Revista de Informática Teórica e Aplicada - RITA - ISSN 2175-2745
Vol. 25, Num. 2 (2018) 85-92

RESEARCH ARTICLE

A Genetic Programming Model for Association Studies to Detect Epistasis in Low Heritability Data

Um Modelo de Programação Genética para Estudos de Associação para Detectção de Epistasia em Dados de Baixa Herdabilidade

Igor Magalhães Ribeiro[1],Carlos Cristiano Hasenclever Borges[2], Bruno Zonovelli da Silva[1], Wagner Arbex[3]*

Abstract: The genome-wide associations studies (GWAS) aims to identify the most influential markers in relation to the phenotype values. One of the substantial challenges is to find a non-linear mapping between genotype and phenotype, also known as epistasis, that usually becomes the process of searching and identifying functional SNPs more complex. Some diseases such as cervical cancer, leukemia and type 2 diabetes have low heritability. The heritability of the sample is directly related to the explanation defined by the genotype, so the lower the heritability the greater the influence of the environmental factors and the less the genotypic explanation. In this work, an algorithm capable of identifying epistatic associations at different levels of heritability is proposed. The developing model is a aplication of genetic programming with a specialized initialization for the initial population consisting of a random forest strategy. The initialization process aims to rank the most important SNPs increasing the probability of their insertion in the initial population of the genetic programming model. The expected behavior of the presented model for the obtainment of the causal markers intends to be robust in relation to the heritability level. The simulated experiments are case-control type with heritability level of 0.4, 0.3, 0.2 and 0.1 considering scenarios with 100 and 1000 markers. Our approach was compared with the GPAS software and a genetic programming algorithm without the initialization step. The results show that the use of an efficient population initialization method based on ranking strategy is very promising compared to other models.

Keywords: Bioinformatics — GWAS — SNP — Genetic Programming — Random Forest — Computational Modeling — Mathematical Modeling

Resumo: Os estudos de associação genômica ampla (genome-wide associations studies - GWAS) visam identificar os marcadores mais influentes em relação aos valores fenotípicos. Um dos desafios substanciais é encontrar um mapeamento não linear entre genótipo e fenótipo, também conhecido como epistasia, que geralmente se torna o processo de busca e identificação de SNPs funcionais mais complexos. Algumas doenças como o câncer do colo do útero, leucemia e diabetes tipo 2 têm baixa herdabilidade. A herdabilidade da amostra está diretamente relacionada à explicação definida pelo genótipo, portanto, quanto menor a herdabilidade, maior a influência dos fatores ambientais e menor a explicação genotípica. Neste trabalho, é proposto um algoritmo capaz de identificar associações epistáticas em diferentes níveis de herdabilidade. O modelo em desenvolvimento é uma aplicação de programação genética com uma inicialização especializada para a população inicial, consistindo de uma estratégia de florestal aleatória. O processo de inicialização visa classificar os SNPs mais importantes aumentando a probabilidade de sua inserção na população inicial do modelo de programação genética. O comportamento esperado do modelo apresentado para a obtenção dos marcadores causais pretende ser robusto em relação ao nível de herdabilidade. Os experimentos simulados são do tipo caso-controle, com nível de herdabilidade de 0,4, 0,3, 0,2 e 0,1, considerando cenários com marcadores de 100 e 1000. Nossa abordagem foi comparada com o software GPAS e um algoritmo de programação genética sem a etapa de inicialização. Os resultados mostram que o uso de um método eficiente de inicialização da população baseado na estratégia de ranking é muito promissor em comparação com outros modelos. .

Palavras-Chave: Bionformática — GWAS — SNP — Programação Genética — Floresta Aleatória — Modelagem Computacional — Modelagem Matemática

[1] *Postgraduate Program in Computational Modeling, Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil*
[2] *Department of Computer Science, Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil*
[3] *Federal University of Juiz de Fora and Brazilian Agricultural Research Corporation, Juiz de Fora, MG, Brazil*
*__Corresponding author__: wagner.arbex@{ufjf.edu.br,embrapa.br}
DOI: http://dx.doi.org/10.22456/2175-2745.79333 • **Received:** 02/01/2018 • **Accepted:** 20/04/2018

1. Introduction

Over the last decade, the studies about the human genome has generated a large amount of information, largely due to the emergence of density-based "chips" technology that has facilitated the measurement of hundreds of thousands of variations of DNA sequences throughout the human genome [1] [2]. The most common form of genomic variation or marker is known as single nucleotide polymorphisms (SNPs). These variations correspond to the alternation (substitution, deletion or insertion) of nucleotides A, T, C and G in a single position of the genome.

The genome-wide association study (GWAS) allows the finding out molecular markers that indicate the risk or predispose to complex diseases. The identification of these markers can help directly or indirectly understand the mechanisms of a particular disease. Directly finding the marker and indirectly indicating the gene, metabolic pathway among other biological characteristics, and the biggest challenge is to interpret and understand the large number of information obtained in the genotyping process, when molecular markers are identified [3][2].

In [4], the authors indicates that one of the types of gene interaction that gives rise to complex diseases is called epistasis. This type of interaction makes the mapping between genotype and non-linear phenotype, that is, one marker can mask or completely alter the behavior of the other generating a completely new characteristic. In this way, the interaction becomes more difficult to detect. Heritability can be estimated by the ratio between the variances of the genotype and phenotype. This ratio measures the proportionality of how much the genetic factor influences the phenotype [5]. The heritability directly interferes with the ability to correctly select markers of interest for the study. The lower the heritability, the less the explanation obtained through the genotype, and the greater the influence of environmental factors.

Several medical conditions or diseases have low heritability, for instance: asthma (0.3) [6], bladder cancer (0.07-0.31) [7], cervical cancer (0.22) [8], leukemia (0.01) [9]; type-2 diabetes (0.26) [10], and so on. Therefore, it is necessary to develop algorithms capable of identifying risk factors at different levels of heritability.

In addition to heritability, there is a complexity in the genotype-phenotype relationship due to distinct gene actions. The works [1] and [11] explain that the linear modeling (linear regression) used in GWAS problems considers only one SNP at a time, in this context, the gene-gene and environment-gene interactions of each marker are ignored. For the identification of more complex genetic actions such as epistasis and dominance, machine learning models that consider multiple markers in classification and regression problems have been presented to identify non-linear interactions between SNPs.

Initialization approaches were used in related works [1, 12, 13], and according [14] the use of expert knowledge can significantly improve the performance in detection SNP-SNP interactions in genetic programming algorithms. These proposed models used a feature selection algorithm called Relief [15] and their variants. The idea of Relief is estimate the feature weight according to their ability to discriminate between individuals and their neighbors . However these algorithms can identify possible SNP interactions, they are susceptible to noise. They may capture marginal effects (single SNP interaction with phenotype) rather than epistatic interactions.

The objective of this work is to develop a model to identify non-linear interaction of functional SNPs, i.e., epistasis, across different levels of heritability. The model proposed is an algorithm of evolution of solutions based on genetic programming (GP) with initialization through random forest. The idea behind the random forest choice is to use a most informative and robust measure in this context. The increase in mean square error (MSE) of predictions can rank SNPs with low noise and more informative to be part of epistatic interactions than Gini index for example, taking between 5%-25% of extra computing time.

2. Proposed Model

We propose a model combining evolutionary computation and machine learning techniques, more precisely, genetic programming to analysis to genotype/phenotype and random forest as expert knowledge guiding the search for SNPs interactions that could lead to a complex disease risk. The expert knowledge algorithm is used to measure of attribute quality and allows SNPs of interest to be inserted into the initial population of GP algorithm.

2.1 Individuals

The structure of individuals is based on a tree – or a tree representation of solutions – was suggested by [16] , where the authors proposed to use multi-valued logic expressions in disjunctive normal form (DNF). A DNF logic expression is disjunctive of one or more monomials, where one monomial consists of a single or a set of literals. In Figure 1, an example of generic tree with DNF logic expression representing a GP individual is shown. The GP grammar adopted is simple and the function set is given by "AND" and "OR" expressions. The terminal set consists of SNPs and their respective alleles, for instance "SNP1 = 0".

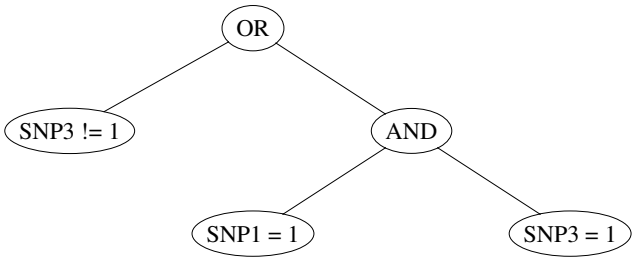

Figure 1. Example of an individual used in GPi. The individuals are expression trees that represents SNP-SNP high order interactions.

2.2 Fitness function

The evaluation of individuals is given by the fitness function f_i, shown in (1):

$$f_i = \frac{T}{VP+VN} + \frac{N_i}{\alpha} \quad (1)$$

where i is an index for an individual, T represents the total of case-control individuals. VP is the true-positives, VN is the true-negatives correctly classified. N_i is the numbers of nodes and α is a parsimony constant (introduced in [17]).

2.3 Operations to generate new individuals

To create a new generation, the genetic operators including crossover and mutation are applied. An overview on the crossover operator is given in Figure 2, where two individuals are selected to crossing. A random node are selected in each individual, then two offspring are created from their combination. The first offspring is a combination of the individual for the cut-off point with the second individual after the cut-off, respectively. The second individual is generated from the inverse composition of the first offspring. In the end, these two individuals are inserted into the new generation.

2.4 Expert knowledge to generate initial population

The initialization mechanism that generate the initial population is based on the importance of the variables according to the random forest algorithm. It is used to capture an isolated effect or a possible genotype-phenotype attribute interaction, generating a ranking of SNPs that predict the phenotype. The measure adopted is most informative than Gini index in this context. The measure represents the increase in MSE of predictions from a sample estimated with an out-of-bag cross validation method.

Usually in GWAS, the parameters are optimized, so, the number of variables to choose from the decision tree nodes and the number of trees that make up the forest need to be defined. The values were defined from empirical tests that presented significant or satisfactory results for the problem in question.

Thereby, the number of variables used in each training subset was the same number of markers used in the simulation and the amount of trees defined by the forest was 1500 for the experiments with 100 markers and 3000 for the experiments with 1000 markers.

To generate the initial population, each terminal node of each individual is submitted to a tournament process in which a marker is selected from among the markers present in the population at random. A comparison of the value assigned to each marker by the random forest algorithm is performed, the one with the highest value is selected to generate the terminal node. Each individual can only have one copy of marker, so if a terminal is populated by a given SNP, it can no longer appear in the solution tree and another tournament is performed until a previously uninserted SNP is found.

2.5 Parameter setting

Table 1 shows the parameter setting used by the algorithms that are part of the model proposed in this work. The parameters such as Population size, Generations, Crossover and Mutation frequency was based on [12] and the functions and terminal sets on [16]. The GP algorithm proposed here has been implement in ECJ [18], and R [19] [20].

Table 1. Parameter setting

Item	Parameter
Population size	4096
Generations	50
Crossover	Single-point
Mutation frequency	0.05
Selection	Tournament

3. Experiments and Results

The following experiments and analyses are conducted on Intel®Core™i7-4770K CPU with 3.50GHz × 8 and 32 GB of RAM. A simulation study was performed to evaluate our model in a GWAS problem. The objective of this simulation is to generate artificial databases capable of capturing the epistatic effects that give rise to phenotypes in cases of low heritability commonly found in genetics. Using GAMETES [21], we could selected heritabilities ranges and created penetrance functions that defines a relationship between the genotype and phenotype. Table 2 exemplifies a penetrance function used to generate a template with epistasis.

Table 2. Example of a penetrance function for a model presenting epistasis.

	AA (0.25)	Aa (0.50)	aa (0.25)
BB (0.25)	0.451	0.214	0.190
Bb (0.50)	0.192	0.164	0.065
bb (0.25)	0.139	0.350	0.463

To the experiments, we developed four penetrance functions where each model has two functional SNPs represents an epistatic interaction and a heritability range between 0.1, 0.2, 0.3, 0.4 respectively. In all scenarios, the minor allele frequency (MAF) was 0.2 and the SNPs represents three alleles (0, 1 or 2).

For each database, functional SNPs were added to other randomly generated markers, defining bases of 100 and 1000 attributes. Each algorithm was run 30 times and counted the number of times that the functional SNPs were correctly selected as the best model of the genetic programming algorithm – we call ”power” the percentage of each algorithm identifies the functional SNPs.

This value represents the predictive power estimation of the proposed method for the phenotype, that is, which frequent

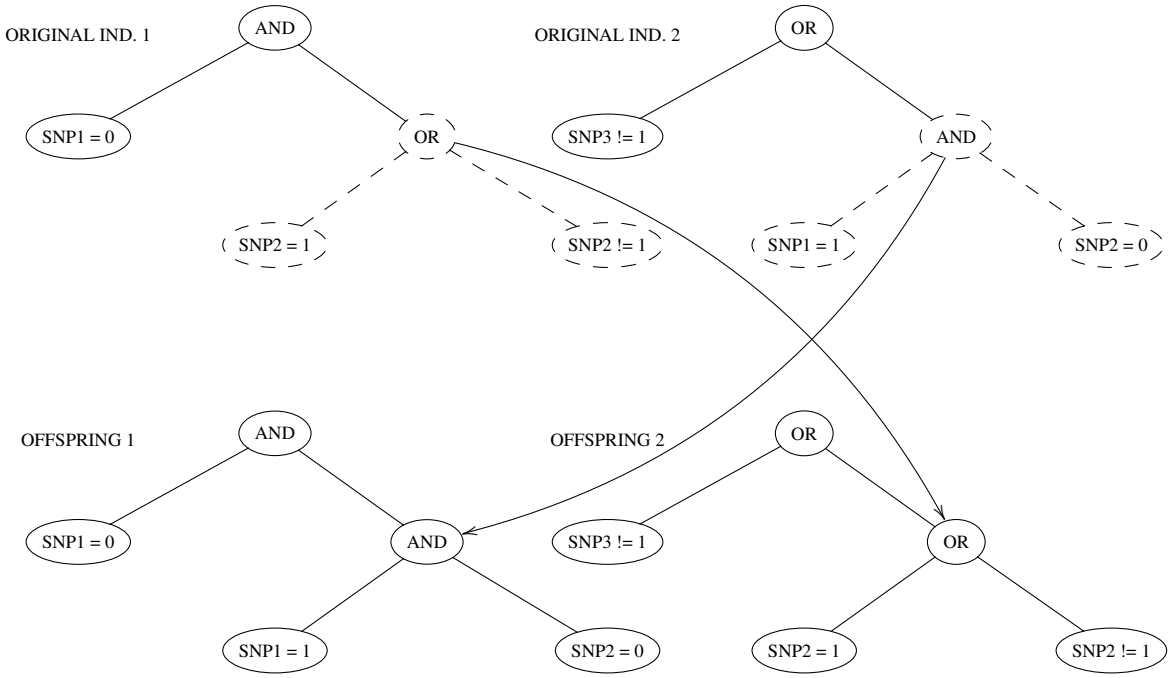

Figure 2. Examples for the crossover used in GPi

the method is able to find the expected solution. The parameter settings used in the simulations were based on [12].

We compared the power GPi algorithm, developed within the scope of this work, against GPi algorithm without initialization step – referred to as GP – and GPAS [16] on estimation of power. We consider the output of each run of GPAS as correct if the best 5 individuals contain the two functional SNPs. This evaluation criterion was used in [12].

As written previously, the simulation data was generated by GAMETES, with all the parameter settings are shown in Table 3:

Table 3. Parameters of the GAMETES simulator

Item	Parameter
MAF	0.2
Population size	2000
SNPs	100 and 1000
Heritability	0.1,0.2,0.3,0.4

For each experiment, a different penetrance function was automatically generated. For example, a possible solution is given by the syntactic tree in Figure 3 for 100 markers and heritability equal to 0.4. The solution tree is generated from the penetrance function given by Table 4.

The results obtained for 100 and 1000 SNPs can be seen respectively in Figures 4 and 5. For the datasets with 100 SNPs, we can observe that GPi – actually, the proposed model – found the correct rules for all heritabilities – even when the heritability dropped to 0.1. The GP algorithm, i.e., without the initialization step, presented satisfactory results. The results achieved by the power of GPAS showed that the algorithm is still satisfactory, presenting a variation in the results only for the case of heritability is equal to 0.1.

Table 4. Penetrance function simulating epistasis effect (database with 100 markers, heritability = 0.4).

	AA (0.25)	Aa (0.50)	aa (0.25)
BB (0.25)	0.535	0.991	0.930
Bb (0.50)	0.998	0.140	0.315
bb (0.25)	0.871	0.433	0.003

Experiments with 100 SNPs indicate that regardless of the methods, functional SNPs can be found. The proposed model obtained a small advantage than the other methods. However, in the databases with 1000 SNPs , the results obtained by each algorithm differ greatly between them. In this scenario, the complexity in finding the functional SNPs has increased. The proposed method obtained significant results even when heritability drops to 0.1. Figure 6 shows the ranking of the SNPs performed by the random forest algorithm. We can note the functional SNPs appear at the top of the all ranking list.

4. Discussion

The identification of SNPs involved directly or indirectly in the gene interactions in scenarios that present low heritability is a fundamental step for the understanding of several complex diseases. The discovery of the biological mechanisms involved in the process can help research directed to the de-

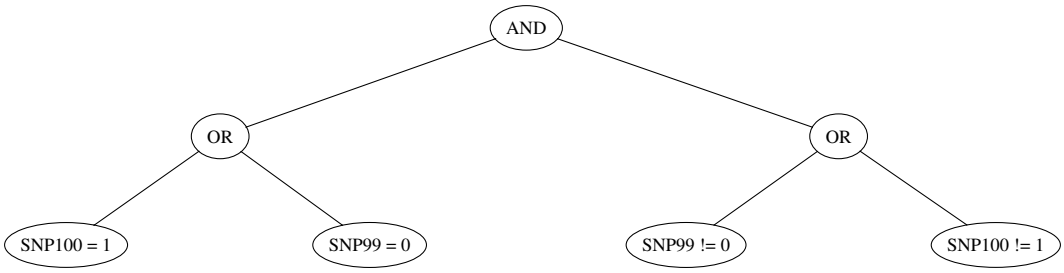

Figure 3. Individual representing a solutions for Table 4.

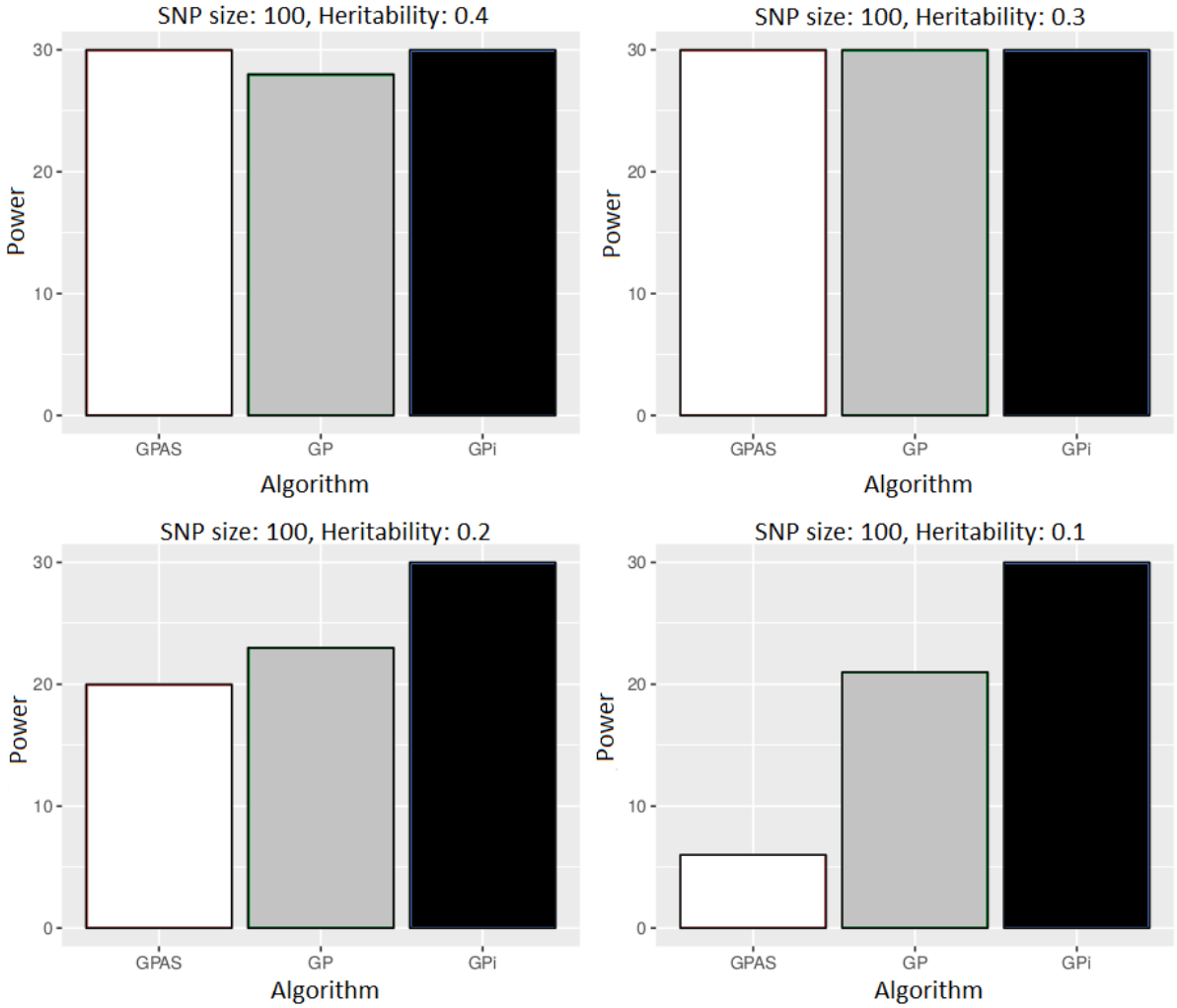

Figure 4. Graphic representing the power of each algorithm (GPAS, GP, GPi) across heritability of 0.4,0.3,0.2,0.1 with a dataset containing 100 SNPs. The power is the number of times that the algorithm identifies the correct two functional SNPs.

velopment of prevention and cure methods.

Non-random initialization methods of the initial population in evolutionary algorithms have been shown to be a strategy to aid in the search for causal SNPs in these scenarios, producing more significant results than algorithms that do not use this strategy. However, we can observe that in cases where heritability is considerable ($\geq 40\%$), initialization strategies may not be the best choice, since the other methods present significant results in this context and do not depend on this step which can be computationally expensive. To provide more conclusive basis for these analyzes, in the future, real datasets could be used, such as GWAS data from different types of complex diseases.

In addition, in order to ratify the results obtained by ana-

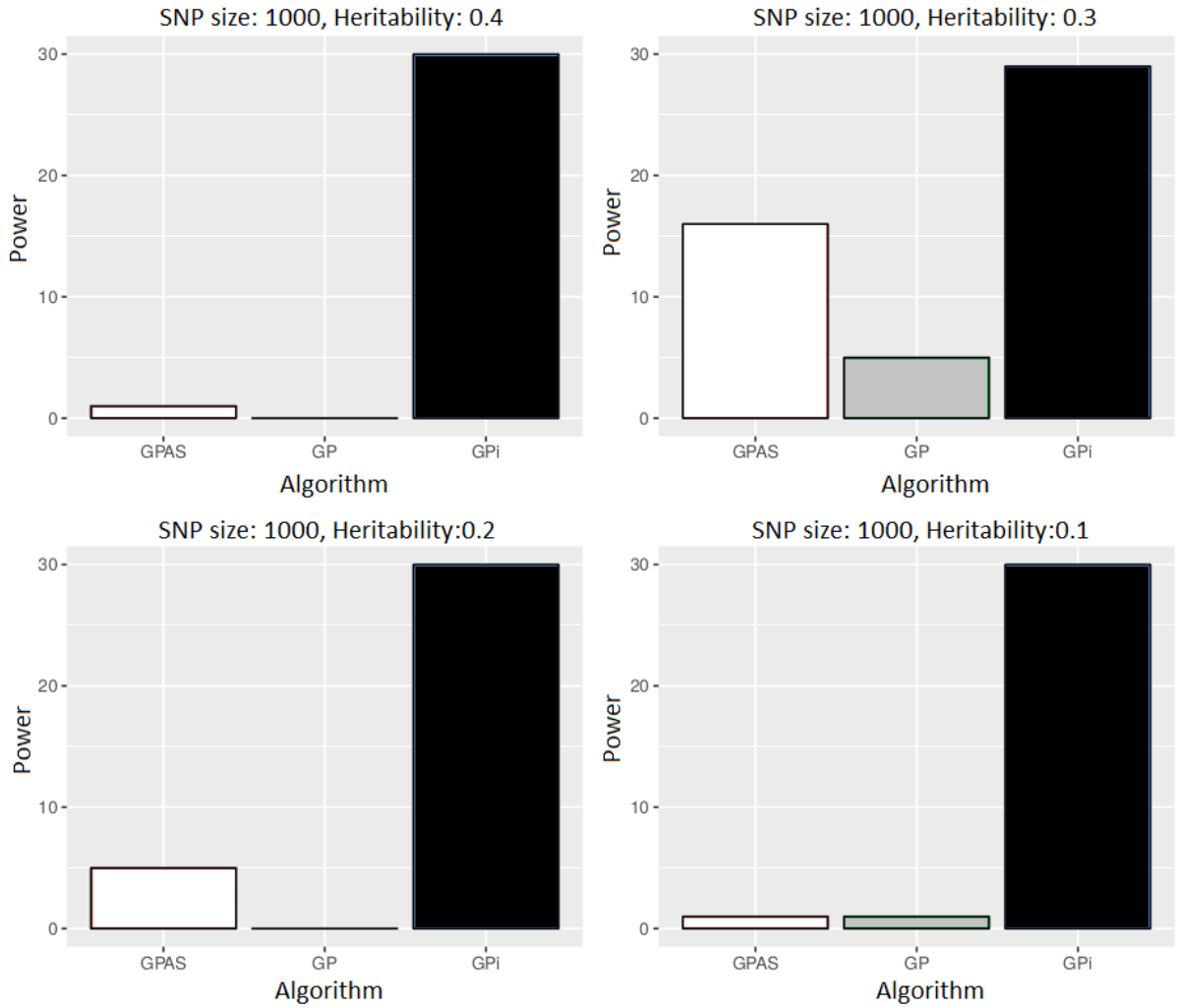

Figure 5. Graphic representing the power of each algorithm (GPAS, GP, GPi) across heritability of 0.4,0.3,0.2,0.1 with a dataset containing 1000 SNPs. The power is the number of times that the algorithm identifies the correct two functional SNPs.

lyzing the algorithm errors in the identification of the markers, other evolutionary algorithms can be used, as well as more efficient GP proficiency functions. The question of the objective function should be better investigated actually since the GP without initialization has generally presented the worst results.

Another two points to take into consideration is the initialization method and size of the databases. In the experiments present, we used databases of 100 and 1000 markers, following the experiments of similar algorithms in the literature. However, a real GWAS database has thousands or even hundreds of thousands of SNPs. This condition implies the need for efficient dimensionality reduction algorithms and / or filters. Furthermore, other classification methods can be combined to improve initialization mechanisms in cases of extremely low heritability (≤ 0.1).

5. Conclusion

A GP algorithm aims to explore all search space. However, due to the large number of possible combinations, this search may be computationally feasible. Expert knowledge approaches are recommended in these cases. in cases.

The results of the methods compared in this work showed that the use of an expert knowledge makes it possible to reduce the search space of the GP algorithm, proving to be effective, even in low heritability dataset. SNPs with higher quality of information are selected and inserted into the initial population of GP using the measure of increase in MSE of the random forest algorithm.

We show that random forest is an option among the algorithms used in other studies as expert knowledge methods and it has shown to be able to capture possible candidate markers for epistatic interactions and to be less sensitive to noises and marginal effects.

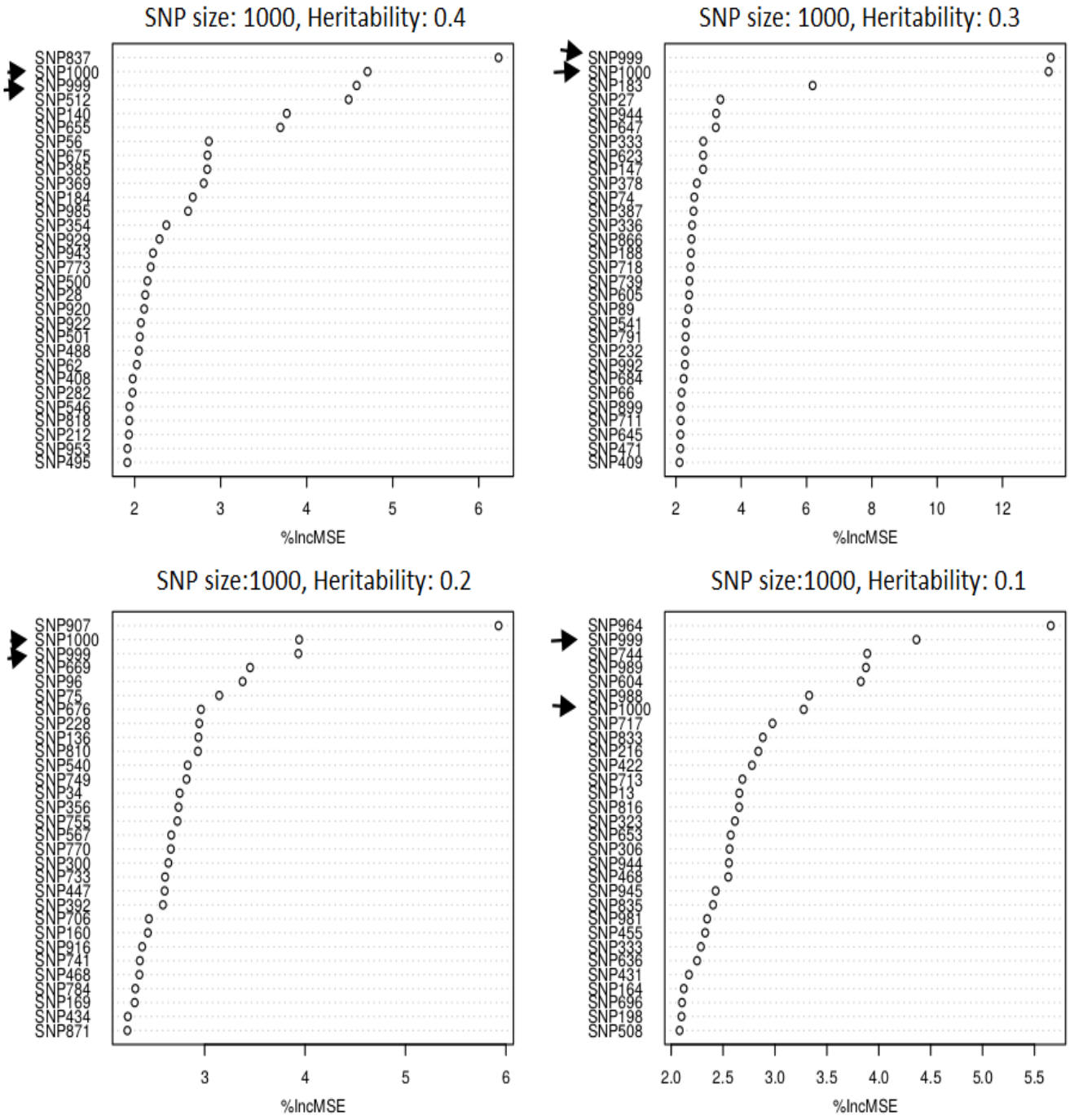

Figure 6. Ranking of the SNPs performed by the random forest algorithm. The results show the initialization step of the initial population in each scenario. The arrows indicate the positions of SNP999 and SNP1000 (the two functional SNPs).

6. Acknowledgment

The authors thanks to reviewers who gave useful comments, and would like to express thanks to the Coordination for the Improvement of Higher Level Personnel (CAPES) and National Council for Scientific and Technological Development (CNPq) and the State of Minas Gerais Research Support Agency (FAPEMIG) for the financial support for the accomplishment of this paper; and to the Postgraduate Program in Computational Modeling of Federal University of Juiz de Fora (UFJF) for the academic support.

7. Author contributions

IMG developed the proposed model, carried out the experiments, analyzed the results and contributed to the methodol-

ogy of this study. BZS structured and provided the infrastructure and computing resources needed to perform the experiments and and contributed to the proposed model. CCHB and WA are the project leaders, proposed the methodology and general approach of this study.

References

[1] MOORE, J. H.; WHITE, B. C. Tuning relieff for genome-wide genetic analysis. In: MARCHIOR, E.; MOORE, J. H.; RAJAPAKSE, J. C. (Ed.). *Proceedings of the 5th European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics*. Berlin, Heidelberg: Springer-Verlag, 2007. v. 4447, p. 166–175.

[2] BUSH, W. S.; MOORE, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol.*, v. 8, n. 12, p. 1–11, 2012.

[3] MANOLIO, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J. Med*, v. 363, n. 2, p. 166–176, 2010.

[4] GRIFFITHS, A. et al. *An Introduction to Genetic Analysis*. 7. ed. New York, USA: W. H. Freeman, 2000. v. 1.

[5] GRIFFITHS, A. J. *Introdução à genética*. 9. ed. Rio de Janeiro, Brazil: Guanabara Koogan, 2008. v. 1.

[6] TAN, H. et al. The estimation of heritability for twin data based on concordances of sex and disease. *Chronic Dis Can.*, v. 26, n. 1, p. 9–12, 2005.

[7] GU, J.; WU, X. Genetic susceptibility to bladder cancer risk and outcome. *Per Med.*, v. 8, n. 3, p. 365–374, 2011.

[8] CZENE, K.; LICHTENSTEIN, P.; HEMMINKI, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish Family-Cancer Database. *Int. J. Cancer*, v. 99, n. 2, p. 260–266, 2002.

[9] CZENE, K.; LICHTENSTEIN, P.; HEMMINKI, K. Environmental and heritable causes of cancer among 9.6 million individuals in the swedish family-cancer database. *Int J Cancer.*, v. 99, n. 2, p. 260–266, 2002.

[10] POULSEN, P.; KYVIK, K. O.; VAAG A.AND BECK-NIELSEN, H. Heritability of type ii (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance – a population-based twin study. *Diabetologia*, v. 42, n. 2, p. 139–145, 1999.

[11] MOORE, J.; WHITE, B. Genome-wide genetic analysis using genetic programming: The critical need for expert knowledge. In: RIOLO, R.; SOULE, T.; WORZEL, B. *Genetic Programming Theory and Practice IV*. 1. ed. Boston, USA: Springer, 2007. (Genetic and Evolutionary Computation, v. 1), cap. 2, p. 11–28.

[12] SZE-TO, H.-Y. et al. Gp-pi: Using genetic programming with penalization and initialization on genome-wide association study. In: RUTKOWSKI, L. et al. (Ed.). *Artificial Intelligence and Soft Computing*. 1. ed. Berlin, Germany: Springer, 2013, (Lecture Notes in Computer Science, v. 7895). cap. 30, p. 330–341.

[13] GREENE, C. S.; WHITE, B. C.; MOORE, J. H. Using expert knowledge in initialization for genome-wide analysis of epistasis using genetic programming. In: RYAN, C.; KEIJZER, M. (Ed.). *Proceedings of the 10th Annual Conference on Genetic and Evolutionary Computation*. New York, NY, USA: ACM, 2008. v. 1, p. 351–352.

[14] MOORE, J.; WHITE, B. *Genome-Wide Genetic Analysis Using Genetic Programming: The Critical Need for Expert Knowledge*. 2007.

[15] KIRA, K.; RENDELL, L. A. A practical approach to feature selection. In: SLEEMAN, D.; EDWARDS, P. (Ed.). *Proceedings of the Ninth International Workshop on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1992. v. 1.

[16] NUNKESSER, R. et al. Detecting high-order interactions of single nucleotide polymorphisms using genetic programming. *Bioinformatics*, v. 23, n. 24, p. 3280–3288, 2007.

[17] BLEULER, S. et al. Multiobjective Genetic Programming: Reducing Bloat by Using SPEA2. In: CEC 2001. *Congress on Evolutionary Computation*. Seoul, South Korea: IEEE, 2001. v. 9.

[18] LUKE, S. et al. *ECJ 16: A Java-based Evolutionary Computation Research System*. 2007.

[19] R.C.R TEAM. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2008.

[20] LIAW, A.; WIENER, M. Classification and regression by randomforest. *R News*, v. 2, n. 3, p. 18–22, 2002.

[21] URBANOWICZ, R. J. et al. METHODOLOGY GAMETES : a fast , direct algorithm for generating pure , strict , epistatic models with random architectures. *BioData Min.*, v. 5, n. 16, p. 1–14, 2012.