Epigenetic modification of genetic algorithm

Kornel Chrominski[1][0000−0001−9875−5812], Magdalena
Tkacz[1][0000−0002−3931−0474], and Mariusz Boryczka[1][0000−0002−7994−8341]

Institute of Computer Science, University of Silesia in Katowice, Bedzinska 39, 40-080
Sosnowiec, Poland
kornel.chrominski@us.edu.pl
magdalena.tkacz@us.edu.pl
mariusz.boryczkaz@us.edu.pl

Abstract. The article presents a new operation in the genetic algorithm. This operation mimics the epigenetic process of cytosine methylation. The Epigenetic processes have a huge impact on the functioning of living organisms, but have not yet been reflected in the operations of genetic algorithms. In a study on the evaluation of the operation mimics epigenetics process were used genetic algorithm for Knapsack issue.

Keywords: genetic algorithms · epigenetics · Knapsack problem.

1 Introduction

For many years man has been trying to imitate nature and use solutions seen in nature. Inspiration processes occurring in nature have also been reflected in computer science. Based on nature observation, algorithms have been developed that use operations that mimic the processes that occur in nature. Algorithms inspired by nature have found a special application in the case of optimization problems that classic deterministic algorithms cannot cope with. Optimization problems are a group of problems for which we are able to define a certain function for which iterative we choose the appropriate values so as to obtain its optimal value. A special type of algorithms inspired by nature - genetic algorithms are the subject of research in this article. They belong to the group of evolutionary algorithms. The idea of genetic algorithms is based on the theory of evolution and processes occurring in the cells of every living organism. In recent years there has been a significant progress in science in the field of biology and molecular genetics, many new mechanisms have been discovered in the process of inheritance and responsible for the evolution of organisms. These mechanisms are not closely related to direct DNA sequence modifications and have been called epigenetic processes. Epigenetic processes have a significant impact on inheritance and are responsible for the influence of external factors on the functioning of living organisms. Classic genetic algorithms were developed in the 1960s, so it was not possible to include all currently known processes regarding inheritance and variability of species. It is these relatively new discoveries in the field of molecular genetics (which are epigenetic processes) that have

become the starting point for the research presented in this article. The starting point for the introduction of new operations in the genetic algorithm (modeled on epigenetic processes) was that since they are of great importance in nature when adjusting living organisms to environmental conditions, it is possible that their proper mapping in genetic algorithms will improve their efficiency.

2   Background and motivation

Epigenetic processes occurring in the living organism have a huge impact on their functioning. Genetic algorithms mimics the operations of living organisms, but epigenetic operations have not yet been implemented in genetic algorithm. The use of operations mimic epigenetic processes is aimed at improving the efficiency of genetic algorithms by reducing the number of new generations being created, which influence the time need to find solution by the genetic algorithm. This article presents the results of the appropriate probability of occurrence the modification mimicking the cytosine methylation process, as well as the results of studies on the impact on the reduction of the number of generations. The time of operation of the algorithm with and without operation mimics epigenetic process was also presented. The research results presented in the article are part of the research on the possibility of using processes that mimic epigenetic processes in genetic algorithms.

2.1   Genetic algorithms

Genetic algorithms [1–5] have found application in solving optimization problems in cases where deterministic approaches did not work well. The idea of a genetic algorithm, as an algorithm inspired by the theory of evolution, was presented by John Holland in the 1960s, and then developed by David E. Goldberg in the 1980s. In genetic algorithms, the terminology was taken from biological sciences and adapted for the needs of genetic algorithms. The pseudo-code [6–8] of the classical genetic algorithm was presented as an Algorithm 1.

The first necessary step in genetic algorithms is choosing the right coding method for the individual - this has a significant impact on whether the algorithm will get the expected result. The next step in constructing the task for the genetic algorithm is to properly construct the function of assessing individuals (fitness function). The fitness function return a value that shows how good the solution is. The value of the fitness function [9–12] is calculated for all individuals in a given population, and its value affects the likelihood of an individual's participation in reproduction – creating a new population. Operations imitating biological processes are carried out on the created population of possible solutions. These operations include individual selection, crossover, and mutation of individuals. In addition to standard operations in the genetic algorithm, there may also be additional operations or modifications of standard operations.

Algorithm 1: Genetic algorithm

Data: population of individuals;
Result: Best individuals;
1 begin
2 | t = 0;
3 | Create an initial population– $Pop(0)$;
4 | Evaluate individuals – calculate the value of the fitness function for each individual in the population P_0;
5 | **do**
6 | | select individuals for the new population $Pop(t)$ – selection;
7 | | perform crossover operation;
8 | | perform the mutation of individuals;
9 | | evaluate individuals;
10 | | replace old population a new one;
11 | | t = t + 1;
12 | **while** *stop condition reached*;
13 end

2.2 Epigenetics

Epigenetics [14–16] is a science that studies the processes of extragenic inheritance, as well as the impact of external factors on the level of gene expression. Gene expression determines the phenotypic characteristics of an individual, i.e. its adaptation to the environment, and behavior, appearance, etc. For a long time, scientists have been wondering, for example, why there is a difference in the appearance and behavior of identical twins, or why cloned individuals, despite identical sequences genes, however, have a different coloration, exhibited a different behavior. Some of these puzzles could be explained by discovering the mechanisms that are the subject of epigenetics research.

Concepts related to epigenetics emerged at the time of the discovery that some changes in the genotype of living organisms are not directly related to the structure of DNA, its changes and inheritance processes. They began to wonder what could cause these changes. The result of the research was the discovery of numerous molecules that affect the processes that occur in living organisms, affecting how the genetic code will be read. In other words - how the individual's phenotype will change without changing its DNA.

It turned out that epigenetic processes play a significant role in differentiating the population and adapting to new conditions. Epigenetic modifications are also a source of some diseases, and also affect individual traits such as, for example, perception of the world or personality traits. It can be stated that the genotype of living organisms is a place of storage of relatively static genetic information, and epigenetic processes are specific dynamic controllers (inhibitors or catalysts) responsible for activating certain information.

Epigenetic processes, through biochemical changes, affect the structure of DNA in living organisms, as well as the level of gene expression. Launching

a given epigenetic process may occur as a result of changes occurring in the body (also randomly), as well as as a result of external (environmental) factors. Current research indicates that the environment in which the organism resides, components of the diet, as well as chemical compounds can be impulses to launch a given epigenetic process in the body. Scientists estimate that a thorough understanding of epigenetic processes will allow the development of effective gene therapies for many diseases. The main modifications occurring in our epigenetic genotype include:

- Modifications resulting from the interaction of free particles – in each cell, apart from the DNA strand, they find numerous short fragments resembling the structure of the DNA strand, which can attach to DNA and modify the protein encoded by a given fragment of the genotype (e.g. inheritance with prion).
- modifications at the level of DNA strand - most often they lead to silencing the expression of a particular gene, i.e. reducing its expression level (e.g. cytosine methylation process);
- modifications at the histone level - histones are small proteins around which the DNA strand is wrapped; their modification may affect the increase or decrease in the synthesis of proteins encoded in a given DNA segment (e.g. deacetylation, allelic exclusion);

Due to the fact that epigenetics is a relatively young field of science, and epigenetic processes themselves have been discovered relatively recently, they have not yet been reflected in evolutionary algorithms. Therefore, the author decided to check whether, as in the case of living organisms (where epigenetics affects faster differentiation of the population and its better adaptation to environmental conditions), also in the case of genetic algorithms, the use of operations imitating epigenetic processes will improve the operation of these algorithms (by faster finding the optimal result by a given genetic algorithm).

3   Genetic algorithm with a modification that mimics cytosine methylation

This section presents the proposed modification that mimics the epigenetic process of cytosine methylation. In the first subsection there is information about the genetic algorithm for solving the Knapsack problem, this algorithm was used as the basis for placing operation mimics the epigenetic process. The next part provides information on the biological basis of modification, followed by the idea of modification and how to implement it in a genetic algorithm.

3.1   Genetic algorithm for solving the Knapsack problem

The formal definition of the Knapsack problem [17, 18] can be formulated as follows:
a backpack with the maximum load capacity V_K is available, and a set of N

elements $\{x_1, ..., x_j, ..., x_N\}$, with what each element of the set has a specific weight of m_j, and the value of c_j, $j \in \mathbb{N}$.

maximize:

$$\sum_{j=1}^{N} c_j x_j \tag{1}$$

Assuming:

$$\sum_{j=1}^{N} m_j x_j \leq V_K, \qquad x_j = 0 \ or \ 1, \ j = 1, ..., N \tag{2}$$

The genetic algorithm used for Knapsack is based on a classic genetic algorithm. Contains standard operations for genetic algorithms such as crossing, mutation, and selection of individuals. Standard operations used in the genetic algorithm for solving Knapsack problem:

1. **Coding of individuals**: The algorithm uses binary coding of individuals. A value of 1 corresponds to the information that the item should be loaded into the backpack, a value of 0 means that the item should not be loaded into the backpack.
2. **fitness function**: The algorithm uses the fitness function calculated on the basis of three parameters:
   - weight of the item;
   - item value;
   - maximum capacity of the backpack.

   The best adapted individual is the one for which the most valuable items will be loaded, without exceeding the maximum capacity of the backpack.
3. **Individual selection**: The algorithm used is a tournament selection method.
4. **Crossing individuals**: One-point crossing operations were used in the algorithm implementation.
5. **Mutation of individuals**: point mutations were used to change individual genes in the genotype of the individual.

3.2   Biological basis of the cytosine methylation process

The operation mimics epigenetic processes proposed in this article is based on the epigenetic process of cytosine methylation. Cytosine methylation is a process involving the attachment of methyl groups (-CH3) to the nitrogenous bases of nucleotides (the basic building block of DNA and RNA nucleic acids). The attachment of a methyl group to nucleotides reduces the expression level of genes encoded by a given DNA fragment. If a large DNA strand is methylated, the sequence fragment may be blocked so that the gene cannot be read. Cytosine methylation may also affect the transfer of genetic information by preventing the transfer of a specific gene in the process of inheritance. This process in living organisms also affects the formation of tissues (cell specialization). It is also extremely important in the overall functioning of the body, because only a

small number of genes have a constant level of expression (so called housekeeping genes responsible for basic life functions). Other genes, as a result of various biological processes, have variable levels of expression, which affects the proper functioning of the body. Cytosine methylation can also help the body get rid of the external DNA code (e.g., from viruses), which protects the body against this code. Disturbance of the methylation process may lead to the development of, for example, cancer or other genetic diseases (e.g. Angelman, Prader-Willi, Beckwith-Wiedeman syndromes).

3.3 Metylation of Cytosine in genetic algorithm

The proposed modification simulated the process of blocking a fragment of a subject's genotype sequence. The epigenetic modification was implemented in the crossover of individuals, modifying this operation. The progeny contain genotype fragments of each parent. If a process simulating cytosine methylation occurs in the crossing operation, a particular fragment of the individual's genotype is blocked, so that he does not participate in the crossing. This is equivalent to the fact that the genome fragment undergoing cytosine methylation will not be passed on to the descendants, i.e. it will not appear in the new population.
Figure 1 shows a diagram of the process of cytosine methylation in the process of crossing individuals. The dark gray fragment of the individual's genotype has undergone the methylation process, i.e. it does not participate in the crossing process.
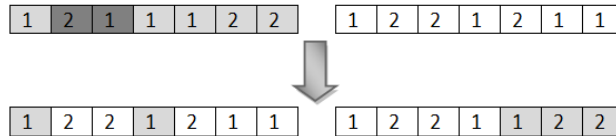

Fig. 1: Epigenetic operation imitating the process of cytosine methylation

In order to prevent the loss of the best solutions in the process of crossing, the process of blocking the sequence occurs in individuals with the lowest values of the fitness function, which aims to transfer to the descendants a larger fragment of the better-adapted individual's genotype. This means that both progeny receive only a specific genotype fragment from the parent having a better fitness function value. Thanks to this, the gene sequence with the poorer fitness function (from the less adapted individual) is "silenced".
Algorithm 2 represents the pseudo-code of the process simulating the process of cytosine methylation. The proposed epigenetic modification occurs in the process of crossing with an appropriate probability, the optimal value of which has been determined empirically.

Algorithm 2: Modified crossover algorithm including simulations of
the cytosine methylation process

   Data: individuals population, methylation probability (P_m), individuals
         fitness function F_p
1 select with probability P_m a pair of individuals to undergo the cross-over
   operation with the Epi modification;
2 foreach *selected pair of individuals* **do**
3   generate a length and location of the blocked part of genotype;
4   **if** $F_p(individual1) > F_p(individual2)$ **then**
5   block a part of genotype in individual2;
6   **end**
7   **else**
8   block a part of genotype in individual1;
9   **end**
10   perform the crossover operation without the blocked part of genotype;
11 end

The presented algorithm of epigenetics modification (Algorithm 2) shows the modified crossover operation at the time of modification, the rest of the algorithm operation corresponds to basic genetic algorithm. Based on the Algorithm 2, a certain group of individuals selected from the entire population undergoes a modified crossing operation, the remaining individuals undergo a standard crossing operation. In the first step, the length of the genotype to be blocked as part of epigenetic modification and the location of the blocking is randomly generated. In the next step, the values of the fitness function of individuals involved in crossing are compared. In an individual with a lower value of the fitness function, transmission of the fragment of the genotype to the descendants, whose location and length was determined in an earlier step, is blocked. Crosses are performed on individuals without the blocked genotype fragment of one of the parents.

4   Experiments

The proposed modification based on the cytosine methylation process has been tested in terms of selecting the optimal probability of occurrence of operations in the genetic algorithm.The optimal probability is the one for which it obtains the biggest reduction the number of generations. In the next step, it was checked how the introduction of the operation (with optimal probability) mimicking the process of cytosine methylation into the genetic algorithm affects the number of algorithm generations. In the last step, the operation time of the algorithms was compared with and without epigenetic modification.

Five data sets were used in the experiments, differing in the possible number of items to be packed and the maximum permissible capacity of the backpack. Data sets are artificial collections created for the purposes of experiments. The

description of the data sets for the load optimization algorithm is presented in Table 1.

Table 1: Datasets used in experiments

dataset	number of items to pack	max capacity of backpack
dataset 1	10	20
dataset 2	20	40
dataset 3	40	80
dataset 4	80	160
dataset 5	160	320

Parameters such as crossing probability, mutation and stop condition have been established for the genetic algorithm used in the study. The parameters set for the load optimization algorithm are standard values recommended in the literature and are presented in the Table 2.

The set parameters for GA for loading optimization were used both in the case of the base algorithm and algorithms modified for operations imitating epigenetic processes.

Table 2: Parameters of the genetic algorithm used in experiments

Parameters	Value
crossover probability	90%
probability of mutation	1%
individual selection method	tournament
stop condition	determined by the length of the set number generation without improving the value of the fitness function

In the first step, it was checked whether the occurrence of epigenetic operations affects the number of iterations needed to obtain the optimal result. The number of generations needed to obtain the best solution depending on the probability of occurrence of the proposed operation for individual test sets is presented in the Table 3. The results presented in Table 3 are the average value of the number of generations needed to obtain the expected result for 100 repeats of each algorithm.

Based on the table 3, it can be seen that the smallest number of generations needed to obtain the best solution for most test sets occurs in the case of 40% probability of epigenetic operation. The exception is test set 3, in which the minimum number of generations was reached in the case of a modification probability of 30%.

Table 3: Number of generations needed to obtain the best result for individual in dataset depending on the probability of epigenetics operations (The smallest number of iterations for a given test set is marked in bold italics)

probability of modification occurring	the number of generations needed to find the best solution				
	dataset 1	dataset 2	dataset 3	dataset 4	dataset 5
0% (without modification)	154	631	973	1355	1513
5%	178	624	968	1454	1574
10%	172	614	972	1286	1530
20%	144	578	690	1103	1588
30%	126	430	*641*	1186	1591
40%	*113*	*406*	644	*889*	*1404*
50%	131	419	658	1016	1459
60%	148	425	741	1021	1589
70%	152	437	951	1165	1598
80%	159	410	963	1302	1581
90%	166	623	967	1398	1598
100%	185	645	980	1465	1597

At the Figure 2 are presented the results of the comparison of changes in the fitness function for the genetic algorithm with epigenetic modification (with the optimal probability of modification occurring), and the algorithm without modification for individual test sets. From the graphs it is possible to read how quickly the value of the fitness function of the best individual increases with next generations of algorithms.

Based on the graphs in Figure 2, it can be seen that in the case of the genetic algorithm with the epigenetic modification, the value of the best individual fitness function increased faster compared to the algorithm without modification. The largest differences in the increase in the value of the fitness function for genetic algorithm with epigenetic operation compared to basic genetic algorithm can be seen for the second and the fifth data sets (Figures 2b and 2e). However, the smallest differences in the change in the value of the fitness function can be observed in the case of set 4 (Figure 2d).

The Table 4 shows the average running time (for 100 repeats) of the algorithm with and without modification. The table also shows the standard deviation value and the percentage value of the algorithm reduction time after application the operation mimicking the cytosine methylation process.

Although the introduced epigenetic modification is an additional operation in the genetic algorithm, a reduction in the total time needed to obtain the expected result by the genetic algorithm has been observed. For most data sets, the time reduction was above 25%. The impact on the reduction of the algorithm's operation time with modification was the reduction of the number of generations in the algorithm.

(a) dataset 1

(b) dataset 2

(c) dataset 3

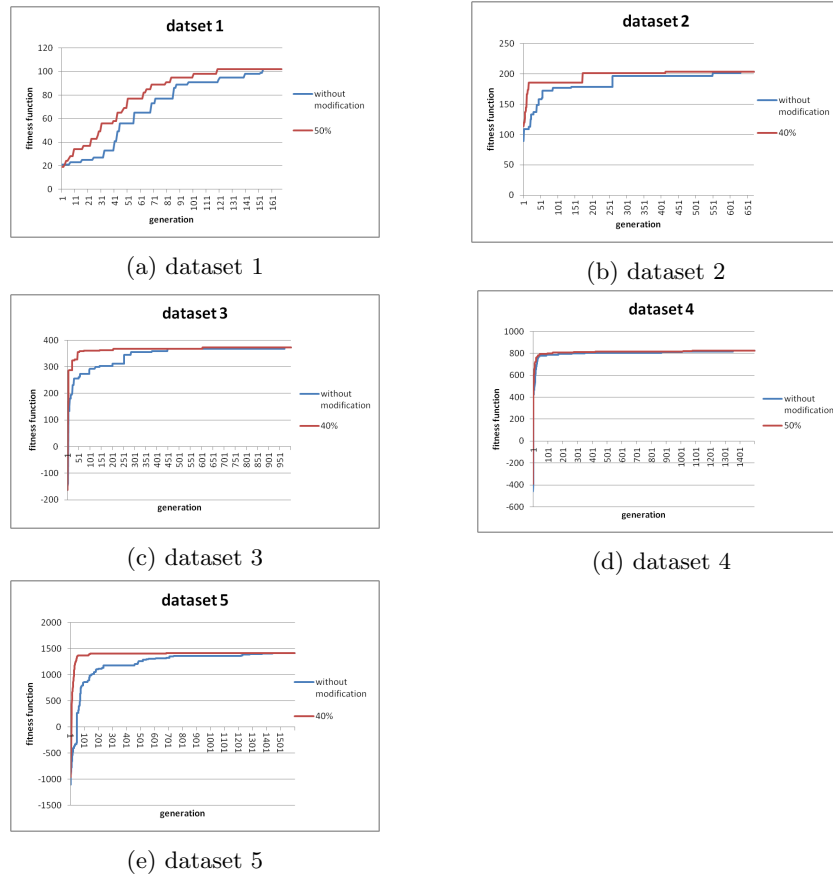(d) dataset 4

(e) dataset 5

Fig. 2: Changes in the value of the fitness function for generation in the case of an algorithm with and without epigenetic modification

Table 4: Comparison of one generation times for a genetic algorithm with and without epigenetic operation

dataset	time [s]				change %
	Epi_GA		GA		
	mean	sd	mean	sd	
dataset 1	7.230	0.68	9.70	1.850	25.46%
dataset 2	32.480	4.470	45.430	6.310	28.51%
dataset 3	70.510	9.62	95.350	20.430	26.05%
dataset 4	160.909	14.224	228.995	29.810	29.73%
dataset 5	475.956	161.460	481.134	173.995	1.080%

The Table 5 shows the T test value for the number of generations for the algorithm with and without epigenetic modification.

Table 5: Statistics for two sample t-test - number of generations in genetic algorithm with modifications and without.

Statistics for two sample t test	
t	-3.0754
df	4
p-value	0.0185

Based on the Table 5, it can be confirmed that the difference in the number of generations for the algorithm with epigenetic modification compared to the algorithm without modification is statistically significant - a p-value below 0.05.

This section presents the impact of a modification that mimics the epigenetic process of cytosine methylation on the operation of the genetic algorithm.

5   Conclusion and future works

The article presents the modification of the genetic algorithm. The modification mimics the epigenetic process of cytosine methylation. Based on the conducted research, it was shown that the application of the proposed modification improves the efficiency of the genetic algorithm. The efficiency is improved by reducing the number of generations in the genetic algorithm, and thus reducing the duration of the genetic algorithm. The presented results are part of the research on the possibility of using epigenetic processes in genetic algorithms. As part of our research, it is planned to examine the assembly of other modifications that mimic other epigenetic processes in various genetic algorithms. Research is also being carried out into the development of an epigenetic algorithm.

References

1. H. L. Abdel-Magid and M. M. Dawoud. Optimal agc tuning with genetic algorithms. Electric Power Systems Research, 38(3):231–238, 1996
2. J. Arabas. Wyklady z algorytmow ewolucyjnych. WNT, 2004 (in polish)
3. D. E. Goldberg. Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Professional, 1989
4. D. E. Goldberg and K. Deb. A comparative analysis of selection schemes used in genetic algorithms. In Foundations of Genetic Algorithms, pages 69–93, Elsevier, 1991
5. W. W. Hager. Computational optimization and applications, 2017
6. R.N. Greenwell, J.E. Angus, and M. Finck. Optimal mutation probability for genetical gorithms. Mathematical and Computer Modelling, 21(8):1–11, 1995
7. O. Abdoun, and J. Abouchabaka. A comparative study of adaptive crossover operators for genetic algorithms to resolve the traveling salesman problem. International Journal of Computer Applications, 31(11), 2012
8. A.E. Eiben, and S.K. Smit. Evolutionary algorithm parameters and methods to tune them. Autonomous Search, pages 25–38, Springer, 2012

9. E. Hopper and B. Turton. Application of genetic algorithms to packing problems. A Review. InSoft Computing in Engineering Design and Manufacturing, pages 279–288, Springer London, 1998

10. A. Hussain, Y. Muhammad, and A. Nawaz. Optimization through genetic algorithm with a new and efficient crossover operator. International Journal of Advances in Mathematics, 2018(1):1–14, 2018

11. K. Jebari. Selection methods for genetic algorithms. International Journal of Emerging Sciences, 3:333–344, 12 2013

12. G. M. Khan. Evolutionary computation. In Evolution of Artificial Neural Development, pages 29–37, Springer, 2018

13. D. Moore and S. David. The Developing Genome: An Introduction to Behavioral Epigenetics. OXFORD UNIV PR, 2015

14. P. Weglenski. Genetyka molekularna. PWN Warszawa, 2006 (in polish)

15. C. Dupont, D. Armant, and C. Brenner. Epigenetics: Definition, mechanisms and clinical perspective. Seminars in Reproductive Medicine, 27(05):351–357, 2009

16. R. Al-Haddad et al., Epigenetic changes in diabetes. Neuroscience Letters, 625:64–69, 2016

17. X. Li, Z. Zhao, and K. Zhang. A genetic algorithm for the three-dimensional bin packing problem with heterogeneous bins. In Industrial and Systems Engineering Research, 2014

18. K. Man, K-S. Tang, and S. Kwong. Genetic Algorithms. Springer London, 2012