AcademicPres

Available online at www.notulaebiologicae.ro

Print ISSN 2067-3205; Electronic 2067-3264

Not Sci Biol 2 (3) 2010, 12-15

Notulae Scientia Biologicae

Results of Evolution Supervised by Genetic Algorithms

Lorentz JÄNTSCHI[1], Sorana D. BOLBOACĂ[2], Mugur C. BĂLAN[1], Radu E. SESTRAŞ[3], Mircea V. DIUDEA[4]

[1] *Technical University of Cluj-Napoca, Department of Chemistry, 103-105 Muncii Bvd., 400641 Cluj-Napoca, Romania; lori@academicdirect.org; mbalan@temo.utcluj.ro*

[2] *"Iuliu Haţieganu" 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 Mănăştur, 400372 Cluj-Napoca, Romania; rsestras@usamvcluj.ro*

[4] *Babeş-BolyaiUniversity, 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.

Received 09 May 2010; accepted 12 August 2010

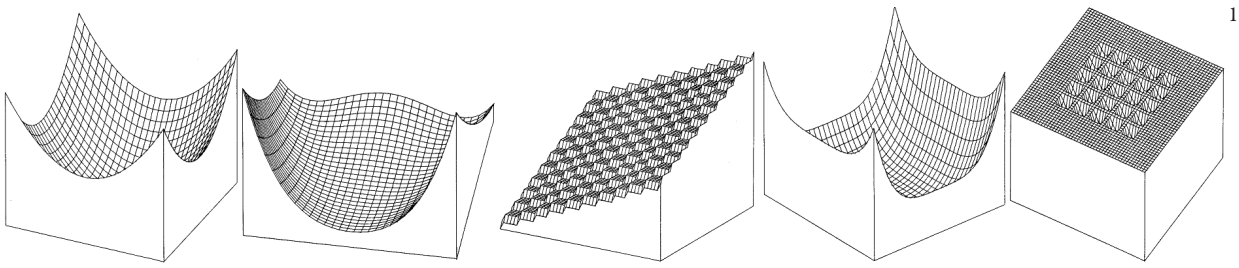

Fig. 1. Representation of 3D test functions from F1 to F5 used for assessing of the genetic algorithm in (de Jong and Holland, 1975)'s work

The doctoral research series which (de Jong and Holland, 1975) generated includes study of classical genetic algorithms (GAs) - and here the role of mutation and recombination is the subject of research (Spears and de Jong, 1998) - basically one of very few similar with the one reported here study (addressing the role of selection and survival), a GA modified form - a cooperative co-evolution (Potter and de Jong, 1997; Wiegand and de Jong, 2003), and where the initial population was divided into sub-populations called islands and evolution occurred on each island, allowing however the migration of individuals from one island to another (Skolicki and de Jong, 2007).

A valuable result of (Spears and de Jong, 1998), capitalized sometime later (Spears, 2001) is illustrated in Figure 2 referring the effect of mutation rate on progress towards equilibrium.

In Fig. 2, the Robbins equilibrium (Robbins, 1918) concerns a population that is recombined with no selection and mutation and the balance evenly cover a population that is repeatedly moved without selection and crossing that leads to a point where the objective function value is canceled (Solan, 2009).

Other doctoral research focused on basic research components include various adjustments made on genetic algorithms to solve special classes of difficult problems holding attention: multi-objective optimization (Zitzler and Thiele, 1999), the Pareto optimization (Knowles and Corne, 2002) and multi-agent systems (Panait and Luke, 2006).

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 (Sastry *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:*

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

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.

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.

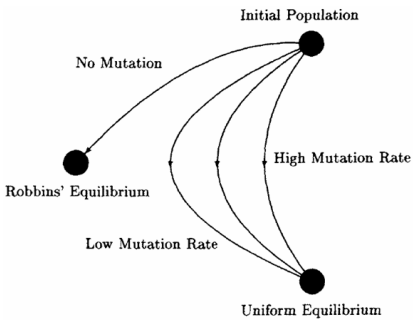

Fig. 2. The effect of the mutation on evolution to the equilibrium in (Spears, 2001)'s work

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 Ghabrial, 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

Selection*	Survival*	Configuration**	Evolution**
Proportional	Proportional	PCB_4044_cfg.txt	PCB_4044_evo.txt
Proportional	Deterministic	PCB_2441_cfg.txt	PCB_2441_evo.txt
Proportional	Tournament	PCB_9878_cfg.txt	PCB_9878_cfg.txt
Deterministic	Proportional	PCB_5108_cfg.txt	PCB_5108_evo.txt
Deterministic	Deterministic	PCB_6369_cfg.txt	PCB_6369_evo.txt
Deterministic	Tournament	PCB_6690_cfg.txt	PCB_6690_evo.txt
Tournament	Proportional	PCB_5828_cfg.txt	PCB_5828_evo.txt
Tournament	Deterministic	PCB_4872_cfg.txt	PCB_4872_evo.txt
Tournament	Tournament	PCB_1758_cfg.txt	PCB_1758_evo.txt

*There are following pairs of selection and survival strategies (PP, PD, PT, DP, DD, DT, TP, TD, TT)
**Files available at: http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival

During the research conducted in (Jäntschi and Sestraş, 2010), following conclusions were drawn:

÷The use of molecular descriptors families on multiple linear regression opens a natural pathway to do the optimization of the regression by using of a genetic algorithm;

÷The classical type of genetic algorithm designed and implemented evolutes relatively fast near to the optimum (in the conducted experiment PDF and CDF of the determination coefficient were obtained; probabilities from CDF to obtain 99% from the optimum in 1000 generations are: TD - 55%, PD - 67%, PP - 68%, TP - 73%, PT - 78%, TT - 80%, DD - 87%, DP - 95%, DT - 97%);

÷Evolution using different selection and survival strategies create populations of genotypes living in the evolution space with different diversity and variability; under a series of criteria of comparisons (number of genotypes, number of phenotypes, number of associations in regressions, top of 23 occurrences from 46 runs of above listed, etc.), these populations were proof to be grouped and the groups were showed to be statistically different one to each other;

÷The investigated evolution objective (determination coefficient of the multiple regressions to maximum) was found to be distributed by the Fisher-Tippett law of extreme values;

÷Obtaining of the distribution laws given the opportunity to construct the Lucky lottery and the Unlucky lottery relative to the chosen strategy of selection and survival;

÷The relative moments of evolution were found to be distributed by a one parameter degeneration of log-Pearson of type III curve, and two pairs of relatives (for relative moments of evolution) were found in strategies (PP and TT and TD and PD);

÷Number of evolutions were found to be distributed by a Fisher-Tippett (again) distribution;

÷The dominance in the Fisher-Tippett distributions of evolution objective are Weibull type III extreme values excepting DP strategy which have dominance of Fréchet type II extreme values during evolution;

÷The Fisher-Tippett distributions of number of evolutions are Weibull type III extreme values (again) excepting TP strategy, which have a Fréchet type II extreme values distribution.

÷The used number of evolutions the variance between strategies were found significantly smaller (4.07^2) than the variance inside strategies (9.68^2).

Acknowledgement

Finalcial support is gratefully acknowledged to CNCSIS-UEFISCSU Romania (Project PNII-Idei1051/202/2007).

References

Aickelin, U. (1999). Genetic Algorithms for Multiple-Choice Optimisation Problems. PhD Thesis (European Business Management) - Supervisor Prof. Dowsland K. University of Wales Swansea, Swansea, UK.

Dai, B. (2007). Simulations-guided design of process analytical sensor using molecular factor computing. PhD Thesis (Chemistry)-Supervisor Prof. Lodder RA. University of Kentucky, Lexington, KY, USA.

de Jong, K. A. (1975). An analysis of the behaviour of a class of genetic adaptive systems. PhD Thesis (Computer and Communication Sciences) - Supervisor Prof. Holland JH. University of Michigan, Ann Arbor, MI, USA.

Droste, S., T. Jansen and I. Wegene (2001). Dynamic Parameter Control in Simple Evolutionary Algorithms, p. 275-294. In:

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

Gantovnik, V.B. (2005). An Improved Genetic Algorithm for the Optimization of Composite Structures. PhD Thesis (Engineering Mechanics)-Supervisor Prof. Gürdal Z. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Jäntschi, L. (2010). Genetic algorithms and their applications. PhD Thesis (Horticulture)-Supervisor Prof. Sestras R. E., University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, Cluj, RO. http://l.academicdirect.org/Horticulture/GAs/Refs/Jantschi&Sestras_2010_Thesis.pdf.

Knowles, J. D. (2002). Local Search and Hybrid Evolutionary Algorithms for Pareto Optimization. PhD Thesis (Computer Science)-Supervisor Prof. Corne D. University of Reading, Reading, UK.

Martin, W. N. and W. M. Spears. (2001). Foundations of genetic algorithms 6. San Francisco: Morgan Kaufmann, p. 342. http://lccn.loc.gov/2001275761.

Matthews, K. B. (2001). Applying Genetic Algorithms to Multi-objective Land-Use Planning. PhD Thesis (Agriculture)-Supervisor Prof. Kraw S. Robert Gordon University, Craigiebuckler, Aberdeen, NSW, Australia.

Osman, N. Y. (2007). The Development of a Predictive Damage Condition Model of Light Structures on Expansive Soils using Hybrid Artificial Intelligence Techniques. PhD Thesis (Engineering and Industrial Sciences)-Supervisor Prof. McManus KAM. Swinburne University of Technology, Melbourne, VIC, Australia

Panait, L. (2006). The Analysis and Design of Concurrent Learning Algorithms for Cooperative Multiagent Systems. PhD Thesis (Computer Science)-Supervisor Assist. Prof. Luke S. George Mason University, Fairfax, VA, USA.

Potter, M. A. (1997). The Design and Analysis of a Computational Model of Cooperative Coevolution. PhD Thesis (Computer Science)-Supervisor Assoc. Prof. de Jong KA. George Mason University, Fairfax, VA, USA.

Prügel-Bennett, A. (2001). The Mixing Rate of Different Crossover Operators, p. 261-274. In: Foundations of genetic algorithms 6, Martin W. N. and W. M. Spears Eds. San Francisco: Morgan Kaufmann, http://lccn.loc.gov/2001275761 .

Robbins, R. B. (1918). Some applications of mathematics to breeding problems. III. Genetics 3:375-389.

Sarmiento-Monroy, C. E. (2006). Taxonomic revision of zelomorpha ashmead, 1900 and hemichoma enderlein, 1920 (braconidae: agathidinae) with a phylogenetic analysis of color patterns. PhD Thesis (Entomology)-Supervisor Prof. Sharkey MJ. University of Kentucky, Lexington, KY, USA.

Sastry, K. M. (2007). Genetic Algorithms and Genetic Programming for Multiscale Modeling: Applications in Materials Science and Chemistry and Advances in Scalability. PhD Thesis (Systems and Entrepreneurial Engineering)-Supervisor Prof. Goldberg D. E. and D. D. Johnson University of Illinois at Urbana-Champaign, Urbana, IL, USA.

Skolicki, Z. M. (2007). An Analysis of Island Models in Evolutionary Computation. PhD Thesis (Computer Science)-Supervisor Prof. de Jong KA. George Mason University, Fairfax, VA, USA.

Solan, E. (2009). Stochastic Games, p. 8698-8708. In: Encyclopedia of Database Systems. Berlin: Springer.

Spears, W. M. (1998). The role of mutation and recombination in evolutionary algorithms. PhD Thesis (Computer Science)-Supervisor Assoc. Prof. de Jong KA. George Mason University, Fairfax, VA, USA.

Spears, W. M. (2001). The Equilibrium and Transient Behavior of Mutation and Recombination, p. 241-260. In: Foundations of genetic algorithms 6 Martin W. N. and W. M. Spears Eds. San Francisco: Morgan Kaufmann, http://lccn.loc.gov/2001275761.

Stender, J. E. Hillebrand, J. Kingdon. (1994). Genetic algorithms in optimisation, simulation, and modelling. Amsterdam and Washington DC: IOS Press (261p.). http://lccn.loc.gov/94077520

Suzuki, H. (1998). Evolution of a Novel Function Fascilitated by Genetic Recombination in Genetic Algorithms. PhD Thesis (Biology)-Supervisor Prof. Iwasa Y. Kyushu University, Fukuoka, Fukuoka, Japan.

Tufail, M. (2006). Optimal water quality management strategies for urban watersheds using macro-level simulation models linked with evolutionary algorithms. PhD Thesis (Civil Engineering)-Supervisor Prof. Ormsbee LE. University of Kentucky, Lexington, KY, USA.

Veith, T. L. (2002). Agricultural BMP Placement for Cost-Effective Pollution Control at the Watershed Level. PhD Thesis (Biological Systems Engineering)-Supervisor Prof. Wolfe ML. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA.

Venard, C. M. P. (2006). The development of colletotrichum graminicola inside maize stalk tissues. PhD Thesis (Plant Pathology)-Supervisor Assoc. Prof. Vaillancourt L. University of Kentucky, Lexington, KY, USA.

Wiegand, P.R. (2003). An Analysis of Cooperative Coevolutionary Algorithms. PhD Thesis (Computer Science)-Supervisor Prof. de Jong KA. George Mason University, Fairfax, VA, USA.

Zhang, C. (2006). Genetic diversity of bean pod mottle virus (bpmv) and development of bpmv as a vector for gene expression in soybean. PhD Thesis (Plant Pathology)-Supervisor Prof. Ghabrial SA. University of Kentucky, Lexington, KY, USA.

Zitzler, 1999. Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications. PhD Thesis (Technical Sciences)-Supervisor Prof. Thiele L. Swiss Federal Institute of Technology, Zurich, Elveția.

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.