Theoretical and Applied Informatics
ISSN 1896–5334
Vol. 25 (2013), no. 3–4
pp. 251 – 262
DOI: 10.2478/thai-2013-0008

EA-MOSGWA – a tool for identifying associated SNPs in Genome Wide Association Studies

ARTUR GOLA [a], MAŁGORZATA BOGDAN [b], FLORIAN FROMMLET [c]

[a]Department of Mathematics and Computer Science, Jan Długosz University in Częstochowa, Poland

[b]Department of Mathematics and Computer Science, Wrocław University of Technology, Poland

[c]Department of Medical Statistics, Medical University Vienna, Austria

Received 2 December 2013, Revised 12 December 2013, Accepted 22 December 2013.

Abstract: This paper presents the current stage of the development of EA-MOSGWA – a tool for identifying causal genes in Genome Wide Association Studies (GWAS). The main goal of GWAS is to identify chromosomal regions which are associated with a particular disease (e.g. diabetes, cancer) or with some quantitative trait (e.g height or blood pressure). To this end hundreds of thousands of Single Nucleotide Polymorphisms (SNP) are genotyped. One is then interested to identify as many SNPs as possible which are associated with the trait in question, while at the same time minimizing the number of false detections.

The software package MOSGWA allows to detect SNPs via variable selection using the criterion mBIC2, a modified version of the Schwarz Bayesian Information Criterion. MOSGWA tries to minimize mBIC2 using some stepwise selection methods, whereas EA-MOSGWA applies some advanced evolutionary algorithms to achieve the same goal. We present results from an extensive simulation study where we compare the performance of EA-MOSGWA when using different parameter settings. We also consider using a clustering procedure to relax the multiple testing correction in mBIC2. Finally we compare results from EA-MOSGWA with the original stepwise search from MOSGWA, and show that the newly proposed algorithm has good properties in terms of minimizing the mBIC2 criterion, as well as in minimizing the misclassification rate of detected SNPs.

Keywords: Evolutionary Algorithm, Genome Wide Association, linear regression

1. Introduction

Recently there has been considerable interest in developing variable selection methods for Genome-wide association studies (GWAS), see for example [7] or [10]. A comprehensive overview can be found in [4], where the methods implemented in the software package MOSGWA were first introduced in the context of quantitative traits. MOSGWA

uses mBIC2, a modified version of the Bayesian Information Criteria which is designed to control the false discovery rate. A fully Bayesian variable selection approach for GWAS based on Markov chain Monte Carlo (MCMC) sampling was developed by [6]. Some review on the latest methods for GWAS analysis can be found for example in [1].

The original MOSGWA package as described in [4] tries to minimize the model selection criterion mBIC2 using some sophisticated stepwise selection procedures. In contrast EA-MOSGWA makes use of some specifically designed evolutionary algorithm [5, 8] to fulfill the same task, hence the name of the new method is EA-MOSGWA. The evolutionary algorithm we present here is very similar to the memetic algorithm used for QTL mapping which was described in [3].

EA-MOSGWA has been developed at the Department of Mathematics and Computer Science of Jan Długosz University in Czestochowa in cooperation with the Wroclaw University of Technology and the Medical University of Vienna. EA-MOSGWA is a module of the larger, more elaborate program MOSGWA, which has been developed for several years at the Medical University of Vienna and Wroclaw University of Technology. The program MOSGWA (Model Selection for Genome Wide Associations) is an advanced tool to analyze GWAS data, and its software architecture is flexible enough to make it relatively easy to incorporate new methods.

2. Theoretical Background

We will focus in this article on the detection of SNPs which are associated with a quantitative trait. The measurement of the trait for the ith individual is denoted as $y_i, i \in \{1, \ldots, n\}$. For each individual the genotypes of p SNPs are assumed to be known, where we use the coding $x_{ij} \in \{-1, 0, 1\}, j \in \{1, \ldots, p\}$. To model the genetic influence on the trait me make use of a classical linear regression model of the form

$$y_i = \mu + \sum_{j \in M} \beta_j x_{ij} + \epsilon_i, \tag{1}$$

where μ is the intercept and M denotes some subset of markers which influence the trait. We assume that the error term is normally distributed, $\epsilon_i \sim \mathrm{N}\left(0, \sigma^2\right)$. Model selection is then performed by trying to minimize the following modification of the Bayesian information criterion

$$\text{mBIC2} := n \log \text{RSS} + k \log(np^2/16) - 2 \log(k!), \tag{2}$$

where $k = |M|$ denotes the size of the evaluated model, n is the number of individuals, and p is the number of SNPs. An extensive justification of this approach is provided in [4].

In principle one might consider to compute the criterion for all possible subsets of SNPs and thus find the model which minimizes mBIC2. However, this approach has exponential complexity in p, and therefore all subset selection is entirely hopeless given the vast number of SNPs in GWAS. In practice it is therefore necessary to perform variable selection by searching only over a subset of models. To this end EA-MOSGWA uses an evolutionary algorithm for searching through the space of all subsets, trying to find a model which minimizes mBIC2. During the search the algorithm stores the mBIC2 values of all visited models, and the gathered information is used to direct the search into regions of the solution space which appear to be most promising. As a result EA-MOSGWA not only reports that model along its search path that minimizes mBIC2, but it provides the information of a large number of interesting models.

3. Implementation

The idea of evolutionary algorithms is inspired by observing the development of species, where over several generations a population is getting better adapted to the environmental conditions in which it resides. In analogy in the case of computer programs evolutionary algorithms are trying to obtain better solutions by allowing a population of existing solutions to evolve in such a way, that the fitness of the population increases [5, 8].

The evolutionary algorithm used by EA-MOSGWA is fairly similar to the Memetic Algorithm (MA) presented in [3]. The memetic algorithm works with a population of individuals, where each individual corresponds to a particular model M according to (1). The population then refers to a collection of such models, where we will denote the population size by u. The fitness of each individual M is measured by the mBIC2 criterion (2). The smaller the value of mBIC2, the fitter is an individual.

MA from [3] was implemented in Matlab, and it was designed to work on experimental populations, where usually less than 500 markers are explored. In contrast our new algorithm works for several hundred thousand markers as in GWAS. The general structure of both algorithms is the same, but due to the much larger number of potential regressors our new algorithm differs in several ways, as will be described in the subsections below.

The general outline of EA-MOSGWA is as follows. The algorithm starts with creating an initial population of size u. Then three evolutionary steps are repeated till some stopping criterion is fulfilled. Each iteration starts with a selection step, where two individuals from the population are chosen as parents. Next comes the recombination step in which with rather large probability a new individual, the so called child, is generated as an offspring from the two parent models. In a third step with rather low probability the child then undergoes mutation. Finally at the end of each iteration local improve-

ment takes place,where the fitness of the child is further improved by some greedy local search.

At the end of each iteration the fitness of the new individual is compared with the weakest member of the current population. If the fitness of the new individual is better, then the population is updated by substituting the weakest individual by the new individual. Otherwise no update occurs. The algorithm terminates after a certain number I_S of consecutive iterations took place without any update ranging among the B best models of the population.

In the following subsections we will describe the main ingredients of our algorithm in more detail.

3.1. Representation of Models

For EA-MOSGWA models have to be coded in a suitable way to become individuals of the memetic algorithm. A model consists of a set of SNPs which might be thought of as causal SNPs. Each set of SNPs uniquely characterizes a linear regression model as specified in (1).

3.2. Initial Population

The first step of EA-MOSGWA is to generate an initial population, which is obtained by repeatedly performing some rather specific greedy forward selection steps, like the multi-forward step described in [4]. To this end all SNPs are first ordered according to their marginal p-values, and then a directed search is performed along this order. Markers are added to the model whenever they lower the original BIC criterion

$$BIC = n \log(RSS) + k \log(n) \; . \tag{3}$$

The number of potential SNPs to be added to the models is reduced by considering only those markers for which the marginal p-values are smaller than 0.15. For computational purposes the size of models which are created in the initial population is limited to 150 SNPs. This procedure is then repeated iteratively to obtain all members of the initial population, where the multi-forward search is performed always over the set of SNPs which has not yet been selected before. We know that BIC has a tendency to select too large models, and we therefore deliberately design the process of generating the initial population in such a way that a large number of potentially interesting SNPs are considered.

3.3. Selection

For the selection step we use the classic method of tournament selection, where the number of participants in each tournament equals to 2. Specifically in a tournament two

individuals from the population are selected randomly, and the model with better fitness is the winner. The winner of the first tournament becomes the first parent. Then a second tournament is performed, and its winner becomes the second parent if it differs from the first parent. Otherwise the second tournament is repeated till two different parents are obtained, which are then used in the following recombination step to generate a new offspring.

3.4. Recombination

In the recombination step we perform some forward selection procedure and some backward selection procedure to generate an offspring. Let S_1 and S_2 be the sets representing the two parents. Let $S_I = S_1 \cap S_2$ be the intersection, $S_U = S_1 \cup S_2$ be the union, and $S_D = S_U \backslash S_I$ be the symmetrical difference of the two parents.

- The forward selection procedure starts from S_I and then consecutively includes markers of S_D in a greedy fashion. The fittest model obtained by this forward search is taken as the child candidate from forward selection.
- The backward elimination procedure starts from S_U and consecutively removes markers of S_D in a greedy fashion. The fittest model obtained by this backward elimination is taken as the child from the backward selection.

The fitter one of the two models obtained above will become the child obtained by recombination. A recombination step is performed with probability $pCross = 0.95$.

3.5. Mutation

In the mutation step two actions can occur with the same probability: either a new marker is added to the child model, or a marker is removed from the child model.

In the case of insertion, a new marker is selected at random from all SNPs which are not in the model. If that new marker is strongly correlated with one of the SNPs which are already in the model (that is $|R| > 0.5$ where R is the pairwise correlation), then the selection of a new marker is repeated.

In the case of removal, one randomly selects a marker from the model to be removed. If the model has only one marker then this SNP is not removed but replaced by another randomly chosen marker.

In the case that in an iteration no recombination was performed than a mutation step is mandatory. Otherwise a mutation step follows with $pMutation = 0.25$.

3.6. Local Improvement

The local improvement step probably differs most from MA of [3]. As in MA we try to improve the model by exploiting the known correlation structure between markers,

but in GWAS the correlation structure is much more complicated than in QTL mapping.

Consider an individual $S = \{s_1, s_2, \ldots, s_k\}$ with k markers. Having calculated mBIC2 for S we try to improve the model by exchanging s_1 with markers which are strongly correlated ($|R| > 0.5$) and close to s_1 (window size 50), while keeping all other markers fixed. We then continue iteratively with such exchange steps for all the remaining markers within the model.

4. Experimental Results

Simulations were carried out using real genetic data from the POPRES sample [9]. We used the imputed genotype data of 23171 SNPs from chromosome 6 from 4077 individuals.

4.1. Causal Model

Table 1: 20 SNPs selected to be causal for the simulation study. All SNPs are from the 6th chromosome of the POPRES sample. The consecutive columns contain: SNP number in the data set, SNP id, position (in base pairs) and the regression coefficient (β_j)

SNP no	SNP Id	Pos	β_j
13	SNP_A-1871676	197772	0.05
1207	SNP_A-1984915	6106312	0.06
2404	SNP_A-1834615	12980206	0.07
3611	SNP_A-1949543	20443039	0.08
4800	SNP_A-2287359	29479863	0.09
6004	SNP_A-1985686	36347031	0.10
7207	SNP_A-1886942	44806243	0.11
8423	SNP_A-2139356	53493302	0.12
9602	SNP_A-1828353	67482625	0.13
10803	SNP_A-2157434	77315837	0.14
12008	SNP_A-1794641	85875697	0.15
13213	SNP_A-1815281	96003236	0.16
14400	SNP_A-2202441	106377385	0.17
15616	SNP_A-2309459	116424331	0.18
16808	SNP_A-2160092	125903635	0.19
18017	SNP_A-1850477	135835399	0.20
19202	SNP_A-2289125	146389095	0.21
20407	SNP_A-1829559	154045512	0.22
21607	SNP_A-2208065	161575001	0.23
22999	SNP_A-1786242	169452387	0.24

We selected 20 SNPs to be causal which were approximately equally spaced along

the chromosome. These selected SNPs were all common (minimum allelic frequency larger than 0.3) and they had pairwise correlation smaller than 0.2. Using this set of SNPs $X_j, j = 1, \ldots, 20$ we simulated 100 different data sets of quantitative traits according to (1), where effect sizes β_j were equally spaced between 0.05 and 0.24. This choice of effect sizes covers the range of very small effect sizes which are in practice not detectable, to very large effect sizes which are detected very easily. The most interesting effect sizes are lying in the middle, where causal SNPs can be detected, but not too easily. Table 1 shows the details on the causal SNPs and the corresponding effect sizes.

When summarizing the results of our analysis we count as True Positive if the algorithm either detects exactly a causal SNP, or a SNP which is strongly correlated with a causal SNP (i.e. if $|R| > 0.5$). If the algorithm detects many SNPs that are correlated with the same causal SNP, only one of them is regarded as a True Positive, while the others are counted as False Positives.

4.2. Preliminary factors

In this section we will study two different preliminary factors which influence the performance of EA-MOSGWA: Clustering of correlated SNPs and the number of iterations I_S which determines the stopping criterion.

EA-MOSGWA uses the mBIC2 criterion to evaluate models, which depends on the total number of available SNPs p. If there is a large number of correlated SNPs than a penalty based on p might be too strict, and as described in [4] it is common practice to work with an 'efficient' number of SNPs. Such an efficient number can be obtained for example by clustering. Similarly as in [4] we compute clusters with the algorithm described in [2], and then replace the value of p in (2) by the total number of clusters, p_C. With the resulting milder criterion EA-MOSGWA will typically yield larger models.

The second factor we consider in this subsection is the number of iterations I_S without any update after which the algorithm terminates. We will compare $I_S = 2000$ with $I_S = 4000$.

Table 2 summarizes the main results of this subsection. We estimate the power as the average number of True Positives divided by 20, the number of causal SNPs. As usual the false discovery rate is estimates as the average of $\frac{\#FP}{\#TP+\#FP}$, where this ratio is set to zero in case of no detections.

The first observation is that the results do not change much whether we use $I_S = 2000$ or $I_S = 4000$ for the stopping criterion. The estimated Power, FDR and number of false positives are almost identical. The average value of mBIC2 also drops only rather insignificantly when using $I_S = 4000$, whereas the runtime becomes much larger. We conclude that a choice of $I_S = 2000$ for the maximum number of iterations without a model improvement is sufficient to ensure that EA-MOSGWA converges.

The second comparison is concerned with the influence of using mBIC2 with the

Table 2: Results for the first two preliminary factors. The first column (Nr) specifies whether the total number of SNPs p or the efficient number of SNPs p_C was used to compute mBIC2. The second column (I_S) refers to the maximum number of iterations without a model improvement that was allowed before stopping. The third and fourth column contain the average values of the runtime and the criterion mBIC2 taken over 100 simulation runs. The remaining columns contain the average of Power, FDR, and number of False Positive detections (FP)

Nr	I_S	Time	mBIC2	Pow	FDR	FP
p	2000	00:40:09	34195,8	0,58	0,015	0,19
p	4000	01:14:53	34195,1	0,58	0,015	0,19
p_C	2000	00:49:53	34166,1	0,63	0,035	0,47
p_C	4000	01:35:06	34165,2	0,63	0,037	0,51

effective number p_C instead of p. Table 2 indicates, as expected, that working with p_C yields larger power, but at the same time also larger Type I error rate. In general an FDR between 3% and 4% seems to be quite acceptable, and therefore we will present in the next section only results obtained with the effective number p_C.

Fig. 1: Frequency of detection of causal SNPs with clustering (mBIC2 with p_C) and without clustering (mBIC2 with p), where $I_S = 2000$.

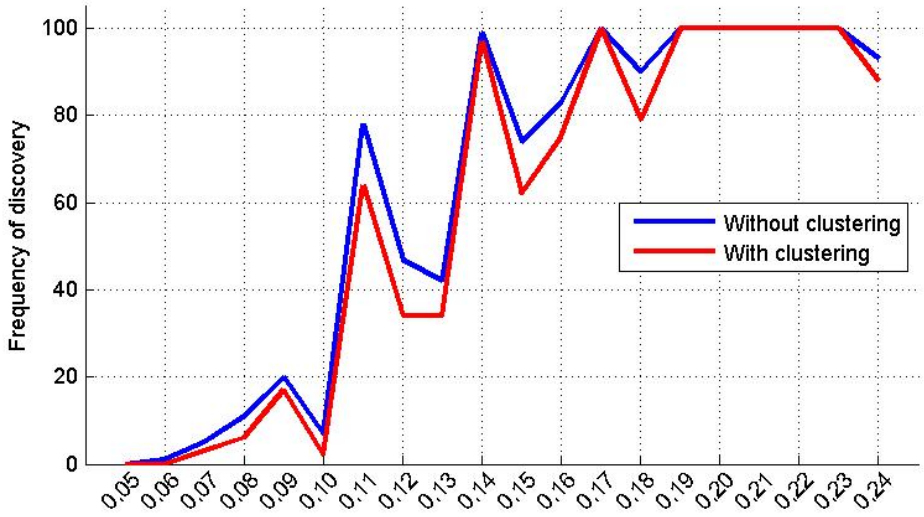

Figure 1 shows the power to detect each causal SNP, where again one can see that working with the efficient number p_C uniformly improves the power compared to working with the total number of SNPs p in the criterion. Figure 1 also illustrates that our

simulation study was designed such that the power to detect SNPs with the smallest effect size is close to zero, whereas SNPs with the largest effect size are detected almost always. The trend is not completely linear, because the power to detect a SNP will also depend on its minimum allelic frequency.

4.3. Influence of population size and update parameter B

In this subsection we will discuss the influence of the population size u and the two parameters B and I_S which are involved in defining the stopping criterion. We consider population sizes $u \in \{10, 20, 60\}$, and for the largest population size we consider various values of B. As previously in the last section we additionally vary the maximal number of iterations without an update between $I_S = 2000$ and $I_S = 4000$. The results for these different parameter settings are presented in Table 3.

Table 3: Dependence of the performance of EA-MOSGWA on the population size u, and the two parameters B and I_S which determine the stopping criterion. As in Table 2 we report the average value of the runtime, mBIC2, Power, FDR, and the number of false positive detections.

u	B	I_S	Time	mBIC2	Pow	FDR	FP
10	10	2000	00:49:53	34166,1	0,625	0,035	0,47
10	10	4000	01:35:06	34165,2	0,625	0,037	0,51
20	10	2000	00:50:13	34164,8	0,626	0,037	0,50
20	10	4000	01:32:49	34166,3	0,622	0,037	0,50
60	5	2000	00:37:18	34164,7	0,628	0,035	0,48
60	5	4000	01:06:19	34164,5	0,629	0,036	0,50
60	10	2000	01:00:54	34164,9	0,626	0,034	0,47
60	10	4000	01:39:14	34164,5	0,628	0,037	0,51
60	30	2000	02:10:38	34164,5	0,630	0,037	0,51
60	30	4000	03:21:25	34164,8	0,629	0,039	0,54
60	60	2000	03:28:34	34164,6	0,628	0,037	0,51
60	60	4000	05:31:11	34164,9	0,629	0,036	0,50

The first observation is that increasing the population size while keeping $B = 10$ fixed allows for obtaining lower values of mBIC2, which goes along with a rather small increase in power, whereas the effect on the false discovery rate is not entirely clear. At the same time the runtime increases only quite moderately.

Changing the value of B which determines the stopping criterion has a huge effect on the runtime, but a rather negligible influence on the performance of our algorithm, which suggests that a small value of B can be recommended. Similarly it makes hardly

any difference whether we work with $I_S = 2000$ or $I_S = 4000$, and therefore like in the last section we can conclude that $I_S = 2000$ is sufficiently large.

4.4. Comparison of EA-MOSGWA with original Stepwise procedure

In Table 4 we report the results illustrating the performance of the deterministic Stepwise procedure implemented in MOSGWA, once again with and without clustering. This procedure is described in full detail in [4].

Table 4: Results for the Stepwise method. The first column (Nr) specifies whether the total number of SNPs p or the efficient number of SNPs p_C was used to compute mBIC2. The second and third column contain the average values of the runtime and the criterion mBIC2 taken over 100 simulation runs. The remaining columns contain the average of Power, FDR, number of False Positive detections (FP), respectively.

Nr	Time	mBIC2	Pow	FDR	FP
p	00:04:15	34197,1	0,60	0,017	0,22
p_C	00:03:46	34168,7	0,63	0,041	0,58

Comparing the results reported in Tables 2 and 4 we conclude that EA-MOSGWA yields (on average) smaller values of mBIC2 than the stepwise procedure. Interestingly, in our simulation study the stepwise procedure has at least in case of working with p a slightly larger power of detecting causal SNPs than EA-MOSGWA, which is however counterbalanced by a slight increase of FDR.

5. Conclusions

The results of our simulation study confirm the ability of EA-MOSGWA to converge, even when the population size is rather small, and that EA-MOSGWA is fairly robust with respect to the choice of different tuning parameters. The simulations also confirm the effectiveness of the modification of mBIC2, with the penalty dependent on the number of clusters rather than on the total number of SNPs.

Due to a very good performance of the Stepwise method in our future work we plan to include the solution provided by this method into the initial population for EA. We also plan to consider the whole information collected by EA for the quantification of the statistical uncertainty related to the choice of the "best model". This task can be accomplished by the comparison of values of mBIC2 for different models visited by EA in the process of the search for the global minimum.

Acknowledgments

This research has been funded by the Vienna Science and Technology Fund (WWTF) through project MA09-007a.

The data set used for simulations in this manuscript was obtained from dbGaP through dbGaP accession number phs000145.v1.p1 at
www.ncbi.nlm.nih.gov/projects/gap/cgibin/study.cgi?study_id=phs000145.v1.p1..

Computer simulations were performed at the Computer Lab, which was established in the research project: DS/IMI/106/2012.

References

[1] F Begum, D Ghosh, G. C Tseng, and E Feingold. Comprehensive literature review and statistical considerations for gwas meta-analysis. *Nucleic Acids Res.*, 40(9):3777–3784, 2012.

[2] F Frommlet. Tag snp selection based on clustering according to dominant sets found using replicator dynamics. *Adv. in Data Anal. and Classif.*, 4:65–83, 2010.

[3] F Frommlet, I Ljubic, H Arnardottir, and M Bogdan. Qtl mapping using a memetic algorithm with modifications of bic as fitness function. *Statistical Applications in Genetics and Molecular Biology*, 11(4):Article 2, 2012.

[4] F Frommlet, F Ruhaltinger, P Twaróg, and M Bogdan. Modified versions of Bayesian information criterion for genome-wide association studies. *CSDA*, 56:1038–1051, 2012.

[5] D. E Goldberg. *Algorytmy genetyczne i ich zastosowania*. Wydawnictwa Naukowo - Techniczne, Warszawa, 2003.

[6] Y Guan and M Stephens. Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Ann. Appl. Stat.*, 5:1780–1815, 2011.

[7] Q He and D Lin. A variable selection method for genome-wide association studies. *Bioinformatics*, 27:1–8, 2011.

[8] Zb Michalewicz. *Algorytmy genetyczne + struktury danych = programy ewolucyjne*. Wydawnictwa Naukowo - Techniczne, Warszawa, 2004.

[9] M R Nelson, K Bryc, K S King, A Indap, A R Boyko, J Novembre, L P Briley, Y Maruyama, G Waterworth, D M amd Waeber, P Vollenweider, J R Oksenberg, S L Hauser, H A Stirnadel, J S Kooner, J C Chambers, B Jones, V Mooser, C D Bustamante, A D Roses, D K Burns, M G Ehm, and E H Lai. The population reference sample, popres: a resource for population, disease, and pharmacological genetics research. *Am J Hum Genet.*, 83:347–358, 2008.

[10] T T Wu, Y F Chen, T Hastie, E Sobel, and K Lange. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25:714–721, 2011.

EA-MOSGWA – narzędzie do identyfikacji przyczynowych SNPów w badaniach asocjacyjnych całego genomu

Streszczenie

W artykule przedstawiony jest aktualny stan rozwoju programu EA-MOSGWA – narzędzia służącego do identyfikacji przyczynowych genów w badaniach asocjacyjnych całego genomu (ang. *Genome Wide Association Studies*, GWAS). Głównym celem tych badań jest określenie tych rejonów chromosomu, które są związane z występowaniem chorób genetycznych (np. cukrzyca, rak) lub wpływają na daną cechę (np. wysokość lub ciśnienie krwi). Sprowadzają się one do przebadania wielu tysięcy polimorfizmów pojedynczego nukleotydu (ang. *Single Nucleotide Polymorphism*, SNP) i powiązaniu ich (pojedynczych lub grupy SNPów) z przypadkami klinicznymi oraz możliwymi do zmierzenia cechami. Kluczową kwestią jest zidentyfikowanie jak największej liczby przyczynowych SNPów przy jednoczesnej minimalizacji fałszywych odkryć.

Program MOSGWA umożliwia detekcje SNPów poprzez wybór zmiennych z użyciem kryterium mBIC2 – zmodyfikowanej wersji Bayesowkiego kryterium informacyjnego Schwarza. MOSGWA stara się zminimalizować mBIC2 przy pomocy metody selekcji Stepwise, podczas gdy EA-MOSGWA wykorzystuje w tym celu zmodyfikowaną wersję algorytmu ewolucyjnego.

W artykule prezentujemy wyniki szeroko zakrojonych badań symulacyjnych, w których możemy porównać wydajność EA-MOSGWA przy użyciu różnych ustawień parametrów. Również bierzemy pod uwagę klasteryzację SNPów, aby złagodzić korekcje wielokrotnego testowania w metodzie mBIC2. Przedstawiamy także porównanie wyników otrzymanych przez EA-MOSGWA z wynikami metody Stepsiwe używanej w programie MOSGWA, aby pokazać że proponowana metoda ma dobre właściwości minimalizacji kryterium mBIC2 oraz minimalizacji wskaźnika fałszywych detekcji.