Contributions to Data Analysis and Processing in Genetic Analysis

Summary of thesis

Dr. Med. Ing. Nicolae Teodor MELIȚĂ

1. Introduction

1.1. Motivation

The significant evolution of technology during the recent decades has a major impact on the research in all disciplines. New advanced methods for data acquisition were developed for the most diverse spheres of theoretical and practical interest, and the data storage capacity increased exponentially. Efficient methods for the analysis, interpretation and integration of data are needed to exploit the new opportunities.

Methods belonging to Artificial Intelligence (AI) were incrementally required to analyze and understand the processes modeled in various fields of study. A domain with a major impact in medicine, which flourished dramatically in recent years, is Bioinformatics. Analysis of differentially expressed genes is one of the essential directions in Bioinformatics, and in this area, the usefulness of AI was certified by spectacular results with impact in medical clinical activity.

Evolutionary Algorithms (EA) are an important part of the direction of artificial intelligence and were often used to interpret data acquired to describe various models. Evolutionary algorithms are built on principles tested for billions of years.

The DNA microarray technology is a widely used method in Bioinformatics. The facile access to real data sets and the corresponding results obtained thru various methods by numerous researchers, provide a major opportunity to develop AI methods into a consolidated framework.

This thesis introduces principles modeled from the biological evolution with the finality to improve the performance and applicability of genetic algorithms (GA).

1.2. Objectives

In conducting the present thesis, I intended to model principles underlying biological evolution to improve the performance and applicability of genetic algorithms. Particularly, we intended to:

1) Model the Incomplete Dominance,

2)Evaluate the opportunity to represent the genotype by a variable number of chromosomes,

3) Model the Random Assortment of Chromosomes during Meiosis and to introduce of a new crossover operator,

4) Test new models and operators in the context of selecting attributes from the data acquired with the DNA microarray technology,

5) Create a software package, integrated in the R and Bioconductor, easily accessible by researchers in the field of DNA microarray data analysis,

6) Model the nonsense mutation,

7) Model the frameshift mutation,

8) Model the deletion,

9) Model the monosomy,

10) Model transposons.

2. Background

2.1. The DNA microarray technology

The DNA microarray technology represents a major progress in genetic analysis. The methodology has been used in numerous studies of genetics and molecular biology. Analysis of differentially expressed genes was probably the most widespread application of the technology, since it was introduced.

The DNA microarrays provide a snapshot of the conditions that describe an instance at a particular time and allow the evaluation of expression for thousands of probes immobilized on a single chip (Fig. 2.1). This approach brings valuable information towards:

1) the detection of valuable markers for early diagnosis,
2) the development of drugs with improved efficiency,
3) improved explanation of different stages in a certain pathology,
4) understanding the consequences of certain pathogens.

a) b)

Fig 2.1 – Example of DNA microarray chip. a)Stanford Technology; b)Affimetrix Technology.
(Source - Academic Dictionaries and Encyclopedias, www.enacademic.com, a public domain.)

The DNA microarray experiments are complex and require a careful implementation (Fig. 2.2). In general, several steps are essential for a successful experiment:

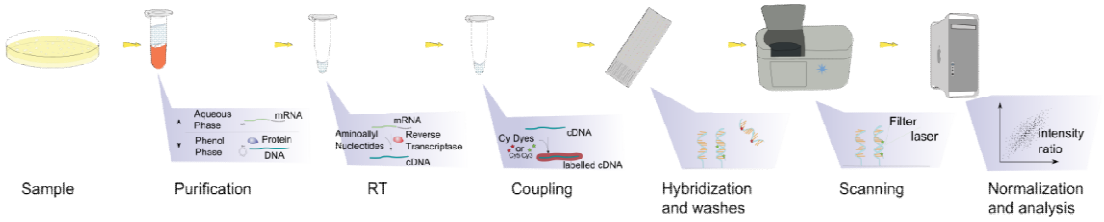


Fig. 2.2 – Stages in a typical DNA microarray experiment.
(Source – wikipedia.com, a public domain.)

To achieve a common objective for the microarray studies, the discovery of a group of genes that may be causally related to a specific pathology, the feature selection methods are paramount. These methods represent the focus of this paper and are addressed with the proposed method.

3. Improved Method for Selecting a Limited Number of Biologically Interpretable Attributes

Evolutionary Algorithms (EA) model principles learned from the biological evolution to address optimization problems. The principles of natural evolution are tested over billions of years of continuous adaptation to the environment. Genetic algorithms (GAs) are part of the EA and have been successfully applied in various optimization problems. Also, the genetic algorithms have become a widespread option for feature selection in different contexts.

Our goal is to select a limited number of biologically interpretable attributes from data acquired with the DNA microarray technology. We propose an improved method, based on GAs, designed to select attributes in a general framework, but optimized for applications with microarray data. We model phenomena underlying the biological evolution, aiming to improve the performance of GAs in this context.

The following approach, revolves around a diploid genetic algorithm, but benefits from original improvements modeled from the biological evolution:

1) an approach inspired from the incomplete dominance for genotype-phenotype mapping,
2) an operator fashioned from the random assortment of chromosomes during meiosis,
3) operators for mutation, inspired from the human genetics, designed to address the requirements of the DNA microarray studies.

3.1. Genetic Algorithms

The Genetic Algorithms have been theorized by John Henry Holland five decades ago. Holland also introduced the concept of schemata and schema theorem [1] to formalize the process of GA evolution. The simple GA approach has been extended in various ways, with diploid chromosome representation [2] and different new operators, many of which were modeled after the biological evolution [3].

The GAs nomenclature is borrowed from genetics to highlight their origins. The population is composed of individuals representing solutions. Individuals are encoded by a genotype that is apparent in the phenotype. The genotype is coding attributes as a string of genes. In the classical binary representation, the genes encode for attributes, have fixed positions in genotype, called loci, and have alleles with the values 0 and 1. The adaptability of a phenotype to the environment is appreciated by evaluating a fitness function.

A context in which genetic algorithms are expected to perform well [4] is outlined several criteria:

1) search space is vast and its configuration is not known,
2) fitness function is affected by noise,
3) finding a local optimum is satisfactory enough.

These terms describe the excellent conditions of the practice of selecting DNA microarray experiments attributes.

An AG perseveres in two activities: exploration and exploitation [4]. The algorithm explores new solutions to better adapt to the environment. Simultaneously, exploits the adaptations already acquired during the search. There must be a balance between these two activities, for a GA optimization to be successful.

The initial population of individuals is generally randomly generated from a discrete uniform distribution. Consequently, replications of a search with a genetic algorithm often converge towards different solutions. This aspect was addressed thru multiple replications of an experiment or deterministic alternatives to random initial population.

The **crossover** models a principle essential for the genetic diversity. During meiosis, an exchange of genetic information occurs between homologous chromosomes. Diachronically, different crossover operators have been proposed to support the evolution in genetic algorithms.

The **mutation** operators in genetic algorithms also model principles from biological evolution. For a change in genotype to be considered mutation, it must be transmitted to heirs. A mutation may transfer desirable characteristics to the next generation, which supports adaptation to the environment or undesirable properties, which on the contrary, alter the heir's fitness. As a result, the chance for a mutation to occur is generally much lower than the crossovers' rate in the genetic algorithms.

The principle of evolution thru selection also originates in a natural law [5]. Individuals that are better adapted to the environment have better chances to survive and consequently to transmit their genetic information. Thus, their genes will be better represented in subsequent generations.

3.2. Improved Method for Selecting a Limited Number of Biologically Interpretable Attributes

The finality of the proposed genetic algorithm is the feature selection from the DNA microarray data. The purpose of such an experiment is not finding a supervised classifier that can perfectly discriminate between two classes of examples. We intend to determine, from a large number of differentially expressed genes, a sub-group that can characterize the two classes of examples. The causal relationship between a condition and a subset of the genes cannot be determined directly by methods of artificial intelligence only. Further biological validation is mandatory for such a relationship to be established.

Supervised classifiers are employed to assess the fitness of the tested individuals. The accuracy of classifiers in discriminating between the ongoing classes in the data was used as fitness value.

Therefore, a genotype, sequence of values 0 and 1, with a length equal to the number of attributes in the data codes for a supervised classifier engaged in learning examples considering only the sub-group of attributes specified by the alleles with the value = 1.

The proposed algorithm is based on a **diploid** GA. Each individual has two haploid sets of chromosomes, therefore, two classifiers using the same learning technique, but considering different sub-groups of attributes.

A key aspect in designing a diploid genetic algorithm is the genotype to phenotype mapping. Generally, this problem was addressed by defining domination schemes, individualized for a specific optimization framework.

In this thesis we propose an original approach, inspired by biological evolution, for genotype to phenotype mapping in diploid GAs. Our proposal is modeled after the incomplete dominance and does not require defining a domination scheme.

The next step is the **condensation of genotypes in chromosomes**. The user can specify the desired number of chromosomes. The distribution of genes on chromosomes is not accomplished evenly. The DNA microarray technology evolution towards user customizable variants, in terms of samples immobilized on the chip, enables a superior approach in selecting attributes. Grouping genes with similar roles on the same chromosome would be a very desirable path, with potentially remarkable results.

In the next stage, the initial population is evaluated thru the accuracies of supervised classifiers in discriminating the classes present in data. The individuals stratified after performances are subsequently subjected to crossover operations between the two haploid sets of chromosomes in each.

The crossovers are performed by an original operator, which models phenomena underlying the biological evolution. The crossover operator implements the random assortment of chromosomes to genotypes a priori affected by two-point crossovers at this stage.

The method for selecting the sets of chromosomes for the next generation exploits the elitism, often implemented genetic algorithms. The sets of chromosomes that encoded the less competent classifier in each individual are removed. A number of haploid sets of chromosomes, which were more fitted in the current generation, are preserved for the subsequent iteration.

The haploid sets of chromosomes selected to be part of the next generation are subject to alteration by mutation, with an incidence specified initialization. During our research, we found that the classical mutation is inappropriate for selecting attributes from DNA microarray data. Consequently, we explored the opportunity for different mutation operators.

The next generation is then created by assembling individuals from haploid sets of chromosomes randomly paired. This new generation is then evaluated and analyzed during a new iteration. These steps are executed repeatedly, over a number of iterations specified at initialization, the termination condition for the algorithm.

A number of replications of the experiment are mandatory to address the stochastic component of the search and get meaningful results.

3.3. The Incomplete Dominance

3.3.1. The Incomplete Dominance in Biology

Every cell in the body, except gametes, includes in nucleus, two copies of each autosome. One copy of each autosome comes from the mother while the other one is inherited from the father. Somatic cells are diploid, have two copies of each autosome. Gametes are haploid and contain a single copy of each autosome. The two copies of each chromosome are called homologs. Each of homologous chromosomes is inherited from one parent and has, at the same loci, genes for the same treatments. Genes present at the same locus in homologous chromosomes are called alleles. Identical alleles are present in the homozygous homologous chromosomes for that locus. Homologous chromosomes that have different alleles are called heterozygous for the specified locus.

In the case of diploid organisms, the question arises how the different alleles present in heterozygous homologous chromosomes, find expression in the phenotype. In 1865, Mendel described a model in which one of the two alleles is expressed in the phenotype (dominant character) and the other one is masked (recessive character). In respect of nomenclature introduced by Mendel, this relationship is called dominance.

In the table below we present an imaginary case in which an organism inherits genes that determine its color them from the previous generation. There are two possible alleles **R** and **a**, for the gene that determines the color of the body. **R** is the dominant allele and determines the red color. Allele **a** is responsible for the blue body color. Table 3.1 shows the possible combinations and the effects on phenotype.

Tabel 3.1 – Complete Dominance

		Inherited from Father	
		R	a
Inherited from Mother	R	RR	Ra
	a	Ra	aa

Tabel 3.2 – Incomplete Dominance

		Inherited from Father	
		R	A
Inherited from Mother	R	RR	RA
	A	RA	AA

An alternative to this model is the incomplete dominance. In this type of interaction between alleles, often found in nature, no allele dominates over the other one. Heterozygous phenotype will be intermediate between the homozygous variants. The principle of incomplete dominance is shown in Table 3.2 for the same imaginary example. The phenotype of heterozygous RA individuals is purple, a combination of the effects induced by the two alleles.

3.3.2. The Incomplete Dominance in genetic algorithms

In designing a diploid GA, the genotype to phenotype mapping is complicated compared to the haploid genetic algorithms. A GA evaluates the fitness of the phenotype. The presence of different alleles, at some loci in an individual, requires further treatment.

The Incomplete Dominance promotes the idea of an intermediate phenotype between homozygous variants. In a diploid genetic algorithm, we can apply this concept as follows. Each set of chromosomes present in the individual can be considered with its particular effects. An individual with two sets of chromosomes has two supervised classifiers, accounting for different attributes, to fit in the same context. The fitness of such an individual is assessed as the average performance of the two classifiers.

3.4. The Random Assortment of Chromosomes

3.4.1. The Random Assortment of Chromosomes during Meiosis

The cells of an organism must divide for different objectives. Somatic cells must divide to regenerate various tissues and support many functions in the body. Germ cells divide to ensure the survival and evolution of the species.

In general, differentiated somatic cells divide for the purpose of producing new identical cells, able to support the same functions. In this context, the genetic material of the cell to divide must be kept intact and changes in the DNA are not desirable. This type of cell division called mitosis.

On the other hand, the germ cells divide in a fundamentally different manner. This process, called meiosis, aims to reproduce the individual and the conduct serves the purpose. The result of meiosis, cells called gametes, contain only one set of chromosomes in the nucleus. While control mechanisms protect the transmission of genetic information through meiosis, a degree of genetic diversity is allowed. The purpose is undoubtedly the evolution of the next generation in order to adapt superiorly to the environment. The phenomena that occur during the cell division are discussed in detail in [6]. One source of genetic diversity during meiosis, the crossing over, has been widely used in genetic algorithms and concretized in different crossover operators. A suggestive illustration of meiosis is shown in Fig. 3.1.

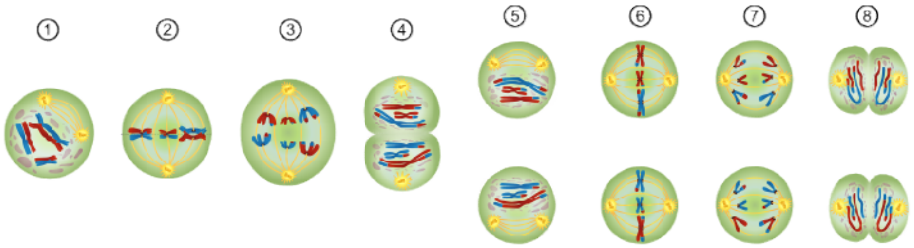


Fig. 3.1 – Meiosis stages (image source: wikipedia.com, rights of free use and modification.)

Another source of variability is the random assortment of chromosomes during meiosis. This source of genetic diversity is extremely important, and the amount of genetic information communicated during this process deserves exploited.

3.4.2. The Random Assortment of Chromosomes in GA

Genetic algorithms traditionally used crossover operators to provide support the evolution. The most popular proposals have been single- or two-point crossovers (Fig. 3.2 and Fig. 3.3). The second phenomenon that ensures genetic diversity, random assortment of chromosomes during meiosis has not been properly exploited in GAs.

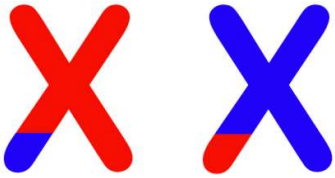


Fig. 3.2 – Single-point crossover.

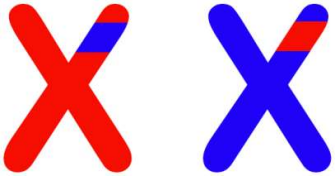
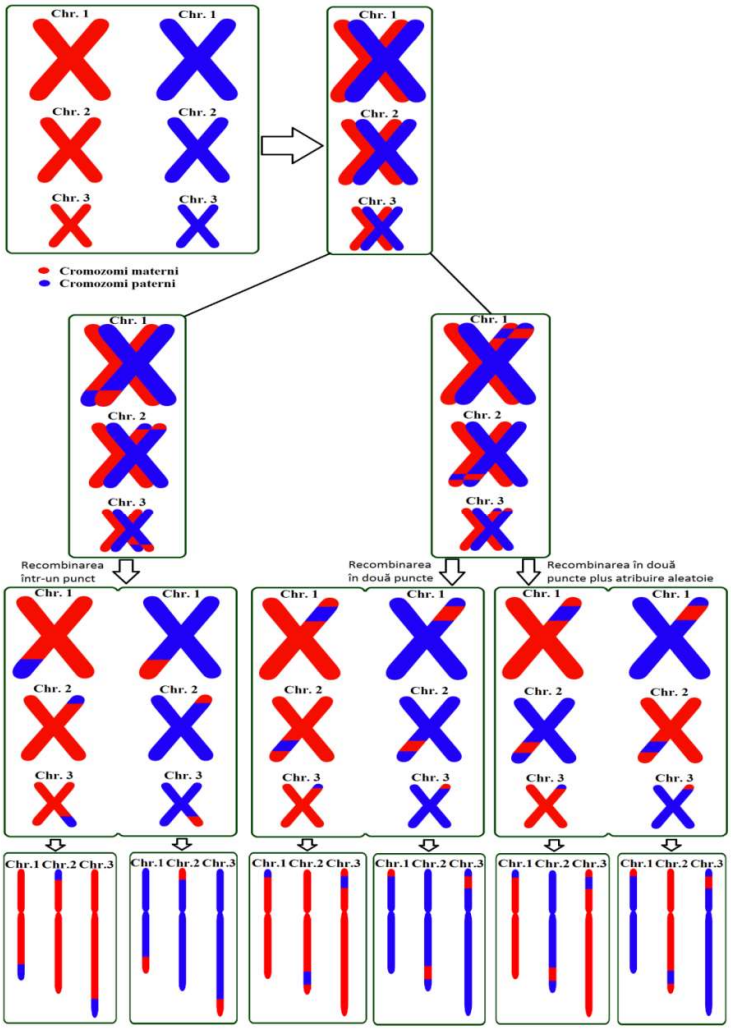


Fig. 3.3 – Two-point crossover.

This PhD thesis proposes an algorithm that benefits from modeling for the random assortment of chromosomes during meiosis. In our approach, the genotype is a priori configured in a variable number of chromosomes. The impact of using the random assortment of chromosomes is illustrated in Fig. 3.4. The genetic information is distributed on three chromosomes for clarity.



(a) (b) (c)

Fig. 3.4 - Genetic diversity (a) Single-point crossover; (B) Two-point crossover; (C) Two-point crossover followed by random assortment of chromosomes.

3.5. Operators for mutation

The operators for mutation originate in biology. The utility of point mutation in the context of DNA microarrays is limited. The number of active genes in an individual increases progressively in the evolved generations.

There are fundamental differences between the genetic code in biology and the principles used in genetic algorithms. The mechanisms of occurrence of genetic mutations cannot be modeled accurately in operators for genetic algorithms. However, principles learned from genetics, can be used to design improved mutation operators. Many versions of operators for mutations have been diachronically proposed by different authors, and genetics has often been the source of inspiration for such methods. I do not consider that the proposals described below are entirely original or that those principles were not previously modeled. I was very interested in the appraisal of these approaches in the context of the newly proposed algorithm for selecting attributes in DNA microarray data.

3.5.1. The Nonsense Mutation

The disorder occurs at the level of a single nucleotide in the DNA. By transcription, it becomes a stop codon in RNA. Consecutively, translation is prematurely terminated, and a protein is not completely synthesized.

In genetic algorithms, the nonsense mutation cannot model accurately the situation from biology. However, the phenomenon behind this type of mutation can be used for designing a valuable mutation operator for GAs. The nonsense mutation operator annuls all the alleles in a chromosome, consecutively to a randomly selected locus.

3.5.2. The Frameshift Mutation

The genetic code does not overlap and is continuous. Therefore, when a nucleotide in the sequence is accidentally deleted or added, all the consecutive substring is erroneously decoded.

The frameshift mutation operator uses this principle, but does not precisely model the biological phenomenon. Starting with a randomly generated position, the string is movement to the left. The last position on the chromosome is further supplemented with the allele 0. The affected chromosomes and interested loci are chosen at random.

3.5.3. The deletion

During meiosis, homologous chromosomes organized in tetrads exchange genetic information. Segments of chromosomes can be completely deleted from a chromosome and added excessive to the homolog. Thus, both the homologous chromosomes become abnormal.

In our implementation for mutation thru deletion, chromosomes and the edges of the intervals are randomly selected. All alleles in that range are annulled.

3.5.4. The monosomy

The operator for monosomy deletes an entire chromosome from a haploid set. The chance for such a mutation to occur is set at the initialization of the algorithm.

3.5.9. Transposons

Transposons are DNA sequences that can change their position in the DNA chain. The operator inspired by transposons randomly selects the chromosomes that suffer the mutation. Subsequently, a locus with an allele 1 is arbitrarily selected and a value for travel distance is randomly generated. The distance can result in negative values, specifying a migration to the left or a positive one, for migration to the right. The operation for mutation of inspired by transposon is illustrated in Fig. 3.5.

```
> chrConf
[1] 1 1 1 1 2 2 2 2
> individualsOriginal
Id 1005_at 1007_s_at 1008_f_at 1009_at 1020_s_at 1030_s_at 1038_s_at 1039_s_at
1  1       1         1         1       0         0         0         0         1
2  1       1         1         1       0         1         0         0         0
3  2       1         1         1       0         1         0         0         0
4  2       1         0         0       1         1         0         0         1
5  3       0         0         1       1         0         1         0         1
6  3       1         0         0       0         1         0         1         1
7  4       0         1         0       0         0         1         1         1
8  4       0         1         1       1         0         0         0         1
> individuals
Id 1005_at 1007_s_at 1008_f_at 1009_at 1020_s_at 1030_s_at 1038_s_at 1039_s_at
1  1       1         1         1       0         0         0         0         1
2  1       1         1         1       0         1         0         0         0
3  2       1         1         1       0         0         1         0         0
4  2       1         0         0       1         1         0         0         1
5  3       0         0         1       1         1         0         0         1
6  3       1         0         0       0         1         0         1         1
7  4       0         1         0       0         0         1         1         1
8  4       0         1         1       0         0         0         0         1
```

Fig. 3.5 – The mutation operator inspired by transposons.

4. Pachetul R dGAselID

The proposed approach for selecting attributes in DNA microarray data is implemented in the dGAselID software package. Although it is designed to select attributes from DNA microarray data, the algorithm may be applied to a wide variety of problems that require the selection of a variable number of attributes from vast data. All proposed methods and tests performed during the thesis are conducted using dGAselID.

Among many software solutions for AI and statistical analysis, a remarkable popularity was gained by packages built upon the R programming language [7]. R was developed on the S programming language, by Robert Gentleman and Ross Ihaka at Auckland University, New Zealand. BioConductor [8] is a joint effort, which supports researchers in bioinformatics, particularly those that focus on genetic analysis. BioConductor was started in 2001 at Fred Hutchinson Cancer Research Center and is currently being developed by Bioconductor core team, composed of researchers from several institutes and universities worldwide.

The proposed algorithm is implemented in the dGAselID software package, developed in R, fully integrated with Bioconductor. R project encompasses a range of methods for statistical analysis and integrates Bioconductor, a very valuable tool in genetic analysis. In addition, both projects are open and receiving contributions from numerous researchers actively involved in these areas. Bioconductor provides free access to many DNA microarray data sets and facilitates the development and comparison of methods for genetic analysis. R and Bioconductor offer many special packages, implemented to facilitate each step in microarray analysis, from the acquisition and preprocessing, until the interpretation of the results. The tools offered by Bioconductor were extensively presented in journals [9]. The DNA microarray research methodology with emphasis on the methods available in R and Bioconductor [10] was detailed in the literature.

During the search, the algorithm implemented in the dGAselID package provides the user with information about the evolution. Data on minimum, medium and maximum accuracy in the current population are displayed after each generation. Also, the researcher is informed about the number of applied mutations and the current stage in the development of the algorithm.

The algorithm also displays a graphical representation of the evolution after each generation. The evolutions of maximum and average accuracy, accompany a histogram of the genes most commonly selected in the graphical representation. A screenshot captured during the evolution of the algorithm is shown in Fig. 4.1.

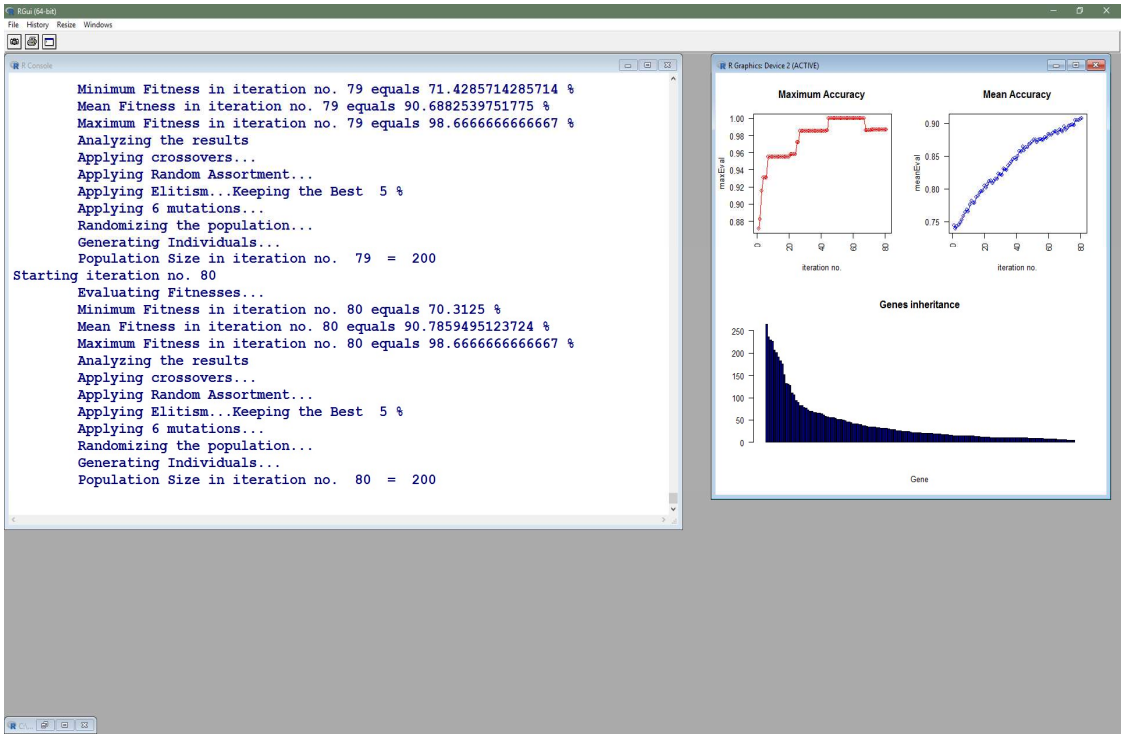


Fig. 4.1 – Screenshot after 80 generations.

The dGAselID software package proposes an original and effective solution to the problem of selecting attributes in DNA microarray data. The package is integrated and benefits from all the methods of analysis and visualization available in Bioconductor. Likewise, the package completes the data analysis methods available in this environment with an alternative to selecting attributes. The proposed method is flexible and applicable to a wide range of problems that require selection of attributes from vast data. The random assortment of chromosomes operator and the incomplete dominance model for genotype-phenotype mapping, favor exploring and produce superior results, improving the methods for selecting attributes available in MLInterfaces.

5. Experiments

The popularity DNA microarray technology enjoyed during the recent years, lead to the acquisition of an impressive amount of data and numerous results. This offers an opportunity for a mutually beneficial relationship between AI and bioinformatics. AI provides improved methods for addressing challenges in Bioinformatics. On the other hand, the affluence of data and results from bioinformatics provides a good framework for AI methods progress, further applicable in various fields.

We tested and evaluated the methods proposed in the thesis:

1) **incomplete dominance** for genotype to phenotype mapping,

2) **random assortment of chromosomes operator** in the context of genotypes with multiple chromosomes and a variable number of genes,

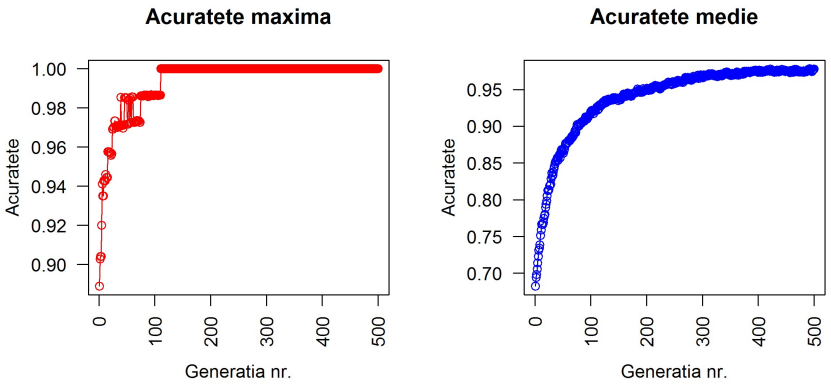
3) Various **mutation operators**.

5.1. The Acute Lymphoblastic Leukemia (ALL) dataset

The experiments in this chapter are conducted mainly with the Acute Lymphoblastic Leukemia (ALL) dataset [11]. The ALL data consists of 128 examples (patients suffering from leukemia) and 12625 attributes (probes on Human Genome U95 chips manufactured by Affymetrix).

5.2. The evaluation of incomplete dominance

The Incomplete dominance for genotype-phenotypes mapping was found to lay a solid base for the evolution of a diploid genetic algorithm. An example of such a progress with a dGA in selecting attributes from the ALL data is illustrated in Fig. 5.1. Commercial chips use multiple copies of the same gene in order to provide a measure of the quality of measurements and results.



a) b)

Fig. 5.1 – Evolution of GA a) The evolution of Maximum Accuracy after 500 generations; b) The evolution of Average Accuracy after 500 generations.

5.3. The evaluation of incomplete dominance version 2

The elitism operator offers an opportunity to address the tendency of AG to converge in a local optimum. In applying elitism, we can consider the adaptability of a genotype or an individual. Selecting the best performing genotypes to be retained in the next generation favors exploitation. Perpetuating in the next iteration of genotypes that were part of the fittest individuals can support exploration in GA.

The diploid genetic algorithm designed with incomplete dominance evolves with or without an elitism operator and benefits from an implicit selection by eliminating, from each individual, the less adapted genotype after each evaluation of the population. Therefore, a selection is inherent in the proposed model. Nevertheless, we named the genetic algorithm with an elitism operator at the level of the individual, version 2 (ID2), to underline this flexibility. In fact, it is the same algorithm with incomplete dominance and a different approach to the operator for elitism. For clarity, we refer to the initial implementation of incomplete dominance in diploid genetic algorithms as version 1 (ID1).

The contrast between the ID1 and ID2 approaches is illustrated by results obtained with the ALL data in Fig. 5.2. The aspects of maximum and average accuracies evolutions' for four initial populations (Fig. 5.3) further demonstrate these characteristics. While there is obvious advantage regarding the average accuracy for the ID1 implementation, ID2 proves to be superior in terms of consistency when examining the results, starting from the same initial population.

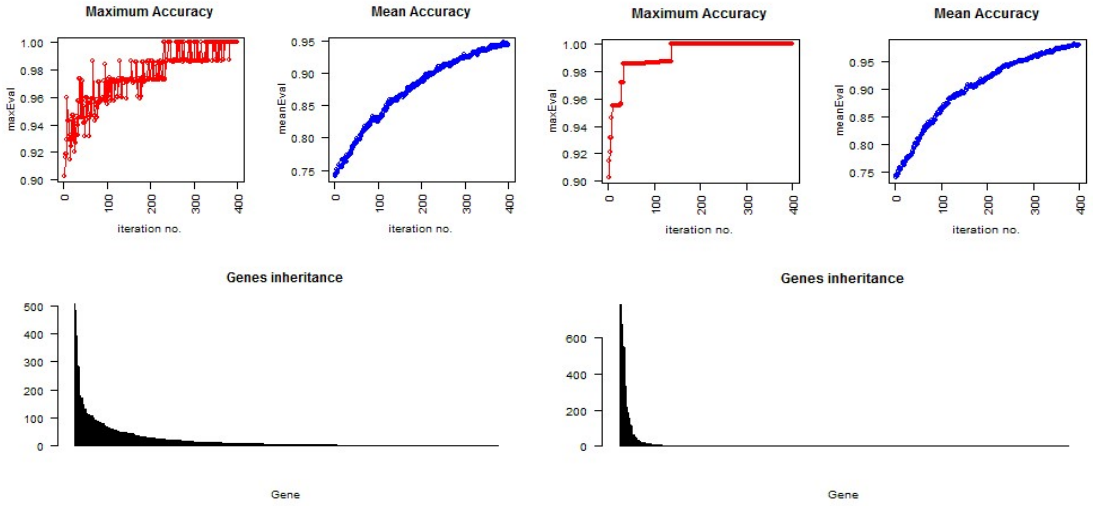


Fig. 5.2 – ID2 vs ID1.

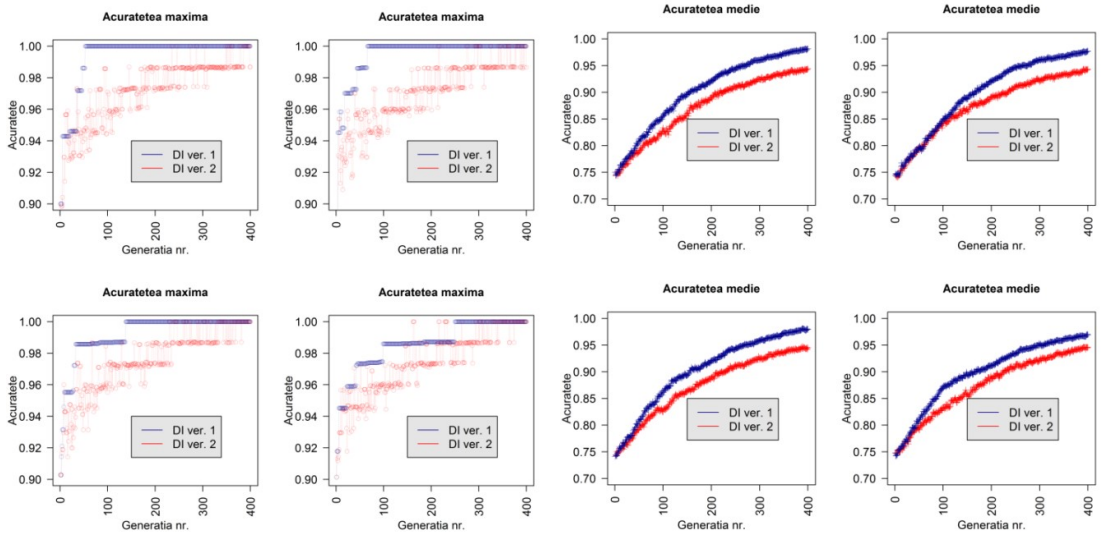


Fig. 5.3 – ID2 vs ID1 on four initial populations.

Clearly, ID2 favors exploration at the cost of exploitation, but, in this context, the selection of attributes in DNA microarray data, provides a desirable equilibrium between the two.

5.4. The Evaluation of the Random Assortment Operator

The exploration - exploitation relationship was advantageously affected by the implementation of Random Assortment of Chromosomes Operator (RAC). A manifest improvement was noticed especially regarding the evolution of the maximum accuracy (Fig. 5.4).

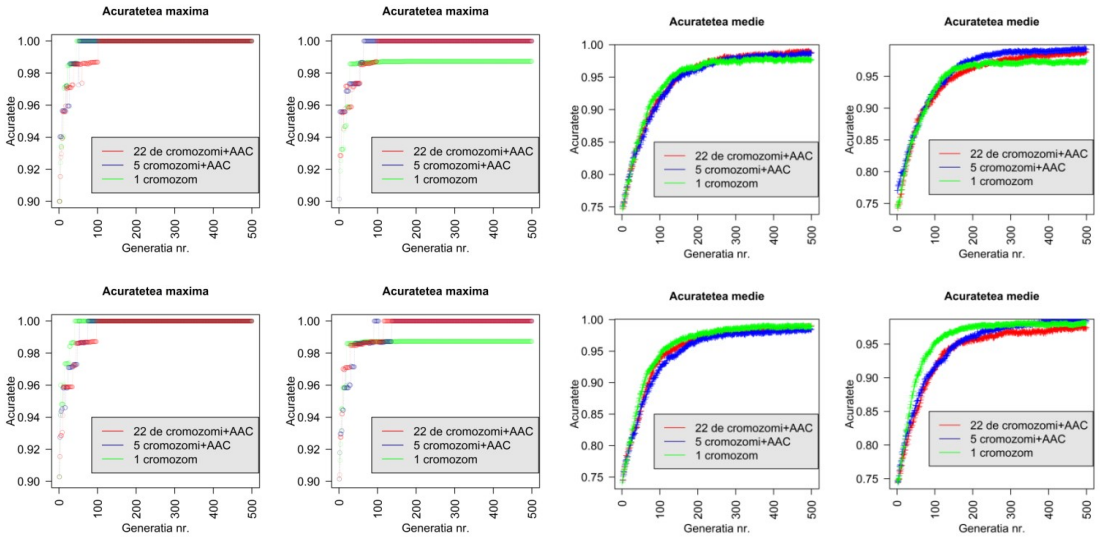


Fig. 5.4 – Evolution of the maximum and average accuracies on four different initial populations.

5.5. The Evaluation of the Transposons Operator

The effects induced by the operator for transposons are depicted in comparison with the alterations determined by the point mutation in Fig. 5.5-5.7. Although the GA's capacity to leave a local optimum is only slightly affected by the operator for transposons, evolutions of the maximum and average accuracies are supported by this approach. Moreover, the number of active genes in late populations does not increase, as is the case with the point mutation.

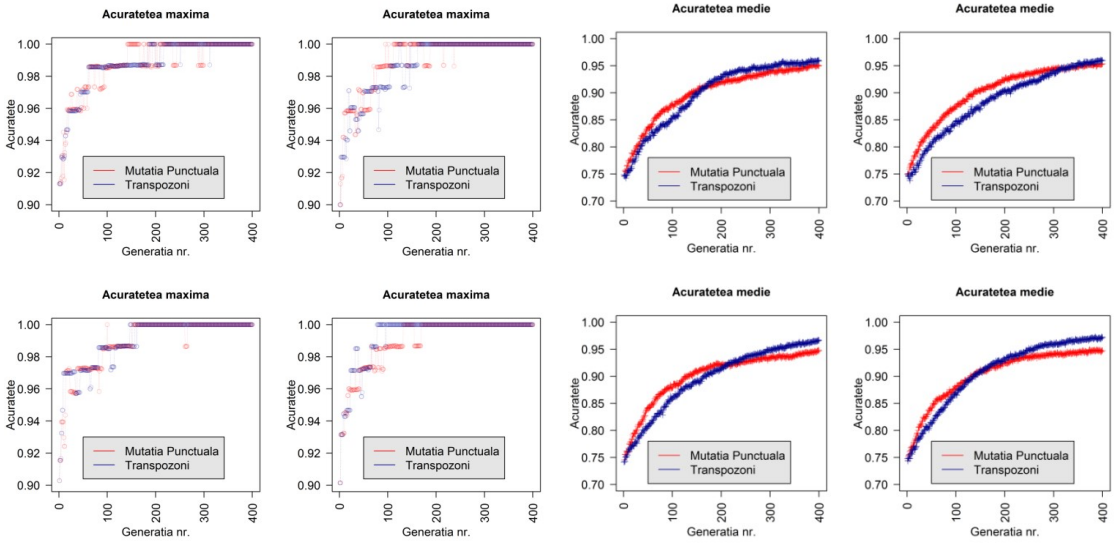


Fig. 5.5 - Transposons vs. Point Mutations.

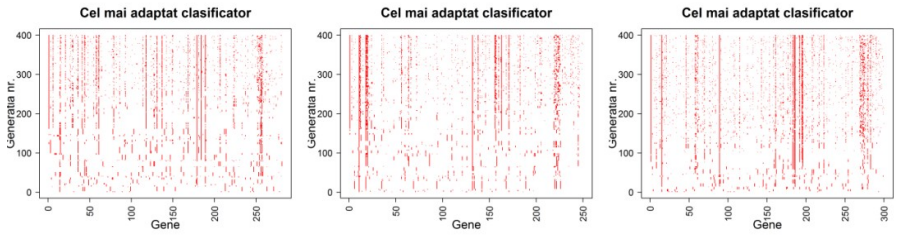


Fig. 5.6 - The evolution of the best classifier in three initial populations, when utilizing transposons.

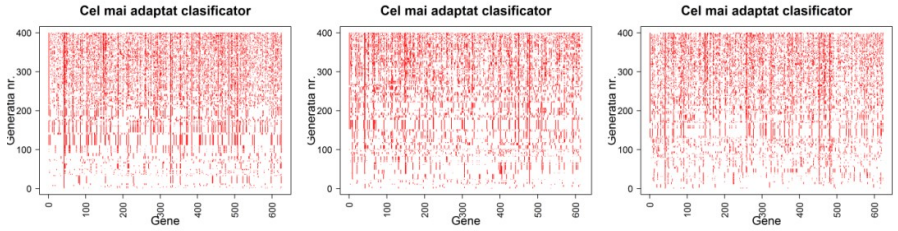


Fig. 5.7 - The evolution of the best classifier in three initial populations, when utilizing point mutations.

Arguments for and against the use of different proposals were determined for each described operator. Desirable effects were observed when employing transposons and monosomy operators. We cannot conclude that any of the proposed mutation operators is superior when selecting attributes from every microarray data. We recommend testing them in the context of a similar experiment to a priori assess the effects on the quality of evolution.

5.6. Evaluation of the cumulative effects of ID2 and RAC

We tested the combined effect of the two approaches, ID2 and Random Assortment of Chromosomes Operator (RAC) for selecting attributes with the GA. The evolution of the maximum accuracy (Fig. 5.8) reveals two extremely important aspects. On one hand, ID2 and RAC algorithm shows a tendency to leave a local optimum, property occasionally observed at the GA with ID1 and RAC. ID1+RAC implementation reached a solution with 100% accuracy during the vast majority of experimental replications, but ID2+RAC always found a perfect classifier, including the situations where the competing implementation failed to do so.

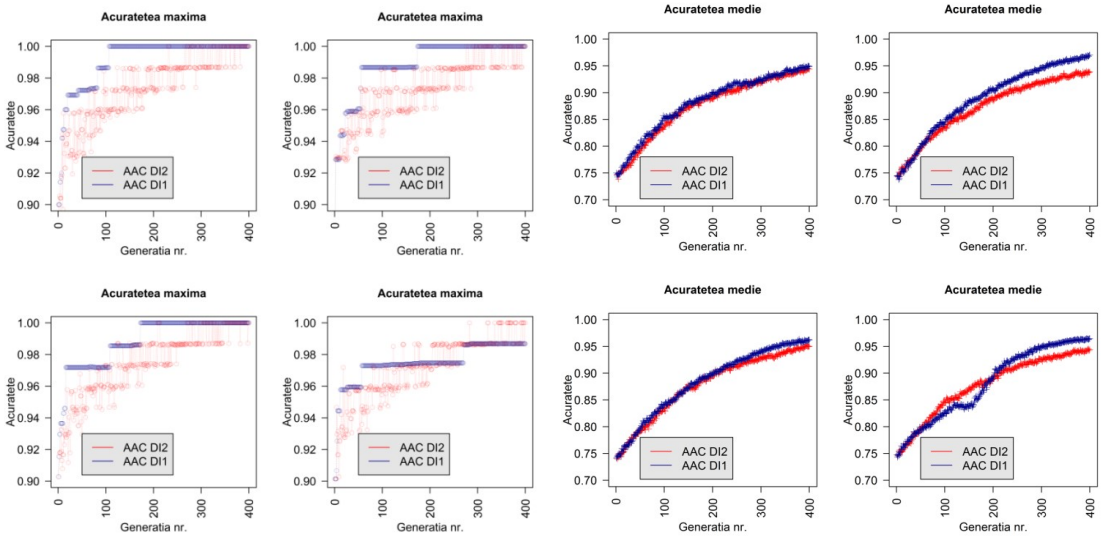


Fig. 5.8 - ID2+RAC vs. ID1+RAC.

6. Conclusions and personal contributions

To accomplish the proposed method, we studied phenomena underlying biological evolution. Starting from the premise that the principles underlying the success of evolution in nature are validated over billions of years, elucidation of these laws and their implementation in evolutionary algorithms will continue to improve the current methods and will boost the design of new approaches.

1) Incomplete Dominance. The proposed model is an alternative to defining a strict domination scheme for designing a diploid genetic algorithm. During the development of experiments it seemed useful to develop two variants, ID1 and ID2 with significantly different properties and applicable in different contexts. The DI2 implementation is more appropriate when selecting attributes in DNA microarray data.

2) Operator for Random Assortment of Chromosomes (RAC). The proposed operator models a phenomenon from biological evolution, which maintains genetic diversity and occurs during meiosis. The experiments confirm the suitability of this model for the genetic algorithms.

3) dGAselID software package. Perfectly integrated into R and Bioconductor, the dGAselID package facilitates access to the proposed method for selecting attributes in the context of genetic analysis and other research areas. The method is thus accessible to a diverse community of investigators in academia.

4) Alternatives to point mutation. We modeled and assessed the impact of the mutation operators in DNA microarray data analysis. The mutation operators implemented in the dGAselID software package are:

1. operator for **nonsense mutation**,
2. operator for **frameshift mutation**,
3. operator for **deletion**,
4. operator for **monosomy**,
5. operator **transposons**.

6.1. Future Work

Building on the results of the PhD thesis, in the near future, we intend to pursue several directions:

1) test the effect of the *partitioning of the genome into a variable number of chromosomes* in selecting attributes from DNA microarray data with **customizable biochips**,

2) evaluate the effects of *incomplete dominance* on the evolution of genetic algorithms employed for selecting attributes from data in **other research areas**, outside the scope of genetic analysis,

3) validate of *RAC operator* with data in **other research areas**,

4) design a more efficient *mutation operator* for selecting features from **DNA microarray data**.

References

[1] J.H. Holland. Adaptation in natural and artificial systems. University of Michigan Press, 1975.
[2] W.D. Hillis. Co−evolving parasites improve simulated evolution as an optimization procedure. Physica D 42:228–234, 1990.
[3] J.R. Levenick. Inserting introns improves genetic algorithm success rate: taking a cue from biology. Proceedings of the Fourth International Conference on Genetic Algorithms. Morgan Kaufmann, 1991.
[4] M. Mitchell. An introduction to genetic algorithms. The MIT Press, Cambridge, Massachusetts, London, England, 1999.
[5] J.M. Baldwin. A new factor in evolution. American Naturalist 30: 441–451, 536–553, 1986.
[6] R. Lewis. Human genetics, Ediţia a XI-a. McGraw-Hill Science/Engineering/Math, pag. 46, 2014.
[7] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
[8] W. Huber, V.J. Carey, R. Gentleman, ..., M. Morgan. Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods, 2015:12, 115.
[9] A. Koschmieder, K. Zimmermann, S. Trissl, T. Stoltmann și U. Leser. Tools for managing and analyzing microarray data. Brief Bioinform 13(1):46–60, 2012.
[10] W. Gregory Alvord, J.A. Roayaei,O.A. Quiñones și K.T. Schneider. A microarray analysis for differential gene expression in the soybean genome using Bioconductor and R. Brief Bioinform. 8(6):415-31, 2007.
[11] X. Li. ALL: a data package. Pachet R versiunea 1.14.0, 2009.