SNPs Selection using Gravitational Search Algorithm and Exhaustive Search for Association Mapping

W A Kusuma[1], L S Hasibuan, M A Istiadi

Department of Computer Science, Bogor Agricultural University, INDONESIA
Jalan Meranti Wing 20 level 5 Kampus IPB, Bogor, West Java, Indonesia

Email: ananta@apps.ipb.ac.id

Abstract. Single Nucleotide Polymorphisms (SNPs) are known having association to phenotipic variations. The study of linking SNPs to interest phenotype is refer to Association Mapping (AM), which is classified as a combinatorial problem. Exhaustive Search (ES) approach is able to be implemented to select targeted SNPs exactly since it evaluate all possible combinations of SNPs, but it is not efficient in terms of computer resources and computation time. Heuristic Search (HS) approach is an alternative to improve the performance of ES in those terms, but it still suffers high false positive SNPs in each combinations. Gravitational Search Algorithm (GSA) is a new HS algorithm that yields better performance than other nature inspired HS. This paper proposed a new method which combined GSA and ES to identify the most appropriate combination of SNPs linked to interest phenotype. Testing was conducted using dataset without epistasis and dataset with epistasis. Using dataset without epistasis with 7 targeted SNPs, the proposed method identified 7 SNPs – 6 True Positive (TP) SNPs and 1 False Positive (FP) SNP- with association value of 0.83. In addition, the proposed method could identified 3 SNPs- 2 TP SNP and 1 FP SNP with association value of 0.87 by using dataset with epistases and 5 targeted SNPs. The results showed that the method is robust in reducing redundant SNPs and identifying main markers.

1. Introduction

Association mapping is a study aiming to detect linkage between genetic polymorphisms and phenotypic variations in existing germplasm. One of the most popular marker to identify genetic polymorphisms is Single Nucleotide Polymorphisms (SNPs), since it allows generation of abundant information on genetic variability at DNA level. The analysis of AM has increased human understanding of phenotipic variations and heritability in human, animals and plants [1][2].

A number of methods for analyzing AM have been proposed before. In the conventional methods, AM is conducted based on investigation of single locus and the interest phenotype [3]. Nevertheless, the methods are not suitable for complex phenotypes with epistases interaction among genes [3][4]. The search for multi-locus association is potential to explain the phenomenon of complex phenotypes and genetic polymorphisms. Multi-locus association is classified as a combinatorial problem, it seems simple to be solved but hard to be implemented because the search space would increase exponentially according to the growth of data.

Multi-locus AM with 1000 SNPs require computer with high performance to evaluate 8.25×10^{12} combinations for five SNPs. Reducing the redundant and uncorrelated SNPs in the search space is the important task. One of the reducing techniques could be applied is using heuristic search and followed by conducting exhaustive search to the remaining SNPs to find the most appropriate combination of SNPs. This approach was implemented using Ant Colony Optimization (ACO) and exhaustive search in the case of case-control phenotype [3].

Gravitational Search Algorithm (GSA) is one of the latest heuristic search that developed based on the law of gravity and mass interaction. It was first introduced by Rashedi et al in 2009 [5]. It becomes to continue growing and being popular in researchers. Many modified and hybrid GSA have already proposed by researchers. They showed that GSA has better performance in terms of computation time and convergence than the previous heuristic search [6]. In this study, we proposed a method combining GSA and exhaustive search to find the most appropriate combination of SNPs that has linkage to quantitative phenotype.

2. Material

The proposed method was tested using two types of simulated dataset which were used by Oliveira et al [7]. The first dataset only has main effects without interaction among SNPs, while the second one with epistasis among SNPs. The dataset were generated by the function of *simulateSNPglm* of the *scrime* package in the R software. No filters HWE and *call rate* in simulated dataset were used, since there were not any missing values and no markers in Hardy-Weinberg disequilibrium.

2.1. Simulated dataset without epistasis 1

The SNP data were given in form of 1, 2 or 3. Notation 1 for homozygous reference sequence, 2 for heterozygous genotype and 3 for homozygous variant sequence. A thousand SNPs were simulated for 250 subjects, with a minor allelic frequency (MAF), simulated for each SNP, based on a uniform distribution with minimum and maximum limits were 0.10 and 0.40 respectively. The phenotype is continous and generated by Equation 1:

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + \beta_4 L_4 + \beta_5 L_5 + \beta_6 L_6 + \beta_7 L_7 + error \tag{1}$$

where *error* was a normal random variable with mean was 0 and standard deviation was 5, L_1 = (SNP1 == 2), L_2 = (SNP10 == 1), L_3 = (SNP20 == 3), L_4 = (SNP30 == 3), L_5 = (SNP40 == 3), L_6 = (SNP50 == 2), L_7 = (SNP60 == 2). The beta coefficients were set as $\beta_0 = 0$, $\beta_1 = \beta_2 = \beta_3 = 200$, $\beta_4 = 900$, $\beta_5 = \beta_6 = \beta_7 = 200$.

2.2. Simulated dataset with epistasis 2

Epistasis could be defined as a form of functional interaction among genes that caused not common phenomenon in phenotypic variations [4]. Illustration about epistatis is showed in Figure 1.

The SNP data were given in form as used before in simulated dataset 1, likewise MAF adjusment. Ten thousand SNPs were simulated for 600 subjects with phenotype as defined in Equation 2:

$$Y = \beta_0 + \beta_1 L_1 + \beta_2 L_2 + \beta_3 L_3 + error \tag{2}$$

where error was a normal random variable with mean was 0 and standard deviation was 1, L_1 = (SNP4 != 2)[3] & (SNP3 != 1), L_2 = (SNP5 == 3), L_3 = (SNP12 != 1) & (SNP9 == 3). The beta coefficients were set as $\beta_0 = 0$, $\beta_1 = \beta_2 = 150$ and $\beta_3 = 40$. [3]The SNP4 != 2 symbol means that the SNP4 does not have the largest allele frequency represented by 2. Figure 2 shows the histogram and boxplot for the datasets.

3. Method

The SNPs selection method proposed in this study consist of two stages, selection using heuristic search in the first stage and exhaustive search in the second one. As mention in introduction, finding the most appropriate combination of SNPs that has linkage to quantitave phenotype is classified as a combinatorial problem, for which the worst case time requirement grows exponentially with the number of SNPs. To address this problem, heuristic search was conducted to select a number of possible SNPs that have linkage to phenotype. In other words, the first stage was aimed to reduce the number of redundant SNPs, this stage was conducted two times. However, solution provided by heuristic seach is not necessary as the best solution for the given problem [6]. Therefore, exhaustive search was needed in the final stage to select targeted SNPs. The association between SNPs and phenotype was conducted using SVR [8] first, SVR generated predict phenotype based on the given combination of SNPs and the predict phenotype was associated to the phenotype in simulated dataset. The correlation between predicted and simulated phenotype was being the evaluation of the fitness function.

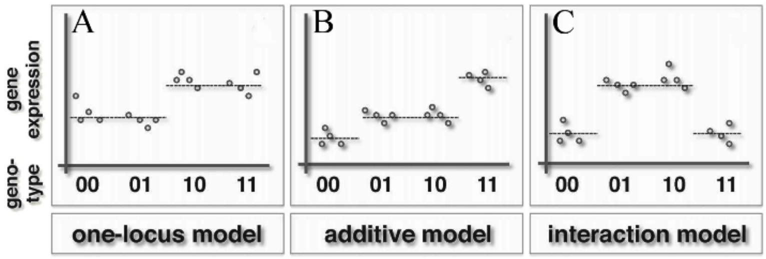

Figure 1. Illustration of the one-locus, additive and epistatic effects models in haploid organisms. The x-axis represent the combined genotype of the pair of loci while the y-axis represent expression level of the underlying gene. **(A)** In the one-locus model the genotype of one locus (here the first in the pair) drives the given phenotype. **(B)** In the additive model both loci contribute to the phenotype in an additive way. **(C)** When an epistatic interaction occurs, the effect of the two loci on the trait is non-additive [4].

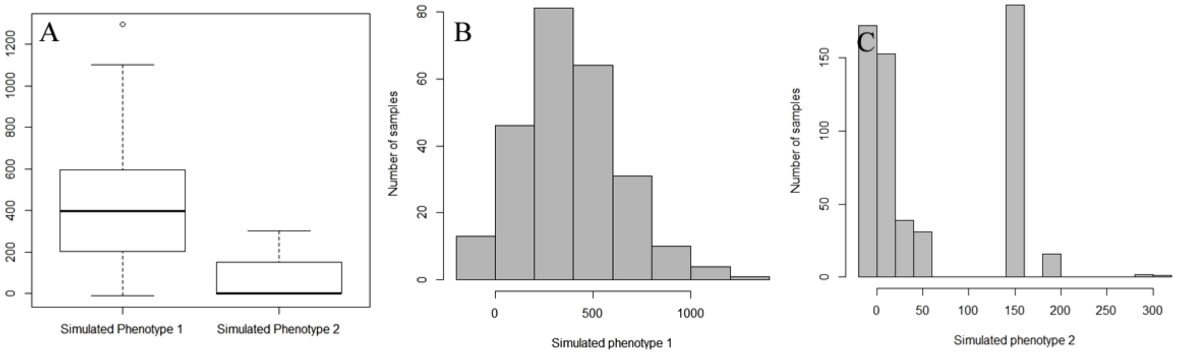

Figure 2. Boxplot and histogram of the datasets

3.1. Gravitational Search Algorithm (GSA)

One of the latest heuristic search proposed by Rashedi et. al. in 2009 is GSA. It was developed based on the law of gravity and mass interactions [5]. As the common heuristic search, GSA is comprised of parallel artificial agents and each agent represents a solution of the problem. Consider an artificial population consist of *N* agents, then solution of the i^{th} agent is defined by

$$X_i = \left(x_i^1, \ldots, x_i^d, \ldots, x_i^n\right) \; for\, i = 1, 2, \ldots, N \tag{3}$$

The solution of the i^{th} agent represents a combination of n SNPs, where x_i^d represents the d^{th} SNP in a combination and n is the dimension, represent the desired number of SNPs for a combination.

All agents attract each other by a gravity force, and this force causes a movement of all agents globally towards the agents with heavier mass. Mass of an agent correspond to the evaluation of the fitness function for the given solution. The attraction and the movement are iterative procedures which would be stopped at a predefined number of iterations, T .

At a specific time t , the force acting on agen i from agen j is defined as the following:

$$F_{ij}^{d} = G(t)\frac{M_{pi}(t)\times M_{aj}(t)}{R_{ij}(t)+\varepsilon}\left(X_{j}^{d}(t)-X_{i}^{d}(t)\right) \tag{4}$$

where M_{aj} is the active gravitational mass related to agent j, M_{pi} is the passive gravitational mass related to agent i, $G(t)$ is the gravitational constant at time t, ε is a small constant, and $R_{ij}(t)$ is the Euclidian distance between two agents i and j. The effects of distance between SNPs is obviated, so that $X_j^d - X_i^d$ is set to 1 and R_{ij} is set to 0. The total force that acts on agen i in a dimension d is randomly weighted sum of dth component of the forces exerted from $Kbest$ agents:

$$F_{i}^{d}(t) = \sum_{j\in Kbest,\, j\neq i} rand_{j} F_{ij}^{d}(t) \tag{5}$$

where $rand_j$ is a random number in the interval [0,1] and $Kbest$ is the set of first K agents with the best fitness value and biggest mass. $Kbest$ is a function of time, initialized to K_0 at the beginning and decreasing with time.

According to the law of motion, the acceleration of the agent i, at time t, in the dth dimension, is given as follows:

$$a_{i}^{d}(t) = \frac{F_{i}^{d}(t)}{M_{ii}(t)} \tag{6}$$

where M_{ii} is the is the inertial mass of the ith agent. The next velocity of an agent is considered as a fraction of its current velocity added to its acceleration. Therefore, its velocity and its position could be calculated as follows:

$$v_{i}^{d}(t+1) = rand_{i}\times v_{i}^{d}(t)+a_{i}^{d}(t) \tag{7}$$

$$x_{i}^{d}(t+1) = x_{i}+v_{i}^{d}(t+1) \tag{8}$$

where $rand_i$ is an uniform random number in the interval [0,1].

The gravitational constant, G, is initialized at the begining and will be reduced with time to control the search accuracy. In other words, G is a function of the initial value (G_0) and time (G):

$$G(t) = G(G_0, t) \tag{9}$$

$$G(t) = G_0 e^{-\frac{\alpha t}{T}} \tag{10}$$

The masses of the agents are calculated using fitness function. A heavier mass means a more efficient agent to solve the problem and moves more slowly. Supposing the equality of the gravitational and inertia mass, the values of masses is calculated using the map of fitness. The gravitational and inertial masses are updated by the following equations:

$$M_{ai} = M_{pi} = M_{ii} = M_{i}, i = 1,2,\ldots,N \tag{11}$$

$$q_i(t) = \frac{fit_i(t) - worst(t)}{best(t) - worst(t)} \tag{12}$$

$$M_i(t) = \frac{q_i(t)}{\sum_{j=1}^{N} q_j(t)} \tag{13}$$

where $fit_i(t)$ represents the fitness value of the agent i at time t, and the $best(t)$ and $worst(t)$ in the population indicate the strongest and the weakest agent according to their fitness evaluation. For a maximization problem:

$$best(t) = max_{j\in\{1...,N\}}\, fit_j(t) \tag{14}$$

$$worst(t) = min_{j\in\{1...,N\}}\, fit_j(t) \tag{15}$$

The principle of GSA is shown in Figure 3.

SNPs of the best agent, $x_j^1,\ldots,x_j^d,\ldots,x_j^n$, in every iteration were collected and then added into detected SNPs set, L_d, at the end of iteration. The reduction of redundant SNP is conducted two times to get L_d with less redundant SNPs at the end of process.

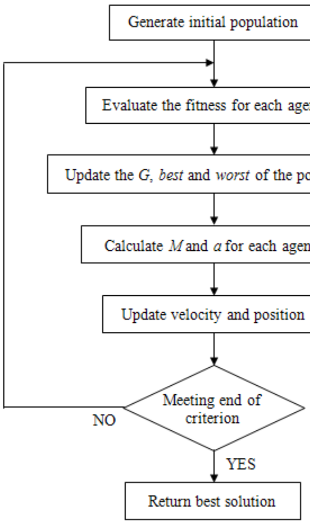
Generate initial population
Evaluate the fitness for each agent
Update the G, best and worst of the population
Calculate M and a for each agent
Update velocity and position
Meeting end of criterion
NO
YES
Return best solution

Figure 3. General principle of GSA [5]

3.2. *Exhaustive Search*

The number of outcome SNPs of the previous stage were less than the number of outset SNPs, $L_d < L_{all}$. Exhaustive search was conducted to select the most appropriate combination of SNPs that has linkage to the interest phenotype. The first step, exhaustive search determined the best combination of two SNPs. The two selected SNP were excluded from L_d and added into solution set L_s. Each remaining SNP in L_d were attempted to be added into L_s iteratively. The addition was allowed if it can improve association value between SNPs in L_s and phenotype. The addition will be stopped when no more SNP in L_d will improve association value if it was added into L_s.

4. Results and Discussions

The performance of the proposed method applied to both simulated dataset was compared to the method proposed by Oliveira et al [7] and generic GSA [5] itself. The comparison was conducted using two simulated dataset, without epistasis and with epistasis.

4.1. Simulated dataset without epistasis 1

Combination of SNPs generated by the proposed method and the other two methods using simulated dataset 1 were showed in Table 1. From the table, it was seen that the generic GSA was able to identify all of targeted SNPs but suffer of many false positive SNPs. The false positive SNPs may lead models' performance to predict phenotype become low, even though all of targeted SNPs consist in the generated combination. As could be seen in Tabel 1, correlation of the generic GSA was the lowest, 0.262, compared to the other methods.

Table 1. Combination of SNPs and correlation of methods in simulated dataset 1

Method	Identified SNPs	Correlation
Oliveira method	**1**, **10**, 15, **20**, **30**, **60**, 158, 177, 269, 274, 391, 446, 516, 673, 686, 693, 717, 725, 739, 825, 930**	0.750
Generic GSA*	**1**, **10**, **20**, **30**, **40**, **50**, **60** and 170 redundant SNPs**	0.262
Proposed Method	**1**, **10**, **20**, **30**, **50**, **60**, 72**	0.830

*)SNPs collection of best agen in every iteration
**)The bold SNPs is the targeted SNPs

The method proposed by Oliveira et al [7] was able to reduce 977 of 993 redundant SNPs, but the given combination still suffer of many false positive SNPs, 16 false positive SNPs. Besides that, there were two other targeted SNPs, the SNP40 and the SNP50, unidentified by the method. The correlation was higher than the generic GSA but lower than the proposed method.

As mention before, the generic GSA [5] was able to identify all of targeted SNPs. Besides that, it was also able to reduce redundant SNPs, 823 of 993 redundant SNPs, although the result was not better than that of the Oliveira method. Because all of targeted SNPs were consisting in the combination generated by generic GSA and the number of redundant SNPs was lower than those of in the outset SNPs, it was possible to conduct exhaustive search to find the most appropriate combination of SNPs. From Table 1, it could be seen that the performance of the proposed method was better than the Oliveira method in term of identifying targeted SNPs. Besides that, the number of false positive SNPs in the proposed method was lowest than the two other methods. Moreover, the correlation yielded by the proposed method was the highest among others, it was 0.83.

There was one SNP, the SNP40, which was not able to be identified by the Oliveira method and the proposed method. Figure 4 showed the relationship between the SNP40 and the simulated phenotype 1. From the Equation 1, it was known that SNP40 would affect to the phenotype when its value was 3. The Figure 4 showed that the number of sample with the value of SNP40 is 3 was less than the number of sample that the value of SNP40 is not 3. Figure 5 showed the relationship between the SNP30 and the simulated phenotype 1, the same thing also happening to SNP30. From the Equation 1, it was known that SNP30 will affect to phenotype when its value is 3. Figure 5 showed that the number of sample that the value of SNP30 is 3 was less than the number of sample that the value of SNP30 is not 3 but the models were able to identified it. From the Equation 1, it could be seen that the beta coeficient of SNP30 was greater almost five folds than the other beta coeficients. It means that, SNP30 is the main marker for the given phenotype. Meanwhile, the SNP40 is not the main marker and only a few samples show effect to phenotype because of this marker. Therefore, the models could not identified the SNP40 except generic GSA.

The comparison of models performance based on true positive rate, false positive rate and correlation was showed in Figure 6. Generally, the proposed method and the Oliveira method [7] were able to reduce redundant SNPs than the generic GSA [5]. The proposed method was better in identifying targeted SNPs than the Oliveira method, it could be seen from TPR and correlation metrics were higher than the Oliveira methods.

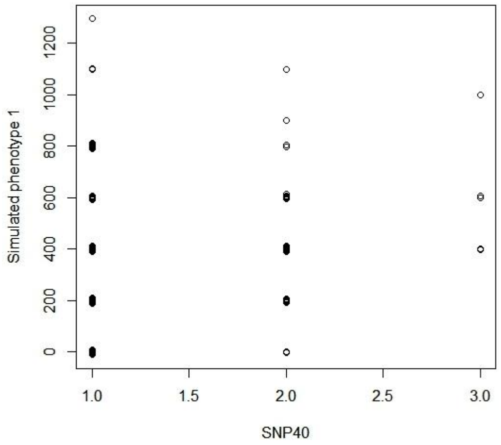

Figure 4. Scatter plot between the SNP40 and simulated phenotype 1

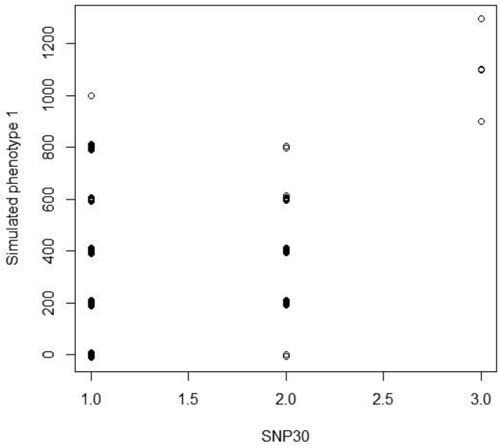

Figure 5. Scatter plot between the SNP30 and simulated phenotype 1

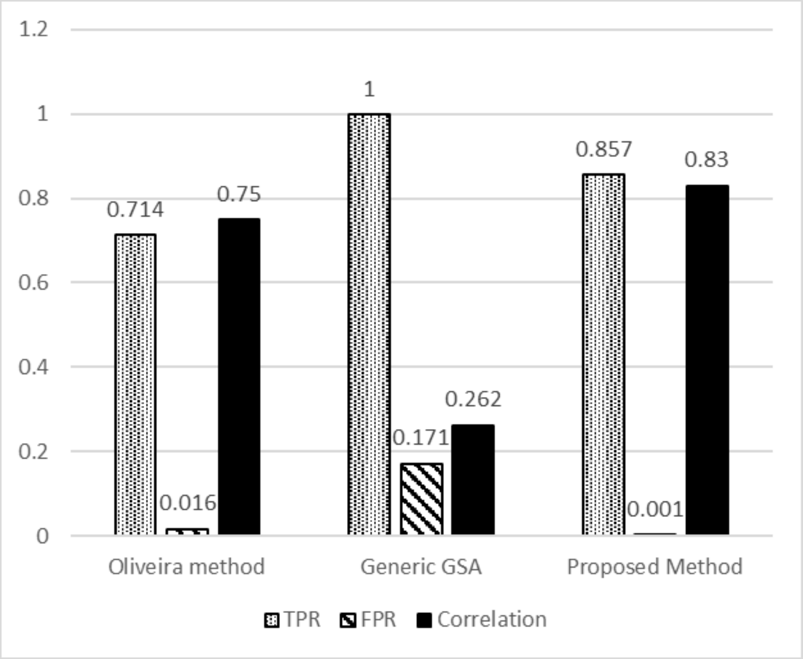

Figure 6. Methods performance comparison in simulated dataset 1

4.2. Simulated dataset with epistasis 2

The performance of the proposed method using simulated dataset 2 was showed in Table 2. From the table, it could be seen that all methods were able to reduce redundant SNPs except the method by generic GSA [5]. Nevertheles, the ability of all methods in identifying targeted SNPs was lower than before.

The generic GSA was able to identify two targeted SNPs, SNP3 and SNP4, while it suffer with 62 false positive SNPs and its correlation was 0.23. The Oliveira method [7] was able to identify one targeted SNP, the SNP3, with no false positive SNP and the correlation was 0.95. The proposed method was able to identify two targeted SNPs, SNP3 and SNP4, with one false positive SNP. The correlation was 0.87. From Table 2, it was seen that only SNP3 could be identified by the proposed methods.

In the simulated dataset 2, there was only one sample with condition (SNP4 != 2) and (SNP3 != 1) and there was no sample with condition (SNP12 != 1) and (SNP9 == 3). According to [7], the simulated dataset 2 was generated with high deviation and asymmetry data. It was done to examine the robustness of the method.

The comparison of models performance based on true positive rate, false positive rate and correlation using simulated dataset 2 was showed in Figure 7. From the figure, it could be seen that TPR of all methods were low and the correlations were high, except the generic GSA method [5]. It is mean that the proposed method and the Oliveira method [7] are robust in reducing the redundant SNPs and identifying main marker affected the simulated phenotype dataset 2.

Table 2. Combination of SNPs and correlation of methods in simulated dataset 2

Method	Identified SNPs	Correlation
Oliveira method	**3**	0.95
Generic GSA	**3**, **4** and 62 redundant SNPs*	0.23
Proposed Method	**3**, **4**, 8029	0.87

*)SNPs collection of the best agen in every iteration
**)The bold SNPs is the targeted SNPs,

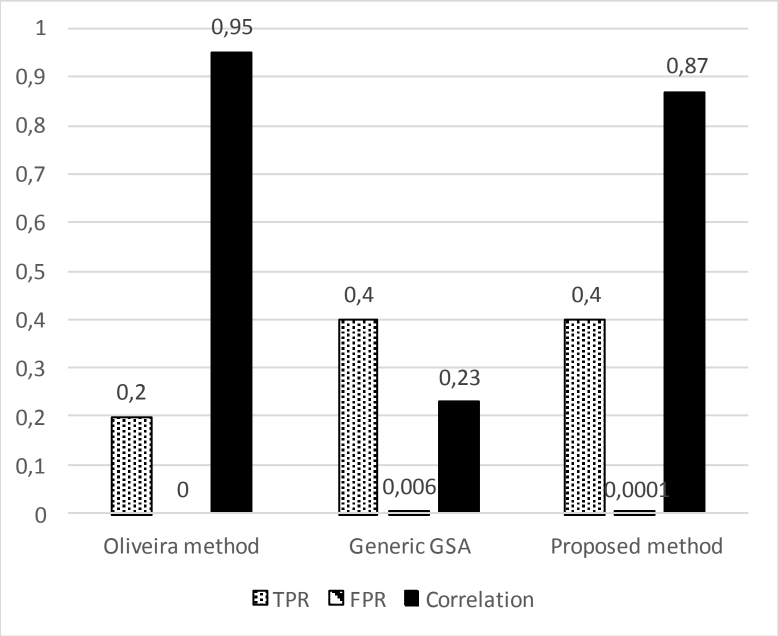

Figure 7. Methods performance comparison in simulated dataset 1

5. Conclusion

The proposed method successfully combined GSA and exhaustive search to solve the association mapping problem. The method was tested using two types of simulated datasets, without epistasis and with epistasis. In both simulated datasets, the proposed method shows the robussnest in reducing the redundant SNPs and identifying the main marker SNP which affect the simulated phenotype. It is concluded that the proposed method has good potentiality to be applied in association mapping with real dataset.

Acknowledgment
The authors would like to thank to Indonesia Ministry of Agriculture for funding this research through the KKP3N 2015 program.

References

[1] Cobb, J. N., DeClerck, G., Greenberg, A., Clark, R., & McCouch, S. 2013 Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement *Theoretical and Applied Genetics* **126(4)** 867-887

[2] Minzhu, X., Jing, L., & Tao, J. Detecting genome-wide epistases based on the clustering of relatively frequent items *Bioinformatics* **28.1 (2012)** 5-12

[3] Wang, Y., Liu, X., Robbins, K., & Rekaya, R 2010 AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm *BMC research notes* **3(1)** 117

[4] Huang, Y., Wuchty, S., & Przytycka, T. M. 2013 eQTL epistasis–challenges and computational approaches *Frontiers in genetics* **4**

[5] Rashedi, E., Nezamabadi-Pour, H., & Saryazdi, S 2010 BGSA: binary gravitational search algorithm *Natural Computing* **9(3)** 727-745A

[6] Sahoo, G. 2014 A Review on gravitational search algorithm and its applications to data clustering & classification *International Journal of Intelligent Systems and Applications (IJISA)* **6(6)** 79

[7] de Oliveira, F. C., Borges, C. C. H., Almeida, F. N., e Silva, F. F., da Silva Verneque, R., da Silva, M. V. G., & Arbex, W 2014 SNPs selection using support vector regression and genetic algorithms in GWAS *BMC genomics*, **15**

[8] Smola, A., & Vapnik, V 1997 Support vector regression machines *Advances in neural information processing systems* ***9*** 155-161.