Results of Evolution Supervised by Genetic Algorithms

Lorenz JANTSCHI¹, Sorana D. BOLBOACĂ², Mugur C. BALAN³, Radu E. SESTRAȘ⁴, Mircea V. DIUDEA⁴

¹ Technical University of Cluj-Napoca, Department of Chemistry, 103-105 Muncii Blvd., 400641 Cluj-Napoca, Romania; lori@academicdirect.org; mbalan@temo.utcluj.ro
² "Iuliu Hațieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania; sbolboaca@umfcluj.ro
³ University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Mănăștur, 400372 Cluj-Napoca, Romania; rsestras@usamvcluj.ro
⁴ Babeș-Bolyai University, Cluj-Napoca, Faculty of Chemistry and Chemical Engineering, Department of Organic Chemistry, 11 Arany Janos Str., 400028 Cluj-Napoca, Romania; diudea@chem.ubbcluj.ro

Abstract

The efficiency of a genetic algorithm is frequently assessed using a series of operators of evolution like crossover operators, mutation operators or other dynamic parameters. The present paper aimed to review the main results of evolution supervised by genetic algorithms used to identify solutions to agricultural and horticultural hard problems and to discuss the results of using a genetic algorithms on structure-activity relationships in terms of behavior of evolution supervised by genetic algorithms. A genetic algorithm had been developed and implemented in order to identify the optimal solution in term of estimation power of a multiple linear regression approach for structure-activity relationships. Three survival and three selection strategies (proportional, deterministic and tournament) were investigated in order to identify the best survival-selection strategy able to lead to the model with higher estimation power. The Molecular Descriptors Family for structure characterization of a sample of 206 polychlorinated biphenyls with measured octanol-water partition coefficients was used as case study. Evolution using different selection and survival strategies proved to create populations of genotypes living in the evolution space with different diversity and variability. Under a series of criteria of comparisons these populations proved to be grouped and the groups were showed to be statistically different one to each other. The conclusions about genetic algorithm evolution according to a number of criteria were also highlighted.

Keywords: genetic algorithm (GA), evolution, genetic operators

Introduction

Simulation of evolution (through different parameters characterizing the sample under development) is a problem insufficiently explored in the literature; genetic algorithms are just one example.

Studies on other key operators for evolution are found in the literature and focus on algorithmic efficiency (seen in terms of speed with which they achieve maximum proximity and global optimum). A collection of representative works of this type is (Martin and Spears, 2001). Thus, various crossover operators are the subject of study in (Prügel-Bennett, 2001), mutation and crossing in (Spears, 2001), and other dynamic parameters in (Droste et al., 2001).

Studies are too often focused on solving difficult problems using genetic algorithms, sometimes dealing with efficiency (execution time, memory resources needed), rarely to the influence of the development of various strategies and objective (and here again especially on algorithm efficiency) and almost never on other parameters characterizing the sample under development.

For linking simulation → optimization a systematic literature search produced only one reference to a monograph (Stender et al., 1994), and literature is much richer but again on the reverse path from simulation to optimization.

Literature Review: Theory

A number of doctoral theses have been conducted on the subject of genetic algorithms in all fields of research and concerns on both basic and applied aspects.

A number of doctoral research of fundamental nature have their starting point the thesis (de Jong and Holland, 1975) supported under the guidance of one of the fathers of modern genetic algorithms - John Henry Holland (born February 2, 1929). Holland is an American scientist, Professor of Psychology, Professor of Electrical Engineering and Computer Science at the University of Michigan, Ann Arbor, he is a pioneer in nonlinear science and complex systems.

Based on the optimization problems, the work (de Jong and Holland, 1975) examines the efficiency of genetic algorithm for some classical problems (Fig. 1), which is known from the literature that classical optimization algorithms often failed.
Genetic algorithms have exceeded the boundaries of informatics domain due to the potential recovery of the computer simulation results.

Thesis with the objective of designing genetic algorithms, evolutionary programming, and implementation of studies based on them are found in practically all fields of research. Further representative works are detailed in this respect.

**Literature Review: Applications**

In the field of agriculture, the GA have found their usefulness in crop planning (Matthews and Kraw, 2001), construction on soil erosion risk assessment (Osman and McManus, 2007), in bioengineering to effectively control pollution in the catchments (Veith and Wolfe, 2002), in chemistry in the design of controlled sensory (Dai and Lodder, 2007).

In economics GA were able to solve optimization problems with multiple options (Aickelin and Dowsland, 1999), to manage the multi-scale modeling processes (Sas et al., 2007), to do mechanical optimization of composite structures (Gantovnik and Gürdal, 2005), and to provide solutions to environmental problems for water quality control strategy (Tufail and Ormsbee, 2006).

Finally, but not least, in biology, two lines come off in terms of development and use of genetic algorithms: the problems of development (Suzuki and Iwasa, 1998) and phylogenetic studies (Zwickl and Hills, 2006).

**Theorem 1** Let $S$ be any string of $L$ alleles: $(a_1, ..., a_L)$. If a population is mutated repeatedly (without selection or recombination) then:

$$\lim_{t \to \infty} p_S(t) = \frac{1}{C}$$

where $p_S(t)$ is the expected proportion of string $S$ in the population at time $t$ and $C$ is the cardinality of the alphabet.

**Theorem 2** Let $S$ be any string of $L$ alleles: $(a_1, ..., a_L)$. If a population is recombined repeatedly (without selection or mutation) then:

$$\lim_{t \to \infty} p_S(t) = \prod_{i=1}^{L} p_{a_i}(0)$$

where $p_S(t)$ is the expected proportion of string $S$ in the population at time $t$ and $p_{a_i}(0)$ is the proportion of allele $a$ at locus (position) $i$ in the initial population.

**Theorem 3** Let $S$ be any string of $L$ alleles: $(a_1, ..., a_L)$. If a population is mutated and recombined repeatedly (without selection) then:

$$\lim_{t \to \infty} p_S(t) = \prod_{i=1}^{L} \frac{1}{C}$$

where $p_S(t)$ is the expected proportion of string $S$ in the population at time $t$ and $C$ is the cardinality of the alphabet.
On the purely applicative, the use of genetic algorithms in agriculture and horticulture, genetic algorithm were found applications in plant growing studies (Venard and Vaillancourt, 2006), on taxonomic classification (Sarmiento-Monroy and Sharkey, 2006) and analysis of genetic diversity (Zhang and Ghahrial, 2006).

The Use of Genetic Algorithms on Structure-Activity Relationships

Optimization problem chosen for the study, namely the structure-activity relationships are at the junction of chemistry with computer sciences and biology. Continuous development of knowledge deposits like those provided by the NIH (National Institute of Health, USA), such as PubMed, PubChem, Genome, etc. stresses the need to have effective tools to articulate this deposited knowledge, and the structure-activity relationships are one of these instruments.

A genetic algorithm (GA) had been developed and implemented in order to identify the optimal solution in term of determination coefficient and estimation power of a multiple linear regression approach for structure-activity relationships. The Molecular Descriptors Family for structure characterization of a sample of 206 polychlorinated biphenyls with measured octanol-water partition coefficients was used as case study.

Probability Distribution Functions (PDFs) and Cumulative Density Functions (CDFs) for a series of observables recorded during GA supervised evolution to the global optimum were seeking in a experimental design in which 46 independent executions were taken into account on every selection and survival strategy as is depicted in Tab. 1 below.

Tab. 1. Simulation results

<table>
<thead>
<tr>
<th>Selection*</th>
<th>Survival*</th>
<th>Configuration**</th>
<th>Evolution**</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportional</td>
<td>Proportional</td>
<td>PCB-4044_cfg.txt, PCB-4044_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Proportional</td>
<td>Deterministic</td>
<td>PCB-2441_cfg.txt, PCB-2441_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Proportional</td>
<td>Tournament</td>
<td>PCB-9878_cfg.txt, PCB-9878_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Deterministic</td>
<td>Proportional</td>
<td>PCB-5108_cfg.txt, PCB-5108_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Deterministic</td>
<td>Deterministic</td>
<td>PCB-6690_cfg.txt, PCB-6690_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Tournament</td>
<td>Proportional</td>
<td>PCB-6690_cfg.txt, PCB-6690_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Tournament</td>
<td>Deterministic</td>
<td>PCB-4872_cfg.txt, PCB-4872_evo.txt</td>
<td></td>
</tr>
<tr>
<td>Tournament</td>
<td>Tournament</td>
<td>PCB-1758_cfg.txt, PCB-1758_evo.txt</td>
<td></td>
</tr>
</tbody>
</table>

*There are following pairs of selection and survival strategies (PP, PD, PT, PP, TP, TD, TT).

References


Foundations of genetic algorithms 6, Martin W. N., W. M. Spears Eds. San Francisco: Morgan Kaufmann.


Zwickl, D.J. (2006). Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion. PhD Thesis (Biology)-Supervisor Prof. Hills DM. University of Texas at Austin, Austin, TX, USA.