)$ matrix with p
 77 the number of SNPs studied, the first column of X contains a 1
- 78 • $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ the coefficients to be estimated.
- 79 • $\boldsymbol{\gamma}_j = (\gamma_1, \dots, \gamma_p)^t$ equals 1 if the SNP j is in the model, 0 otherwise. The
 80 operator \cdot corresponds to the product of the term-by-term vectors.
- 81 • \mathbf{e} are Gaussian residuals assumed to be independent and identically dis-
 82 tributed (i.i.d.) with zero mean and variance σ_e^2 .

83 Parameters $\boldsymbol{\gamma}$, σ_e^2 , β_0 and $\{\beta_j : \gamma_j = 1, 1 \leq j \leq p\}$ have to be estimated.

84 In this proposed modeling, as in many approaches from the literature, animals are
 85 considered to be independent.

86 However, unlike human studies, familial relationships exist between individuals;
 87 these are described by means of a deep pedigree. This is important information,
 88 which must be taken into account to avoid, for example, considering SNPs as sig-
 89 nificant when they are not, and thereby increasing the number of false positives.

90 We proposed integrating these familial relationships, using the pedigree, through a
 91 linear mixed model based on equation (1). A term $Z\mathbf{u}$ is added to this equation in
 92 order to introduce correlations between observations. This leads to equation (2).

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j \gamma_j x_{ij} + \sum_{k=1}^q z_{ik} u_k + e_i, \quad i = 1, \dots, n, \quad (2)$$

$$\mathbf{y} = X(\boldsymbol{\beta} \cdot \boldsymbol{\gamma}) + Z\mathbf{u} + \mathbf{e},$$

93 where $Z\mathbf{u}$ are the random effects representing familial relationships (animal model).
 94 These effects serve to reduce the number of false positive SNPs detected due to fa-
 95 miliary relationships [8].

96

97 The objective here was to estimate the parameters $\boldsymbol{\beta} = (\beta_0, \dots, \beta_p)$ and
 98 $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)$. As $\boldsymbol{\gamma}$ is a discrete parameter belonging to $\{0, 1\}^p$, determining
 99 the γ_j values is equivalent to determining variables that participate in the regres-
 100 sion model. This problem is a typical feature selection problem, well known in data-
 101 mining, and which may be seen as a combinatorial problem. Hence it can be ad-
 102 dressed using combinatorial optimization methods. In what follows, such a method
 103 is proposed for this task.

104

105 2.3 Validation

106 We compared the proposed approach with two classic regression methods used in
 107 genomics considering SNPs as fixed effects (as in our approach): *elastic net* [9] and
 108 *Lasso* [10]. These methods are shrinkage regression approaches, meaning that they
 109 shrink regression coefficients toward 0, which leads to select variables. The main
 110 difference with our approach is that they cannot take familial relationships into
 111 account.

112 Since the objective was to create a low density chip, we fixed the maximum number
 113 of markers selected by using our approach to 96, i.e. a classic low density chip size.
 114 We also compared our approach with the two previous methods bounded to 96
 115 selected markers.

116

117 On the Qualvigen dataset, 100 young bulls were selected to constitute a vali-
 118 dation set, leading to a training sample made up of 1,007 individuals. In order to
 119 generalize the results obtained, this split was performed 30 times, each time differ-
 120 ently (generating 30 instances). We evaluated the performance of the two proposed

121 models: the first with a multiple linear regression (eq. (1)) and the second with a
122 mixed model integrating familial relationships (eq. (2)).

123 We ran each method on the 30 generated instances. Results are evaluated both
124 in terms of RMSEP (root mean square error of prediction - to minimize) on the
125 validation set and in terms of correlation (to maximize) between the estimated trait
126 and the real trait.

127

128 On the pig dataset, 100 subjects were selected (from the 3,534) to form the val-
129 idation set and this split was performed differently 10 times in order to obtain 10
130 different instances. We studied the performance of the approaches for the five traits.
131 As the results for both methods, elastic net and Lasso were similar, we present only
132 those obtained using Lasso.

133

134 Our approach was developed in C++ using the PARADISEO platform [11]
135 (<http://paradiseo.gforge.inria.fr/>), for metaheuristics setting. For classic ap-
136 proaches (elastic net, Lasso), we used R software with the “*glmnet*” procedure. The
137 λ parameter for both methods was determined using “*cv.glmnet*” and the α param-
138 eter for elastic net by a 3-fold cross-validation. In our approach, the evaluation of a
139 variable selection with a mixed model was performed using the BLUPF90 program
140 in FORTRAN by Mistzal [12].

141 The results presented were computed at the regional cluster financed by Lille 1
142 University, the CPER Nord-Pas-de-Calais/FEDER, France Grille and CNRS.

143 **3 OPTIMIZATION APPROACH**

144 Evolutionary algorithms are search methods based on natural evolution [13] and the
145 most popular one, used in this study, is the genetic algorithm [14]. The objective in
146 this study was to search for a relevant subset of variables (markers) among a large
147 amount of possible subsets.

148 Figure 1 shows the general scheme of the algorithm.

149 It starts with the initialization of a population of n individuals where an individual
150 is an encoding version of a candidate solution (in our study a solution describes
151 a subset of variables). Each solution is evaluated and $n/2$ couples of solutions are
152 selected. Each couple generates two new solutions through the crossover operator

153 and then a mutation operator might be applied in order to diversify these new so-
154 lutions. A replacement strategy chooses, among initial solutions and new solutions,
155 the solutions of the next population. These successive steps correspond to a gen-
156 eration. The algorithm stops when it reaches a given stopping criterion. Different
157 selection, crossover, mutation and replacement operators or stopping criteria may
158 be used. We present the choices we made below.

159 3.1 Encoding of a solution

160 The representation of a solution plays a major role in the implementation of a
161 metaheuristic since it influences the choice of the operators and the evaluation
162 function. We chose to use a binary vector indicating whether a variable is selected (1)
163 or not (0) since it is very close to the statistical models (1) and (2) presented above
164 (this is equivalent to the vector $(\gamma_1, \dots, \gamma_p)$). In addition, this encoding provides a
165 simple but effective design of the neighborhood. Example of a solution:

1	0	0	1	1	0	1	0
---	---	---	---	---	---	---	---

166
167 In this solution, variables 1, 4, 5 and 7 are selected. The size of a solution (8 in this
168 case) is equal to the total number p of variables in the dataset studied.

169 3.2 Objective function

170 The aim of the optimization method is to effectively explore a large search space
171 matching all possible subsets of variables. Therefore, this method uses an evaluation
172 criterion (fitness function) able to associate one quality measure with each solution.
173 In our context, the objective was to identify the best subset of variables, in other
174 words, the one which will provide the best predictive model. A well-known difficulty
175 in data-mining is how to assess the model's ability to predict a trait from data that
176 were not used to develop the model (validation sample). The objective function
177 used, depending on the model considered (a multiple linear regression model or a
178 mixed model) will be described later.

179 3.3 Initialization

180 The classic initialization of solutions of an evolutionary algorithm is to set solutions
181 randomly (in a uniform manner). As the representation of a solution is a binary
182 vector, the purpose here is to set up each bit to 0 or 1. To obtain diversified ini-
183 tial solutions, we wish to have solutions of different sizes (with different numbers

184 of selected variables), while remaining below the maximum allowable number of
185 variables (96 variables here). Therefore, for each solution, its number k of selected
186 variables is uniformly chosen in a predefined interval $[min., max.]$. Moreover, to
187 accelerate convergence, and in order to obtain interesting initial solutions, we com-
188 pared pure random initialization with guided initializations. Three configurations
189 have been tested.

- 190 • The first consists in uniformly choosing the k variables of each solution of the
191 initial population.
- 192 • The second configuration consists in initializing all the solutions of the initial
193 population using the variables selected by the Lasso method. Indeed, the Lasso
194 method (not limited in number of selected variables) allows us to obtain a
195 subset of *a priori* interesting variables. So, for each solution (individual) of
196 the initial population, we uniformly selected variables among those obtained
197 by the Lasso method. If the number k of variables desired for the solution is
198 greater than the number of variables extracted by the Lasso method, we choose
199 all variables identified by the Lasso method, and add variables uniformly
200 selected from among the others.
- 201 • The third configuration consists in combining the two described above. To
202 do this, we separate the initial population into two parts. For the first half,
203 solutions of the initial population are randomly generated, while solutions for
204 the second half are constructed using variables selected by Lasso.

205 Experimentations on simulated data [15] show that initializations based on the asso
206 method (configurations 2 and 3 presented above) give better results than a pure
207 uniform initialization. As there was no significant difference between the two con-
208 figurations based on Lasso method, we chose to use the third configuration, that is
209 an initialization based on Lasso for 50% of the solutions of the initial population,
210 in order to maintain diversity.

211

212 3.4 Selection

213 The selection process of an evolutionary algorithm aims to determine the individuals
214 that will breed and how many children each couple will generate. This is equivalent
215 to determining the subset of variables which will be used for the creation of new

216 subsets. Several selection strategies are possible including roulette wheel selection,
217 stochastic universal sampling or tournament selection [3]. We chose tournament
218 selection as it does not converge too fast and also helps to maintain diversity. Tour-
219 nament selection consists in randomly selecting m individuals, m being the size
220 of the group tournament. The best individual among the m individuals will be
221 retained. The selection of n individuals requires n executions of a tournament.

222

223 3.5 Reproduction

224 Once the parents are selected, the reproduction phase applies variation operators
225 such as crossover and mutation to generate children. The choice of binary encoding
226 of solutions would allow us to use classic crossover and mutation operators [16].
227 Nevertheless, the choice of **crossover** operator may depend on the problem stud-
228 ied, in order to ensure an efficient one. Indeed, in the context of feature selection,
229 traditional operators such as 1-point or 2-point crossovers may have a negative ef-
230 fect since they may “break” some interesting blocks. Therefore, we chose to use
231 a crossover operator adapted to the problem of feature selection, the Subset Size-
232 Oriented Common Feature (SSOCF [17]). The principle is described in Figure 2.

233 Variables in common to both parents are kept by the children. The others are in-
234 herited from the parents with the probability $(n_i - n_c)/n_u$ where n_i is the number
235 of variables selected by the i^{th} parent, n_c is the number of variables selected jointly
236 by both parents and n_u the number of variables unshared by the parents (variables
237 selected by one of the parents, but not both). The objective of this method is, on
238 the one hand, to keep the blocks of useful information and on the other hand, to
239 keep for the children the variables shared by their parents.

240 The **mutation** is a unary operator (one input solution) applied to an individual
241 to change it slightly. In a binary representation of solutions, the mutation typically
242 used is a bit-flip. Two types of mutation were used in our algorithm based on the
243 number of selected variables in the current solution:

- 244 • flip a (small) percentage of bits uniformly determined among all variables
245 when the number of variables in the selected current solution is less than the
246 maximum desired number of variables (\Rightarrow addition or deletion of variables).

- 247 • flip a (small) percentage of bits uniformly determined among selected vari-
248 ables (bit = 1) when the maximum number of desired variables is reached
249 (\Rightarrow deletion of variables).

250

251 At the reproduction step, crossover and mutation are not applied consistently.
252 Indeed, the crossover rate is used to define the probability that two selected parents
253 are crossed to generate children. Similarly, the mutation rate is the probability of
254 applying a mutation to a solution. We compared the performances of the algorithm
255 on simulated data [15] using low (0.2) and high (0.8) crossover and mutation rates.
256 We finally chose to keep a crossover rate of 0.8 and a mutation rate of 0.8.

257

258 3.6 Replacement

259 The population size must be constant over generations. Hence, when children are
260 generated, all parents and children cannot be kept. The replacement procedure,
261 the last step of a generation, will help to define the survivors among parents and
262 children generated. The replacement procedure that we chose here was to keep a
263 child only if it is better than the worst of the remaining parents. When a child is
264 preserved, the worst parent is deleted. The worst parents are replaced progressively
265 by the best children.

266 3.7 Stopping criteria

267 The evolutionary algorithm is an iterative approach for which it is necessary to set
268 a stopping criterion. Here, we set a maximal number of generations, determined
269 empirically depending on the evolution curve of the best solution of the population.

270

271 3.8 Diversification

272 During the evolution of the evolutionary algorithm, a failure that can be observed
273 is the stagnation of the search. To avoid this, diversification methods are proposed
274 such as the stochastic diversity of migration or “Random Immigrant” [18]. The idea
275 is to replace a portion of the population by individuals generated uniformly when
276 the best individual of the population has not been improved for a given number
277 of generations. In our algorithm, when the best individual of the population does

278 not change for a fixed number of generations, all individuals whose fitness is lower
279 than the average fitness of the population are replaced by new individuals uniformly
280 generated.

281

282 3.9 Parallelization

283 During the evolutionary algorithm, for a generation, several solutions (children gen-
284 erated) have to be evaluated. The evaluation may take time as a regression (com-
285 putation of the coefficients of each marker) has to be performed. Thus, to reduce
286 execution time, we proposed making these evaluations in parallel. We therefore im-
287 plemented a synchronous parallel version of the algorithm with the SMP module
288 of PARADISEO [11]. The aim is to parallelize, at every generation, the evaluations
289 of children (solutions) of the evolutionary algorithm using the scheme “master /
290 slave”. Once all children are generated, their evaluations are independent so they
291 are performed in parallel. During the evaluation phase, the master sends one solu-
292 tion to evaluate per slave and they send back the fitness of the solution received.

293 4 A STATISTICAL FITNESS FUNCTION

294 As we saw in Section 3.2, such an optimization method is based on a fitness function
295 which evaluates the quality of solutions. The quality of a solution (a subset of
296 variables) was assessed according to the quality of the underlying model (i.e. how
297 best it fit the data). We defined a fitness function for each of the models proposed
298 previously: multiple linear regression (1) and mixed model (2).

299 4.1 Multiple linear regression

300 Through multiple linear regression, a range of model selection methods is available
301 in the literature (e.g. [19]). The most commonly used criteria are the AIC criterion
302 (Akaike Information Criterion) [20], the BIC criterion (Bayesian Information Crite-
303 rion) [21] and cross-validation. Unlike the AIC criterion, the BIC criterion tends to
304 penalize complex models more heavily and therefore seems more appropriate to our
305 objective of variable selection in high dimension. In a previous study, we compared
306 three criteria [15]: BIC and two types of cross-validation (k-fold and leave-one-out)
307 on simulated data and the BIC criterion gave the best results. We used it in this
308 study.

309 4.2 Mixed model

310 For our second model, the quality of a solution was evaluated with a 3-fold cross-
311 validation. Indeed, calculating the BIC requires calculating the likelihood of the
312 model. However, the method of maximum likelihood is not suitable for mixed mod-
313 els and the use of restricted maximum likelihood (REML) is recommended for this
314 type of model. An adaptation of the BIC has been proposed by [22] under repeated
315 data but this is not the case of our data, so we chose to use 3-fold cross-validation.

316

317 5 RESULTS

318 In order to analyze performance of the proposed methods, we compared them on
319 the two presented datasets using elastic net and Lasso approaches without and with
320 a restriction on the number of selected markers. Figure 3 illustrates our results on
321 the cattle data (Qualvigène project).

322 Our approach based on multiple linear regression (LM) allowed us to obtain re-
323 sults comparable to classic approaches bounded to 96 selected markers. Adding
324 familial relationships using mixed models improved the results of our method from
325 a correlation of 0.48 with LM to a correlation of 0.56 with MM. Moreover, this new
326 approach outperformed classic approaches (the Student's test on the mean predic-
327 tion error concluded with a significant difference between MM and EN96 or MM
328 and L96) and became comparable to unlimited approaches (mean RMSEP equal to
329 0.49 for MM against 0.41 for the Lasso method (Las), for example).

330

331 Figures 4 to 8 illustrate results (RMSEP and correlation) for the five traits on
332 the pig dataset. We observed that whatever the trait, our approach based on a mul-
333 tiple linear regression performed slightly better than Lasso limited to 96 selected
334 markers. Moreover, performance is improved with our second approach including
335 familial relationships so as to outperform the classic approach (the Student's test
336 concluded with a significant difference between L96 and MM for all traits). For
337 example, on the trait T1, the mean prediction error of the Lasso method limited to
338 96 variables selected was equal to 0.55; it decreased to 0.51 for our first approach
339 and to 0.43 for our MM approach. On this trait, the prediction error of our last

340 approach outperformed that of the Lasso method (0.46).

341

342 In order to evaluate the influence of heritability on the performances of the differ-
343 ent methods, results were compared on pig data on 5 traits with different heritabil-
344 ities: 0.07, 0.16, 0.38, 0.58 and 0.62. For traits T2, T3 and T4, taking into account
345 familial relationships improved the results (significant Student's test). However,
346 although whatever the trait MM is always better than LM, sometimes the dif-
347 ference is small. For trait T1, the difference observed between LM and MM was
348 significant in terms of RMSEP ($p - value = 0.03$) but not in terms of correlation
349 ($p - value = 0.07$). For the trait T5, the difference between LM and MM was not
350 significant.

351

352 The execution times of our approaches were slightly higher than those of classic
353 approaches especially because they required the execution of the Lasso to be initial-
354 ized. Table 1 shows the execution times for the different methods on the Qualvigène
355 dataset.

356 We observed that the evaluation using a mixed model takes much longer than the
357 multiple linear regression due to the high number of parameters to be estimated.
358 Indeed, the actual execution time of the algorithm (once the initialization time was
359 removed) with the mixed model was 7 minutes compared with 10 seconds for the
360 linear regression. However, the execution times of our approaches were reasonable
361 compared with the time taken to collect and pre-process data.

362 6 DISCUSSION

363 On the real datasets used, our approaches lead to similar or even better results
364 than classic approaches. This enabled us to validate the relevance of combining a
365 combinatorial optimization method and a regression to solve our problem.

366 We observed that methods unlimited in the number of selected markers (EN and
367 Las) obtained the best results (with a correlation of around 0.6). However, they
368 selected too many variables (≈ 580 and 300 respectively) and were not suitable for
369 our problem. Indeed, selecting a large amount of variables results in a more accu-
370 rate model, but our objective was to select a limited number of variables. The first
371 model proposed in this study, based on a multiple linear regression (LM), assumes

372 that animals are independent. This is not the case in our data so we proposed the
373 second model including familial relationships using a mixed model (MM). As the
374 assumptions of LM are not met in our data but those of MM are, we expected
375 to have better results with MM than with LM. This was confirmed by the results
376 obtained on real cattle and pig data, which showed the importance of including
377 familial relationships for these datasets. Regarding the results obtained on the pig
378 dataset, as we have 5 traits with low to high heritability, we can measure the impact
379 of the heritability of the trait on the performances of the different approaches. First,
380 if we look at the results in terms of correlation (which are comparable from one
381 trait to another), the methods performs better on low heritability traits. Moreover,
382 if we compare LM and MM the difference in terms of correlation is not significant
383 for the less heritable (T1) and the most heritable (T5) traits but significant for the
384 others. It seems interesting, therefore, for this type of data, to integrate familial
385 relationships for trait with moderate heritabilities but not necessarily for very low
386 or high heritability.

387

388 In an evolutionary algorithm, it is difficult to fine-tune parameters. For each oper-
389 ator, we tested several possibilities (the most popular regarding this kind of data),
390 evaluated their performance on simulated data and chose the best. Our approach is
391 flexible regarding the statistical model used to evaluate subsets of variables. Indeed,
392 we first performed a multiple linear regression and next a mixed model. However,
393 this can be easily changed, for example by combining a multiple linear regression
394 to start the search and a mixed model to refine the search. It could be also possible
395 to test other approaches, such as Bayesian models.

396

397 In order to evaluate the quality of our approach, we extracted 100 bulls from our
398 original datasets. We decided to choose these 100 bulls from among the young ones
399 given our end objective of predicting performance on young bulls. Another outcome
400 could be to extract a family if the objective was to predict the trait under study
401 for an animal unrelated to those in the study.

402

403 Results were presented in terms of correlation and Root Mean Square Error of
404 Prediction (RMSEP). Indeed, although the majority of genomic selection studies

405 present the results in terms of correlation, this measure is less accurate than the
406 RMSEP. The RMSEP evaluates the difference between each prediction and real val-
407 ues whereas the correlation only looks at their distribution and evaluates whether
408 they go the same way. So, if for all subjects the trait is estimated with a lower value
409 than the real one, the correlation will be good whereas the RMSEP will be bad. If
410 the objective is only to select the best animals, the correlation is a good indicator
411 but if it is also interesting to have a good estimation for the trait, RMSEP is more
412 accurate.

413

414 In genotyping data, some markers are in high Linkage Disequilibrium (LD). In
415 our approach, if a marker is in high LD with another one already in the model, as
416 it is not adding more information to the actual regression model, it is likely that it
417 will not be selected. In the final model, the LD between markers is low so that they
418 explain different parts of the trait.

419

420 Our objective was to find a predictive model based on a small number of markers
421 (96). This kind of very low density chip could be a decision tool for breeders in order
422 to select animals to be genotyped on a 54K chip for example or in a 6K chip with
423 imputation. Results on cattle and pig datasets showed that our approach obtains
424 better results than elastic net or Lasso method in a reasonable computational time.
425 Some may argue that a very low density chip (96 SNPs), specific to a given trait,
426 became less interesting once imputation and “low” density (6K) chips were used.
427 This may be true for cattle. But for livestock such as pigs or poultry, 96 SNP chips
428 still appear to be interesting tools.

429 **7 CONCLUSION**

430 The objective of this study was to select a subset of relevant markers to predict
431 a quantitative trait. We proposed a novel approach based on an evolutionary al-
432 gorithm combined with a statistical model. We compared two statistical models,
433 the first without familial relationships and the second integrating them using the
434 pedigree information.

435 We first showed the importance of including of familial relationships in the statistical
436 model as prediction on a validation set was better on the real datasets tested. Due

437 to its powerful exploration of the search space, the optimization approach makes
438 it possible to find the SNPs of interest. Our approaches performed as effectively as
439 the most efficient approaches used in the field and sometimes outperformed them.

440 **Competing interests**

441 The authors declare that they have no competing interests.

442 **Authors' contributions**

443 JH performed the analysis and drafted the manuscript. JH, CD, JJ and GE designed the study and developed the
444 method. GE and RD prepared phenotypic and genotypic data. CD, JJ, GE and RD revised the manuscript. All
445 authors read and approved the final manuscript.

446 **Acknowledgements**

447 We highly appreciate and thank the technical staff at the CRI-Lille 1 center for their strong and helpful support. We
448 thank the ANR AGENAE and APIS-GENES for allowing us to use the Qualvigène data.

449 **Author details**

450 ¹Inria Lille - Nord Europe, Lille, France. ²Gènes Diffusion, 3595 rte de Tournai, Douai, France. ³Laboratoire Paul
451 Painlevé / CNRS, Université Lille 1, Lille, France. ⁴LIFL, Université Lille 1, Lille, France.

452 **References**

- 453 1. Hayes, B.J., Pryce, J., Chamberlain, A.J., Bowman, P.J., Goddard, M.E.: Genetic architecture of complex traits
454 and accuracy of genomic prediction: Coat colour, milk-fat percentage, and type in holstein cattle as contrasting
455 model traits. *PLoS Genet* **6**(9), 1001139 (2010)
- 456 2. Corne, D., Dhaenens, C., Jourdan, L.: Synergies between operations research and data mining: The emerging
457 use of multi-objective approaches. *European Journal of Operational Research* **221**(3), 469–479 (2012)
- 458 3. Talbi, E.-G.: *Metaheuristics*. John Wiley & Sons, Inc., Hoboken, NJ, USA (2009)
- 459 4. Darwin, C.: *On the Origin of the Species by Means of Natural Selection: Or, The Preservation of Favoured
460 Races in the Struggle for Life*. John Murray, London (1859)
- 461 5. Allais, S.: *Détection et validation de marqueurs génétiques impliqués dans la qualité de la viande bovine*. PhD
462 thesis, AgroParisTech (January 2011)
- 463 6. VanRaden, P.M., Wiggans, G.R.: Derivation, calculation, and use of national animal model information. *Journal
464 of dairy science* **74**(8), 2737–2746 (1991). PMID: 1918547
- 465 7. Cleveland, M.A., Hickey, J.M., Forni, S.: A common dataset for genomic analysis of livestock populations. *G3:
466 Genes|Genomes|Genetics* **2**(4), 429–435 (2012)
- 467 8. Habier, D., Fernando, R.L., Dekkers, J.C.M.: The impact of genetic relationship information on
468 genome-assisted breeding values. *Genetics* **177**(4), 2389–2397 (2007)
- 469 9. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67** Part 2,
470 301–320 (2005)
- 471 10. Tibshirani, R.: Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series
472 B* **58**, 267–288 (1994)
- 473 11. Cahon, S., Melab, N., Talbi, E.-G.: ParadisEO: a framework for the reusable design of parallel and distributed
474 metaheuristics. *Journal of Heuristics* **10**(3), 357–380 (2004)
- 475 12. Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D.H.: BLUPF90 and related programs
476 (BGF90). In: *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production: August
477 2002, Montpellier*, pp. 1–2 (2002)
- 478 13. Pal, S.K., Bandyopadhyay, S., Ray, S.S.: Evolutionary computation in bioinformatics: a review. *IEEE
479 Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews* **36**(5), 601–615 (2006)
- 480 14. Holland, J.H.: *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to
481 Biology, Control, and Artificial Intelligence* vol. viii. U Michigan Press, Oxford, England (1975)

- 482 15. Hamon, J.: Optimisation combinatoire pour la sélection de variables en régression en grande dimension :
483 Application en génétique animale. PhD thesis, Université des Sciences et Technologie de Lille - Lille I
484 (November 2013)
- 485 16. Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning, (1989)
- 486 17. Emmanouilidis, C., Hunter, A., MacIntyre, J.: A multiobjective evolutionary setting for feature selection and a
487 commonality-based crossover operator. In: Proceedings of Congress on Evolutionary Computation, pp. 309–316
488 (2000)
- 489 18. Grefenstette, J.: Genetic algorithms for changing environments. In: Parallel Problem Solving from Nature 2:
490 Amsterdam, pp. 137–144 (1992)
- 491 19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning - Data Mining, Inference, and
492 Prediction, Second Edition, (2009)
- 493 20. Akaike, H.: A new look at the statistical model identification. IEEE Transactions on Automatic Control **19**(6),
494 716–723 (1974)
- 495 21. Schwarz, G.: Estimating the dimension of a model. The Annals of Statistics **6**(2), 461–464 (1978)
- 496 22. Delattre, M., Lavielle, M., Poursat, M.-A.: BIC selection procedures in mixed effects models (2012)

497 **Figures**

Figure 1 Evolutionary algorithm. General scheme of an evolutionary algorithm.

Figure 2 SSOCF. Subset Size-Oriented Common Feature : a crossover operator.

Figure 3 Performances on Qualvigène dataset. Boxplot evaluating performances of classical approaches (elastic-net (EN), lasso (Las)), classical approaches limited to 96 SNPs selected (EN96, L96) and our two approaches (based on multiple linear regression (LM) and on mixed model (MM)), in term of RMSEP (to minimize) on the left and of correlation (to maximize) on the right.

Figure 4 Performances on PIC, trait T1 ($h^2 = 0.07$).

Figure 5 Performances on PIC, trait T2 ($h^2 = 0.16$).

Figure 6 Performances on PIC, trait T3 ($h^2 = 0.38$).

Figure 7 Performances on PIC, trait T4 ($h^2 = 0.58$).

498 **Tables**

Figure 8 Performances on PIC, trait T5 ($h^2 = 0.62$).

Table 1 Execution time of different methods on cattle data

EN	Lasso	EN96	L96	LM	MM
35 min.	3 min.	16 min.	1 min.	3 min. 10	10 min.