University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca
Doctoral School
Faculty of Horticulture

Lorentz JÄNTSCHI

Summary of PhD Thesis

Genetic algorithms and their applications

Scientific Advisor:
Prof. Univ. Dr. Radu E. SESTRAŞ

Cluj-Napoca
2010

Genetic algorithms and their applications

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor in Philosophy, Horticulture - Genetics and Plant Amelioration
at University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca

by

Lorentz JÄNTSCHI

BA, Informatics, Babeş-Bolyai University, 1995
BA, Chemistry and Physics, Babeş-Bolyai University, 1997
PhD, Chemistry, Babeş-Bolyai University, 2000
MS, Agricultural Sciences, UASVM Cluj-Napoca, 2001

Advisor: Radu E. SESTRAŞ, Professor
Dean of Faculty of Horticulture

Summer Semester 2010
University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca
Cluj-Napoca, Cluj

Contents

Introduction

The thesis entitled "Genetic algorithms and their applications" having the aim of "Simulating the evolution with genetic algorithms in structure-activity relationships optimization problems" is an interdisciplinary approach to fundamental research.

The subject is optimization of the quantitative relationships (between the structure of chemical compounds and their biological activity) hard problems (those with exponential complexity).

The framework for the construction and application of a genetic algorithm to solve optimization problem were built. It was built within a defined genetic algorithm. The genetic algorithm was implemented in an evolutionary program, was applied on an experimental data set and the evolution was recorded.

Experimental design was done in order to make the transition to the problem of simulation from the optimization problem - namely simulating evolution using different selection and survival strategies. A 3 by 3 contingency of selection and survival strategies (following proportional, tournament and deterministic algorithms) were created and evolution were recorded over 20,000 generations repeated 46 times for each strategy pair.

Statistical inferences were analyzed in qualitative and quantitative observables of the evolutionary process controlled by the different development strategies using different variables that evolutionary program was set to record the values.

The results are mainly of fundamental research nature. Statistical analysis of evolution simulation results offered responses to questions like: *What is the distribution law for evolution objective?*, *What is the distribution law for evolution moments?*, *What is the distribution law for number of evolutions?*, *How it is influenced the genotypic variability and diversity by the choice of evolution strategy?*, *How early developments occur in relation to the chosen evolution strategy?*, *How often evolution occur in relation with the chosen evolution strategy?*, *How spread are the evolution in relation with the chosen evolution strategy?*, *How predictable are the evolutions in relation to the evolution strategy chosen?*, *What are the similarities and differences between evolutions following different strategies?*, etc.

A series of results of applicative nature were obtained: implementation of the (classical) genetic algorithm in a evolutionary program able to solve a hard problem of structure-activity relationship optimization by using families of structure descriptors; implementation of software modules for automating the molecular geometry optimization; implementation of software modules for Anderson-Darling statistic usage for agreement between observation and a model; implementation of Grubbs procedure for identifying and removal of observations in error relative to a model.

Thesis gives also solutions for technological transfer, covering answers to a series of problems like: I want to make an evolution to reach an objective; I am interested to know which strategy to apply in order to reach the objective.

Problems of structure-activity relationships optimization

Mathematical approach of SAR (structure-activity relationships) for BAC (biologically active compounds), started in nineteen century, were capitalized through the born of the quantitative structure-activity relationships (QSAR) concept (Hammett, 1935), a mathematical tool describing the quantitative link between chemical structure and biological activity of a given set of compounds. SAR records were communicated in scientific literature since 1868, when (first) Crum-Brown & Fraser were given the idea to seen the activity of compounds as a function of chemical structure and composition (Crum-Brown & Fraser, 1868), but only after almost forty years the QSAR paradigm were found practical useful in agro-chemistry, pharmaceutical chemistry, toxicology, etc (Hansch & Leo, 1979). Scientific literature contains numerous reports on usage of SQARs in the methodology of designing new BACs, and the monograph (Diudea & others, 2001) covers a good part of it.

Simulating evolution with genetic algorithms

Hard (Weismann, 1893) and soft (Lamarck, 1809) inheritance, selection and survival (Darwin, 1859), traits (Mendel, 1866) and genes (Morgan & others, 1915) crossover, a long and

contentious debate over the 19th century (Fisher, 1954) are all pieces from a puzzle building today the modern genetics (Ayala & others, 1994) and being the sources of inspiration for genetic algorithms (GAs). First simulating of evolution studies are of Nils Aall BARRICELLI (Barricelli, 1954). Few times later, Alex FRASER (1923-2002) gives a series of studies about simulation of artificial selection of organisms having multiple loci controlling a measurable trait. Fraser's simulations (Fraser, 1957-1970) include all essential elements of modern GAs.

The genetic algorithms intrinsic methodology

The tool for developing genetic algorithms is informatics, and thus we should call for it here. Usually, in day by day life issues as in scientific research we operate with ***problems***. In informatics and its relatives (as chemo-informatics and bioinformatics) a problem has a precise meaning, close to the meaning of the algorithm. An ***algorithm*** is essentially a recipe specifying what to do in certain circumstances to reach an objective. An algorithm requires two resources to ***solve*** a problem: time (in the sense of execution time, correlated with the number of elementary instructions) and space (to store entry data and its variables). Not all problems are of same ***complexity*** and the same for solving algorithms. Some problems have ***exponential complexity*** (the best possible algorithm solves the problem - giving the exact solution(s) - in an execution time growing exponential with the size of the entry data), being called ***hard***, because even the best available (or possible) algorithm will be probably un-useful when are feed with entry data from practice (Falkenauer, 1998). If a problem is hard, then the search for the optimum often goes out of available time for real applications. But fortunately, a series of hard problems does not call for the optimum, a ***good solution*** being enough. For a variety of hard problems, one or more ***heuristics*** were designed. Heuristics and sets of rules designed to solve a given problem usually based on common sense (relative to the expected solution) by avoiding gross errors; they are not designed to give always a exact solution and to give a solution for any entry data. Even if the most of the heuristics are ad-hoc and dependent on the given problem, together with developing of the informatics, the researches were succeeded to formulate three heuristics being very general (able to be applied to a large variety of hard problems), called (because of their generality) ***meta-heuristics***, all three being stochastic in their nature (implies one or more random variables; implies the chance or the probability), two of them being inspired from natural processes having place around us from all times, one of them being genetic algorithms. Even if first studies are in year 1954, systematic researches started after 1970 (Bosworth & others, 1972; Holland, 1975) and were reinvented after 1990 (Davis, 1991; Holland, 1992), together with the progress of computation tools. An important issue about heuristics is the NFLT (No Free Lunch Theorem) on ***algorithmic complexity*** (Wolpert & Macready, 1995&1997; English, 1996), theorem stating that using three evaluation criteria (speed, precision and scope) all algorithms are equivalent (for algorithms A and B, for every set of data for which A is performs better than B it exists a set of data for which B performs better than A). Genetic algorithms serves in phylogenetic (Jäntschi & others, 2008-PTA) and gene sequence (Jäntschi & others, 2009-GSA) analysis, hard problems of dynamics of processes (Jäntschi & others, 2009-DPA) and in any other category of decision, classification, optimization or simulation (Falkenauer, 1998) hard problems.

Research frame

Continuing growth of knowledge banks like the ones administered by the NIH, such as PubMed, PubChem and Genome underlines the necessity to posses efficient tools to relate this knowledge; the SARs are one of such kind of tools.

The research question "*How the evolution can be observed and characterized via different parameters characterizing the sample simulated to evolutes?*" are not enough explored in the specialty literature on genetic algorithms subject. Studies of different operators essential for evolution are focused mainly on algorithmic efficiency - and representative for this approach is the collection from (Martin & Spears, 2001).

Very few studies are about the influence of the evolution strategy on evolution objective, and almost nothing about the influence of different parameters characterizing the evolving sample on evolution objective.

The GAs passed out long time ago the border of the informatics field, because of its potential of results capitalization. PhD theses having objectives of projecting, implementation and use of genetic algorithms are found practically in all fields of research. Thus, in agriculture GAs were found useful to crop planning (Matthews & Kraw, 2001), in constructions to assess the risk of soil damage (Osman & McManus, 2007), in bioengineering to efficient control of pollution at a hydrographic basin level (Veith & Wolfe, 2002), in chemistry at design of sensor-based controlled processes (Dai & Lodder, 2007), in economics at optimization of problems with multiple options (Aickelin & Dowsland, 1999), in management at multi-scale processes modeling (Sastry & others, 2007), in mechanics at optimization of composite structures (Gantovnik & Gürdal, 2005), in environment at strategy chousing for water quality control (Tufail & Ormsbee, 2006), in biology in phylogenetic analysis (Zwickl & Hills, 2006) and evolution studies (Suzuki & Iwasa, 1998). Only uses of GAs embedded in evolutionary programs are reported in (Venard & Vaillancourt, 2006) for studies of vegetables growing, in (Sarmiento-Monroy & Sharkey, 2006) for taxonomic classifications and in (Zhang & Ghabrial, 2006) for genetic diversity analysis.

Research aim and objectives

The research aim covered projecting of a GA, implementation of an evolutionary program based on it, and then the analysis of the influence of different selection and survival strategies on evolution controlled by the genetic algorithm feed with data for structure-activity relationships optimization in a series of biologically active compounds. Three objectives were followed:

1. (method) design of the GA (including defining of the hard problem); formulation of the problem in genetic terms; projecting of the GA; implementation and documentation of the evolutionary program embedding the GA;
2. (results) simulation of the evolution (defining of the observables; defining of the contingency between selection and survival strategy; projecting of the statistical experiment; run of the experiment;
3. (analysis) analysis and interpretation of the runs results about qualitative observables and about evolution objective (was set to r2 - determination coefficient) - quantitative observable during evolution.

Definition of the QSAR optimization problem taken

The chosen set of molecules for the study is the PCBs data set (with 209 compounds in the series). For this set of data log(K_{ow}) were available measurements in same conditions of experiment for 206 compounds (Eisler & Belisle, 1996; Mullins & others, 1984); (Jäntschi & others, 2007-Chromatogr). Kolmogorov-Smirnov (Kolmogorov, 1941; Smirnov, 1948), Anderson-Darling (Anderson & Darling, 1952; Scholz & Stephens, 1986 & 1987), and Pearson-Fisher Chi-Square (Pearson, 1900; Fisher, 1922-X2; Fisher, 1924; Fisher, 1935) statistics were used to measure the agreement between observed data and normal distribution model. Grubbs test (Grubbs, 1969) was used to identify an outlier. HyperChem (licence v. 8.0/2007) was used (using AMBER molecular mechanics model, POLAK-RIBIERE optimization algorithm, and AM1 method for semiempirical energy calculations). Molecular Descriptors Family (Jäntschi, 2004; Jäntschi, 2005; Jäntschi & Bolboacă, 2007-Results) were used to create the population of structure descriptors from which to feed the genetic algorithm. The search was started for multiple linear regressions with four descriptors members of MDF relating the observed log(K_{ow}) of 206 PCBs.

Definition of the genetic problem created

Evergy gene codifies an operator used in construction of the chromosome of a molecular descriptor. Every descriptor (of a family of descriptors, such as MDF) is a genotype and all together is the genetic material of the family. Folowing table gives the search space created by MDF:

Family	Gene	Genome																												
MDF	D_M	t	g																											
	A_P	C	H	M	E	G	Q																							
	I_D	D	d	O	o	P	p	Q	q	J	j	K	k	L	l	V	E	W	w	F	f	S	s	T	t					
	I_M	r	R	m	M	d	D																							
	F_C	m	M	D	P																									
	S_M	m	M	n	N	S	A	a	B	b	P	G	g	F	f	s	H	h	I	i										
	L_O	I	i	A	a	L	l																							

The working methodology af genetic algorithms suppose a initial prelevation (at random or using a strategy) of a ***sample*** of chromosomes from the genetic material - in this case a array of MDF members - from X_1 to X_p which enters in ***cultivar*** for conducting the ***evolution process***. The ***genetic algorithm*** operates on the sample which is changed (in part) in every ***generation***. Every set of `n` descriptors (where n is the multiplicity order for MLR) is a point in the ***search space*** and a ***possible solution***. The operators which change the genetic code are crossover and mutation. ***Crossover*** of two genotypes suppose chousing of a part from the stream of genes to be cross over (at random or using a strategy) and the values of the parts are swithched one in the place of the other, and two descendents are produced. ***Mutation*** of a genotype supposes the changing of the value of a (or more) gene with other allowed value from the list of possible values for the given gene. Both crossover and mutation produces ***descendents***. The ***selection*** of the genotypes is the operation which mutation and crossover calls for, are based on a ***strategy*** and uses a score function (***selection score***). At least a part of the descendents is ***viable*** (descriptors), being able to be part of a ***viable solution*** (MLR) in the next generation(s). Viable descriptors replace a part from the sample through a ***survival*** process. As selection process, survival process uses a score function (***survival score***) and uses a ***strategy***. The evolution objective are recorded during evolution using a score function (***objective score***). Once in every generation the individuals which gives the best objective score (enters in the best MLR) are ***marked***. An option is to automatically qualify for the next genetation the marked individuals (no survival strategy applies on it). Not all individuals of a generation (including parents and descendants) survive and will be present in the next generation. The reasoning to do this is for keeping constant the sample size (thus the number of replaced individuals is equal to the number of viable descendants).

Selection and survival based on selection and survival scores are applied through a selection and survival strategies, using an ***algorithm*** for every different strategy. ***PS*** algorithm constructs a ***proportional strategy*** using an array of scores and gives to an individual a chance (to be selected in selection process or to be killed in survival process) proportional with the score, and returns a given number N_Sel of individuals using their chances. ***DS*** algorithm constructs a ***deterministic strategy*** returning the N_Sel individuals with the first N_Sel highest scores (if is necessary applies a random qualification at equal scores). ***TS*** algorithm constructs a ***tournament strategy*** using the array of scores and qualifies N_Sel individuals through a repeated N_Sel times tournament of two individuals.

The genetic algorithm acts as follows:

÷ the sample of the given size (N_Gen) is created (containing predefined or random individuals);
÷ repeat steps 1..6 until *objective score is satisfactory* or *a number of generations are eshausted*;
÷ Step_1: Computes selection scores, survival scores and ojective scores (and eventually include in the next generation the marked individuals);
÷ Step_2: Select (using selection strategy) N_Cro pairs of individuals;
÷ Step_3: For every one from 2×N_Cro, using p_Par (low) probability and a discrete uniform distribution pick a number of N_Mut genes and make a mutation on it (parents); save the result (whatever mutated or not, 2×N_Cro individuals);
÷ Step 4: For every one from N_Cro, using a discrete uniform distribution pick the sequence of genes to be crossover, do crossover; save the result (replace the previous one, 2×N_Cro individuals);
÷ Step_5: For every one from 2×N_Cro, using p_Chi (low) probability and a discrete uniform distribution pick a number of N_Mut genes and make a mutation on it (childs); save the result (whatever mutated or not; replace the previous one, 2×N_Cro individuals);
÷ Step_6: Replace (sing survival strategy) a part of N_Gen with a part of 2×N_Cro;

Definition of the simulating evolution obtained

The ***parsimony*** principle is the essence staying at the basis of the link Optimization(SAR) → Evolution (Observables). The principle were applied in the simulating of the controlled (under given parameters) evolution of the sample toward the evolution objective under the different

selection and survival strategies given above. The principle was applied in order to compare the evolutions under different selection and survival strategies.

The parameters kept constant during the (parsimony) experiments are given in the following table.

Class	Parameter	Value
Topology of the family of molecular descriptors	Genes	mp/fc/oi/id/ap/dm
	Addre	fc/ap/id/oi/dm/mp
	mp	mMnNSPsAaBbGgFfHhIi
	fc	mMDP
	oi	RrMmDd
	id	DdOoPpQqJjKkLlVEWwFfSsTt
	ap	CHMEGQ
	dm	gt
Topology of the informational infrastructure	Mydb	MDFSARs
	TabE	PCB_lkow_data
	TabM	PCB_lkow_tmpx
Topology of the sample	sn0_SAMPLE_Size	12
	a_v_ADAPT_Variance	0.1
	ajb_ADAPT_JarqueBera	0.1
	a_c_ADAPT_Correlation	0.1
	g_r_GENERATIONS_first_rich	Yes
	b_k_RUNS_kepp_best_in_sample	Yes
	b_f_RUNS_get_best_from_file	No
Crossover	cn0_CROSSOVER_Pairs	2
	m_m_MUTATION_Genes	2
	mpp_MUTATION_Parent_probability	5%
	mcp_MUTATION_Child_probability	5%
Evolution objective	rn0_REGRESSION_Multiple	4
	b_p_SELECTION_parameter	r2
	b_o_SELECTION_objective	max
Experiment	e1n_GENERATIONS_max	20000
	e0n_RUNS_number	46
Selection	sfn_FITNESS_normalized	No
	sfr_FITNESS_ranks	No
	sfa_FITNESS_accuracy	10000
	sff_FITTEST_function	r2_min
	sfo_FITTEST_objective	max
	fr2_FITTEST_r2_p	1.0
	fse_FITTEST_se_p	1.0
	fMt_FITTEST_Mt_p	1.0
	fHr_FITTEST_Hr_p	1.0
Survival	v_p_SURVIVAL_phenotyping_p	1.0
	v_g_SURVIVAL__genotyping_p	1.0
	vfr_SURVIVAL_ranks	No

Two parameters (*sfs_FITNESS_strategy* and *vfs_SURVIVAL_strategy*) were taken different values once at the time for the parameters kept constant (the above table), nine executions of the program being independently started, and the results were recorded in separate files (two files per execution, table above).

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

Benford test checking the output data

Frequencies for number of distinct & viable genotypes (num_obs) and for total number of viable genotypes (sum_obs) in the generations producing evolution for every strategy of selection (Sel) and survival (Srv) from the list {P(Proportional), T(Tournament), D(Deterministic)} for frames of thousands of generations were extracted from the observed execution results and place

together -2(observables)×9(strategies)×20(milenia) = 360 frequencies. Kolmogorov-Smirnov (K-S) and Pearson-Fisher Chi-Square (C-S) statistics were used to measure the agreement between observed data and Benford Law for first three digits of the frequencies. With 5% risk being in error C-S did not rejected the hypothesis of Benford law distribution for first thee digits: $X^2(d_0, df=9\text{-}1\text{-}1) = 11 < 14.7 = \chi^2(df=7, p=5\%)$; $X^2(d_1, df=10\text{-}1\text{-}1) = 6.6 < 15.5 = \chi^2(df=8, p=5\%)$; $X^2(d_2, df=10\text{-}1\text{-}1) = 7 < 15.5 = \chi^2(df=8, p=5\%)$. With 5% risk being in error K-S did not rejected the hypothesis of Benford law distribution for first thee digits: $D\sqrt{n}(d_0) = {}^{14}/_{80} < {}^{31}/_{80} = K(df=9, p=5\%)$; $D\sqrt{n}(d_1) = {}^{17}/_{152} < {}^{56}/_{152} = K(df=10, p=5\%)$; $D\sqrt{n}(d_2) = {}^{31}/_{255} < {}^{94}/_{255} = K(df=10, p=5\%)$.

Analysis of variability

Fequency of the genotypes in the sample during evolutions allow making of remarks regarding the capacity of adaptation, and serves to characterize the variability of the genetic material of the sample in relation with the selection and survival strategy used.

A contingency of observables were created {Top23, Total}×{Dist, Sum, Part} where Top23 - all over 23 occurences, Total - no lower limit of occurrences, Dist - number of distinct genotypes, Sum - number of genotypes, Part - number of genotypes having at least a phenotype with which a MLR was created. Six times the contingency Selection = (P, T, D) × Survival = (P, T, D) were used (for every observable defined), and in every case the Pearson-Fisher Chi-Square (C-S) statistic were used to measure the independence of the observable on selection and survival strategy.

According to (Fisher & Mackenzie, 1923-Treatment) the product formula for calculating expectations in contingency table under assumption of independence is an approximation for the solution of a polynomial equation minimizing the disagreement with the assumption that the observation has a probability given by the product of two probabilities given by the two implied events applied on the observable. Thus the formulation of the statistical hypotheses and their answer (after analysis with C-S statistic) are:

÷ Selection and survival strategy are independent events when the number of distinct viable genotypes are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 70$);

÷ Selection and survival strategy are independent events when the total number of viable genotypes are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 135$);

÷ Selection and survival strategy are independent events when the number of viable genotypes having phenotypes in regression equations are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 187$);

÷ Selection and survival strategy are independent events when the number of distinct viable genotypes with over 23 occurences in 46 runs (Top23) are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 14.6$);

÷ Selection and survival strategy are independent events when the total number of viable genotypes with over 23 occurences in 46 runs (Top23) are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 420$);

÷ Selection and survival strategy are independent events when the number of viable genotypes with over 23 occurences in 46 runs (Top23) having phenotypes in regression equations are observed in the generations which produces evolution? - Answer: NO (with $X^2(df = 4) = 440$);

A linear relationship between the numbers implied in the statistics from above was found; thus for the numbers cumulating the total frequencies of {Dist, Sum, Part} determination coefficients are: $r^2(N\_Dist, N\_Sum) = 0.982$; $r^2(N\_Dist, N\_Part) = 0.982$; $r^2(N\_Sum, N\_Part) = 0.999$.

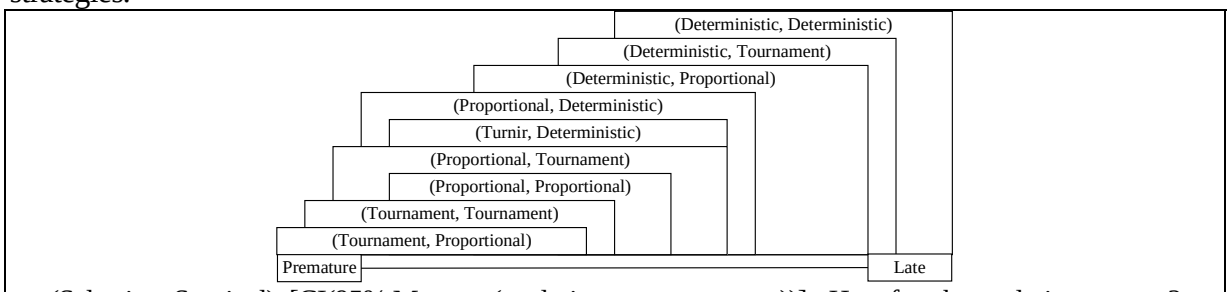
The contingency of observables {Top23, Total}×{Dist, Sum, Part} were applied for every run (from run 1 to run 46) recording the numbers; on the obtained data, mean and standard deviation together with their 95% confidence intervals were used to compare selection and survival strategies. A series of important remarks was extracted from the analysis, such as:

÷ Independent of survival strategy the deterministic selection has as effect the decreasing

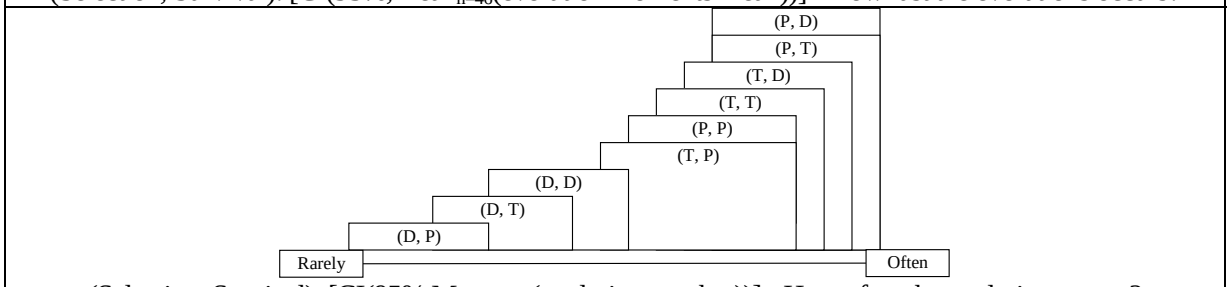
(statistically significant) of the number of disctict genotypes;

÷ When survival is deterministic, excepting the deterministic selection all other increases (statistically significant) at the Total all observables (Dist, Sum, Part);

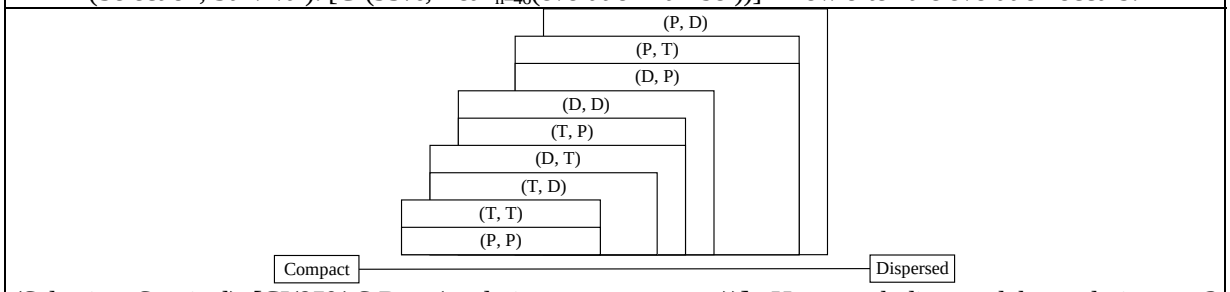
Two measures were defined and used for the records in the given interval of generations (0..20000) for which the observation were made: the number of the evolutions n(·,·) - as a measure of adaptation capacity, and the mean of the evolution observed moments m(·,·) - as a measure of adaptation speed (where the dots are places for selection and survival strategies). Means and standard deviations were calculated with 95% confidence intervals. Following plots rescaled these values from min to max, keeping proportions, and served for comparison between different strategies.



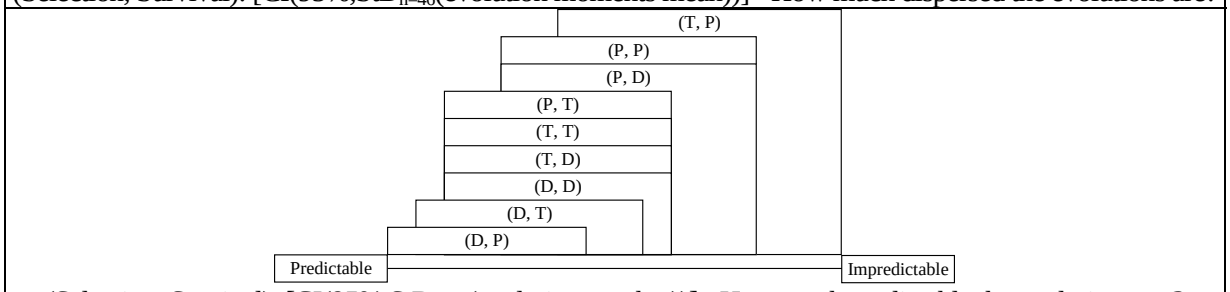
(Selection, Survival): [CI(95%,Mean$_{n=46}$(evolution moments mean))] - How fast the evolutions occurs?



(Selection, Survival): [CI(95%,Mean$_{n=46}$(evolution number))] - How often the evolution occurs?



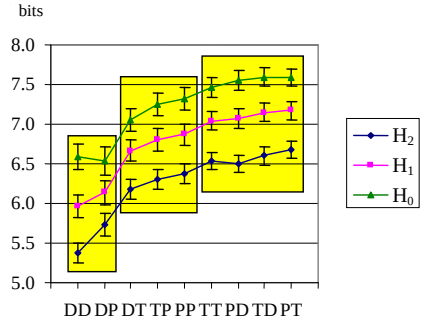
(Selection, Survival): [CI(95%,StD$_{n=46}$(evolution moments mean))] - How much dispersed the evolutions are?



(Selection, Survival): [CI(95%,StD$_{n=46}$(evolution number))] - How much predictable the evolutions are?

Analysis of diversity

The diversity of the genotypes can be quantified by the informational entropy. A family of entropic measures - given by the expression of the generalized (or Rényi) entropy H_α (Rényi, 1961) - are available. H_0, H_1, and H_2 were used to measure the genotipic diversity during evolution (see figure).



If the observations are put together by selection and again by survival strategies, computing again the average

(from 46 experiments) and its confidence interval, the results are like in following table (for H_1).

Strategy of	Means and confidence intervals at 5% risk being in error
Selection	D(6.26±0.10)
	T(7.00±0.07)
	P(7.04±0.07)
Survival	P(6.61±0.10)
	D(6.73±0.12)
	T(6.96±0.08)
D: deterministic; T: tournament; P: proportional	

Measuring agreement between observed distributions

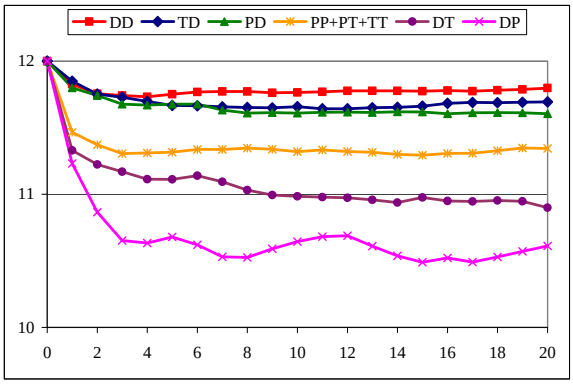
The average number of viable genotypes, number of phenotypes (from a genotype may descent no more than six phenotypes given by the linearization operator L_O, not all viable - last entry in first table) and number of presences of a genotype in a MLR were calculated by thousands of generations (average from 46 experiments) for every pair of survival and selection strategy in order to find answer to the following questions:

÷ In which degree the average number of viable genotypes are (in)dependent of selection and survival strategy?

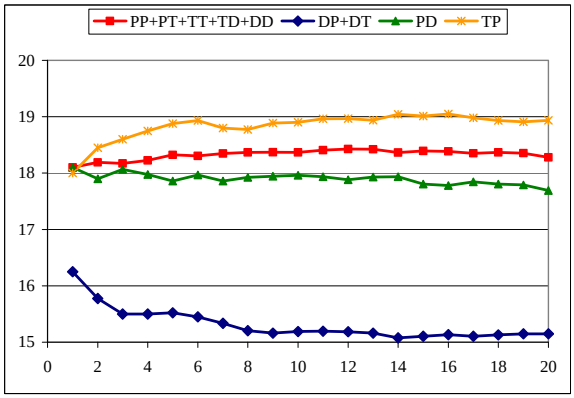
÷ In which degree the average number of phenotypes are (in)dependent of selection and survival strategy?

÷ In which degree the average number of presences of a genotype in a MLR are (in)dependent of selection and survival strategy?

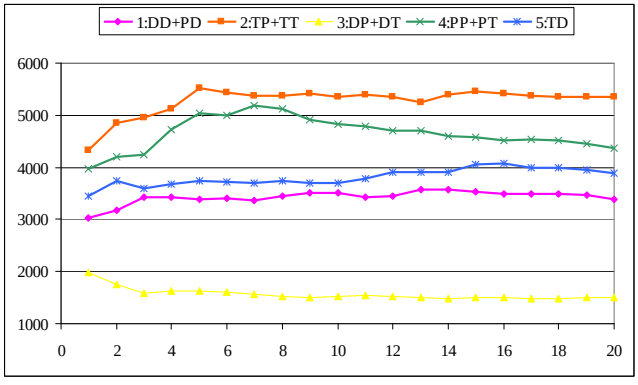
The k-Sample Anderson-Darling test were used to measure the agreement between observations (502 statistical inferences for a research question). Following three figures gives the Monte-Carlo experiments based on the results and evidencing the observed agreements.



Genotypes



Phenotypes



Associations

Distribution of evolution objective's

The recorded data were used to reconstruct the value of determination coefficient in all generations (because between evolutions determination coefficient in a generation is equal with the determination coefficient from previous generation).

First analysis was conducted using a random sample of generations (on which the hypothesis of discrete uniformity was verified) and DataPlot software for likelihood estimation of parameters and statistical agreement between observed data and a pool of 7 probabilitity density functions (PDF).

The sample of 9 generations from discrete uniform distribution 0..20000 was: {9221, 4182, 14283, 15329, 8875, 4599, 994, 8620, 7404}. The pool of PDF was: DE - double exponential, LG - logistic, LN - log-normal, NO - normal, UN - uniform(0,1), EX - exponențial, GU - extreme values of type I (Gumbel), from which UN and EX were easily rejected. Following table gives the statistic of PDFs not rejected at 1% risk being in error by the series of data from the sample of generations.

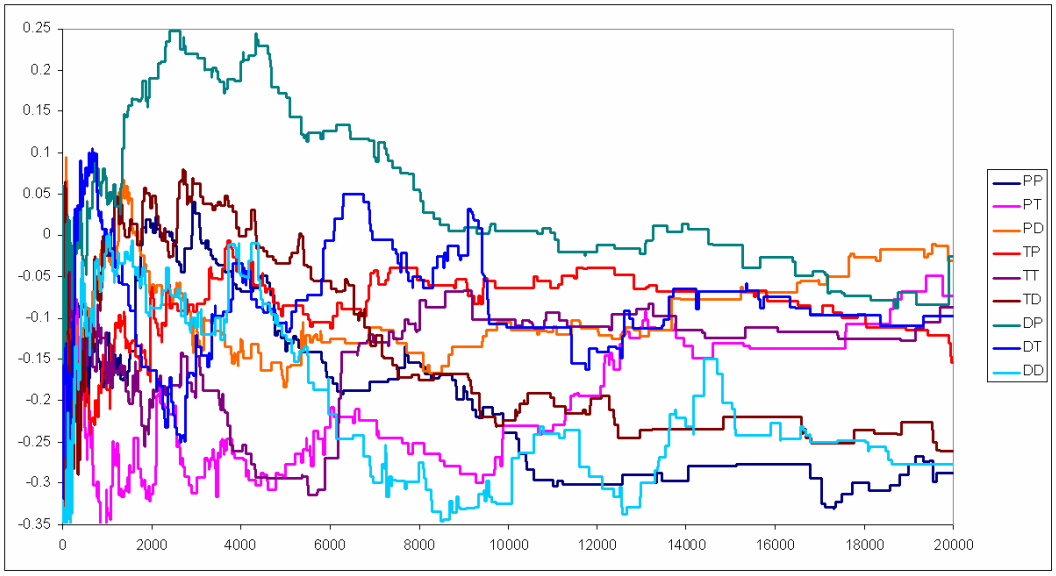
1-5-10 %	PP	PT	PD	TP	TT	TD	DP	DT	DD
DE	9-5-3	9-4-1	9-8-5	8-6-3	7-7-1	9-4-1	6-4-3	5-0-0	3-0-0
LG	9-8-6	9-2-1	9-6-6	9-8-5	9-9-6	8-1-0	6-0-0	6-0-0	1-0-0
LN	9-9-7	7-2-0	9-6-3	9-5-2	9-9-9	7-1-0	0-0-0	4-0-0	2-0-0
NO	9-8-7	7-2-0	9-6-3	9-5-2	9-9-9	7-1-0	0-0-0	4-0-0	2-0-0
GU	9-9-7	3-1-0	9-9-5	9-9-9	9-7-7	9-9-6	9-7-7	9-5-2	2-0-0

At 1% first is GU with 68 (from max 81) and second is LG with 66; at 5% GU with 56, DE with 38; at 10% GU with 43, LG with 24. Taking by strategy, for PP most likely are LN & GU (9-9-7), for PT most likely is DE (9-4-1), for PD most likely is GU (9-9-5), for TT most likely is NO & LN (9-9-9), for TD most likely is DP (9-7-7), for DT most likely is GU (9-5-2) and for DD most likely is DE (3-0-0). Since the distributions were not accepted at a reasonable risk being in error, we must draw the conclusion that the determination coefficient during evolution is not distributed by the models given by the list of seven. Further investigations were made on GU (which is accepted at a reasonable risk being in error by PP, PD, TP, TT, TD, DP and DT - 7 out of 9). An important result derived from the study till this point: *the mean is not a sufficiency statistic*, since normal distribution was rejected to be the population distribution. Another important result were obtained from the study of the GU alternative of distribution: *GU (extreme values of type I) is not general enough* to agree with observed data.

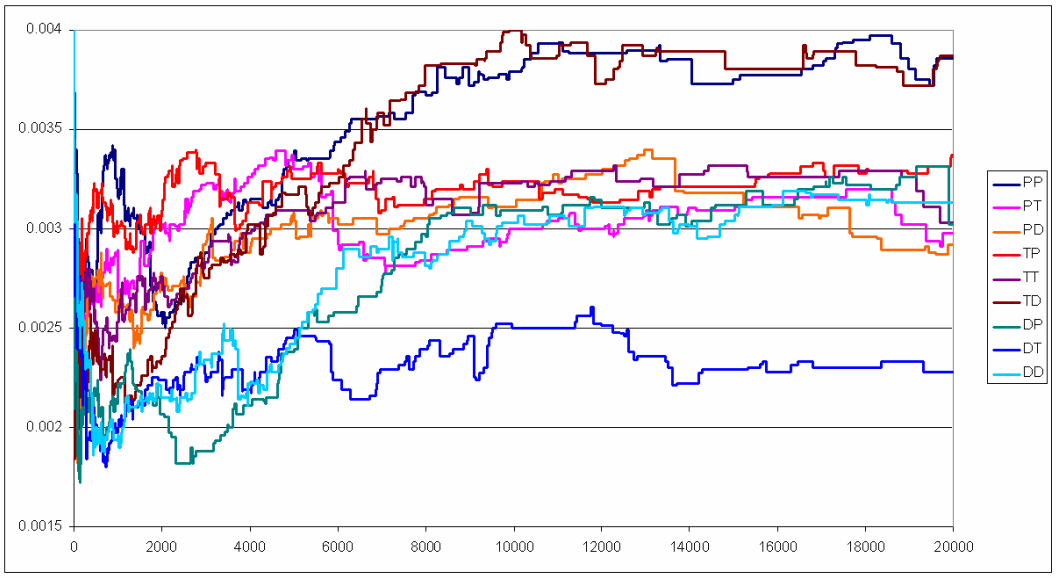
Second analysis was conducted using whole avaliable data (20000 samples - of generations - every which 46 observed values of r^2), a pool of over 50 PDFs as alternatives of distribution, and three statistics (C-S, A-D & K-S) for measuring the agreement with observed data. The pool of distributions were easily shortened at Beta, Johnson, Kumaraswamy, Pert, Power, Reciprocal, Triangular and Uniform (bounded) and Fisher-Tippett, Pareto and Log-Pearson type III (generalized). The study (conducted using EasyFit) shown that the Fisher-Tippett distribution (generalized extreme value) is general enough to agree with observed data in 98.8% of the cases at 1% risk being in error.

$$FT_{PDF}(X)=\begin{cases}\frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}, & k<0 \quad \text{Weibull}\\ \frac{1}{\beta}\exp\left(-\frac{x-\lambda}{\beta}-\exp(-\frac{x-\lambda}{\beta})\right), & k=0 \quad \text{Gumbel}\\ \frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k}, & k>0 \quad \text{Fréchet}\end{cases}$$

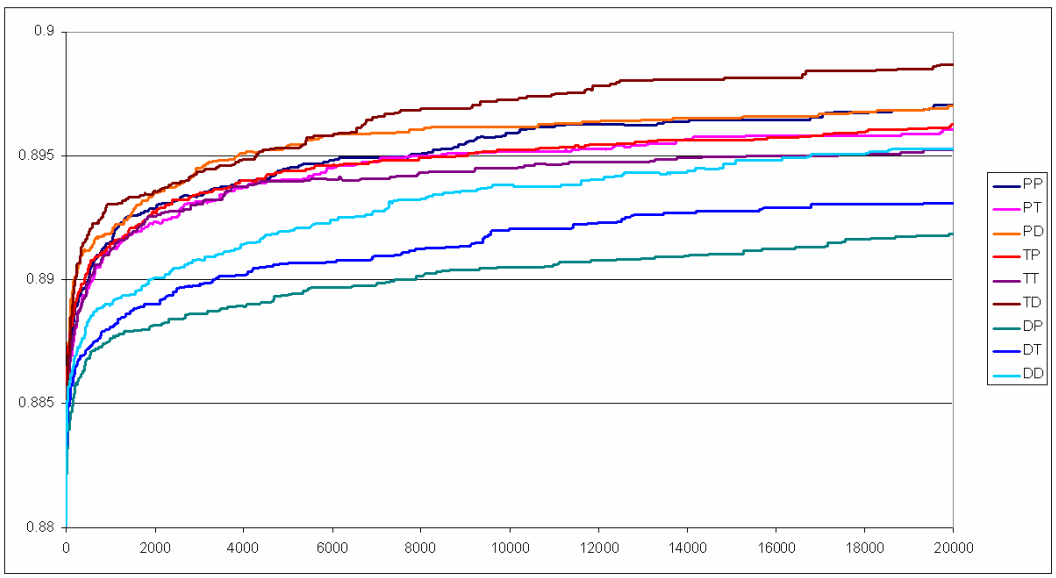
The shape (k), scale (β) and location (λ) parameters of FT distribution were estimated (using MLE) with EasyFitXL for every generation (0..2000) and strategy (PP, PT, PD, TP, TT, TD, DP, DT, DD). Statistica were used for exponential smoothing. Following three figures gives the estimation results (no smooting here); on the abscissa are the generation and on the ordinate are the parameters values.



Shape k=k(G)



Scale $\beta = \beta(G)$



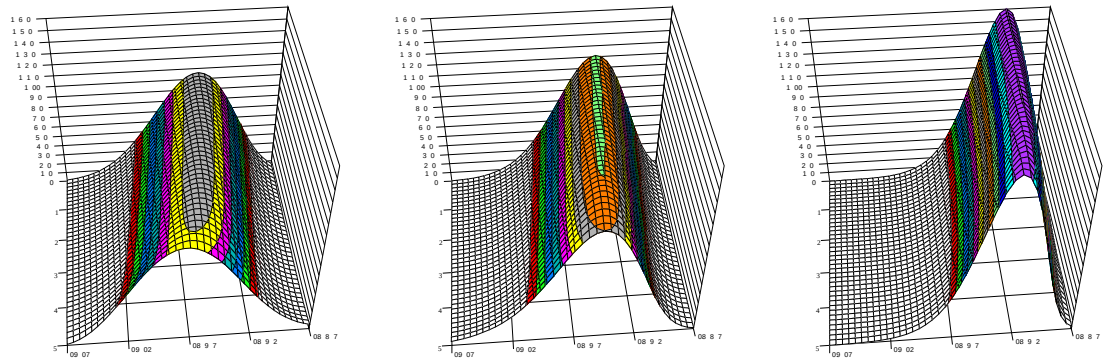
Location $\lambda = \lambda(G)$

FT(r^2; k,β,λ):Fisher-Tippett distribution of evolution's objective

In the obtained results (for k, β, and λ) a search for trend were conducted. Following table gives the obtained results.

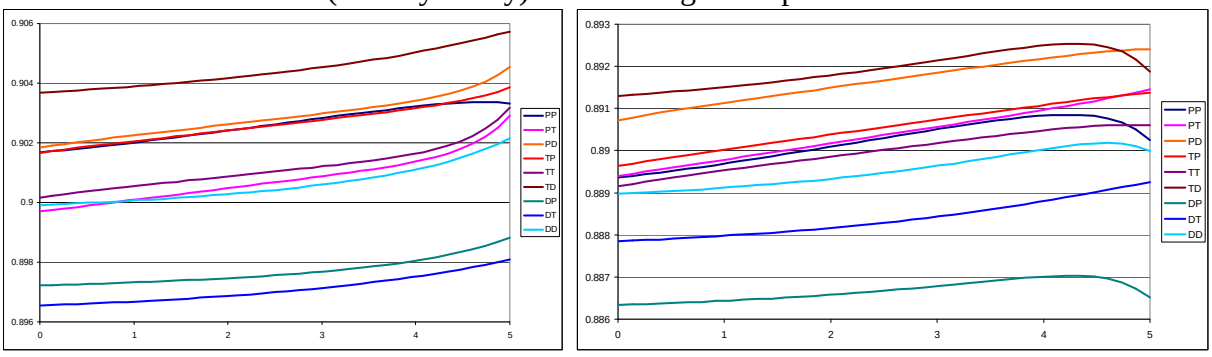
SS	$k(G) = a_0 + a_1 \cdot G$		$\beta(G) = a_0 + a_1 \cdot G$		Trend $\lambda(G)$	a_0	a_1	a_2
PP	-0.1912	$-1.47 \cdot 10^{-6}$	3.541E-3	5.5E-9	$\lambda(G) = a_0 + a_1 \cdot \ln(G+a_2)$	0.89357	$1.82 \cdot 10^{-4}$	0.867
PD	-0.0961	$3.12 \cdot 10^{-7}$	2.983E-3	1.9E-9		0.89422	$1.55 \cdot 10^{-4}$	-0.344
TP	-0.0833	$1.24 \cdot 10^{-7}$	3.192E-3	8.9E-10		0.89333	$1.54 \cdot 10^{-4}$	-0.213
TT	-0.1476	$5.58 \cdot 10^{-7}$	3.072E-3	2.9E-9		0.89286	$1.40 \cdot 10^{-4}$	-0.348
PT	-0.2108	$1.08 \cdot 10^{-6}$	2.996E-3	8.2E-10	$\lambda(G) = a_0 + a_1 \cdot \ln(G)$	0.89309	$1.69 \cdot 10^{-4}$	
TD	-0.1352	$-1.47 \cdot 10^{-6}$	3.419E-3	7.9E-9	$\lambda(G) = a_0 + a_1 \cdot \text{pow}(G,a_2)$	0.89465	$6.84 \cdot 10^{-4}$	0.117
DP	-0.0193	$-1.32 \cdot 10^{-6}$	2.730E-3	7.1E-9		0.88916	$2.02 \cdot 10^{-4}$	0.171
DT	-0.0797	$-1.35 \cdot 10^{-7}$	2.296E-3	6.1E-10		0.89016	$3.19 \cdot 10^{-4}$	0.151
DD	-0.0207	$-9.52 \cdot 10^{-7}$	2.745E-3	5.6E-9		0.89173	$2.93 \cdot 10^{-4}$	0.172

The trend equations for shape, scale and location were used to plot the trend of PDFs. Following figures depicts the PDFs for three strategies (out of nine) where from right to left axis is evolution objective (r^2) and in perspective axis - from 0 to 5 - is $\log_{10}(G)$.



$FT\text{-}PP_{PDF}$ $FT\text{-}TT_{PDF}$ $FT\text{-}DT_{PDF}$

Using again the trend equations for shape, scale and location the 95% and 5% probability borders (from CDF) were obtained. Note that the chance to be upper to 95% border are reserved only for 5% of the observed cases (lucky lottery) and the chance to be below 5% border are reserved for 95% of the observed cases (unlucky lottery). Next two figures depict these borders.



Lucky lottery (CDF = 95%) Unlucky lottery (CDF = 5%)

Using estimations for shape, a statistic regarding the type of the extreme modeled by the Fisher-Tippett distribution were constructed (next table - observed cases and percents).

Type of the extreme	PP	PT	PD	TP	TT	TD	DP	DT	DD
I ($\|k\|<10^{-2}$) ≈ Gumbel	778 (3.9%)	0 (0%)	317 (1.6%)	63 (0.3%)	23 (0.1%)	992 (5%)	3237 (16.2%)	1091 (5.5%)	292 (1.5%)
II ($k>10^{-2}$) = Fréchet	324 (1.6%)	0 (0%)	299 (1.5%)	0 (0%)	36 (0.2%)	2158 (10.8%)	9012 (45.1%)	1619 (8.1%)	0 (0%)
III ($k<-10^{-2}$) = Weibull	18899 (94.5%)	20001 (100%)	19385 (96.9%)	19938 (99.7%)	19942 (99.7%)	16851 (84.3%)	7752 (38.8%)	17291 (86.5%)	19709 (98.5%)

We can note that (in average) in the best case Gumbel is observed below 4%, Fréchet in about 7.5%, and Weibull in over 88% of the cases. Also, the table shows that the DP strategy is the only one with dominance of extreme type II values (Fréchet).

The distribution law for relative moments of evolution

Using generations in which evolution occurs, a transformation like in following table were applied.

Generation number	0	15	136	188	246	528	5423	11887
Evolution moment	1	16	137	189	247	529	5424	11888
Time till the next evolution	15	121	52	58	282	4895	6464	?
Relative time frame	15.00	7.56	0.38	0.31	1.14	9.25	1.19	-
Data from run 1 (of 46) using DP strategy (PCB_5108_evo.txt data file)								

First, an answer for "*Which distribution follows the relative moments of evolution independent on evolution strategy?*" were given. 11347 relative moments of evolution were obtained joining together all 46 runs and 9 strategies. EasyFit software were used having over 65 alternatives for distribution and three statistics (C-S, A-D, K-S) for measuring the agreement with observed data. Following table contain first three distributions sorted by rank of agreement by C-S statistic.

The most probable distribution laws for relative moments of evolution (all data)

Dist\Stat	K-S	p_{K-S}	Rank	A-D	p_{A-D}	Rank	C-S(df)	p_{C-S}	Rank
Log-P-3	0.01197	0.07683	1	2.4264	0.05617	1	41.731(13)	7.3E-05	1
Burr	0.01635	4.57E-03	3	6.7901	3.23E-04	3	46.345(13)	1.25E-05	2
Burr-4P	0.01592	6.27E-03	2	6.0813	7.48E-04	2	51.408(13)	1.71E-06	3
Dist: Distribution law; Stat: Statistic; Rank: Rank of the statistic in the list of 65 alternatives Log-P-3: log-Pearson of type III									

The results from the above table strongly suggest that if there is a distribution law out of the 65 alternatives, then it is LP3 (only C-S rejects the agreement with Log-P-3 at 5% risk being in error; all other distributions are rejected at 5% risk being in error by all three statistics).

Same experiment was conducted for observations coming from a given strategy (nine samples). Agreement with Log-P-3 was measured (table below).

SS	nr.Obs	K-S	p_{K-S}	A-D	p_{A-D}	C-S/df	p_{C-S}
TT	1379	0.02284	0.46	0.63251	0.47	12.3/10	0.27
TD	1429	0.01224	0.98	0.23477	0.75	3.3064/10	0.97
TP	1318	0.02691	0.29	1.2118	0.24	14.35/10	0.16
DT	996	0.02845	0.39	0.73496	0.41	10.628/9	0.30
DD	1084	0.01919	0.81	0.34184	0.66	8.1401/10	0.62
DP	851	0.02416	0.69	0.6234	0.47	6.8598/9	0.65
PT	1463	0.0203	0.58	0.70531	0.43	12.512/10	0.25
PD	1474	0.03055	0.13	0.93998	0.33	8.6564/10	0.56
PP	1353	0.01212	0.99	0.23201	0.75	3.5574/10	0.97
SS (DD, DP, DT, PD, PP, PT, TD, TP, TT): strategy Stat (nr.Obs, K-S, p_{K-S}, A-D, p_{A-D}, C-S(df), p_{C-S}): statistic							

The agreement from table above is excellent - no rejection at 1%, 5% and 10% risk being in error; two rejections at 20% risk being in error (p_{K-S} for PD & p_{C-S} for TP) from 27 cases. Thus, there is no statistical evidence to reject the hypothesis that the relative moments of evolution follow the log-Pearson of type III distribution. More, the agreements from table above correlated with C-S disagreement for all data joined together suggests that *log-Pearson of type III is the distribution law for relative moments of evolution, and its parameters depends on selection and survival strategy*.

Degeneration of log-Pearson type III to uniparametrical for describing relative moments of evolution

The values of shape (α), scale (β) and location (γ) from MLE for all 9 strategies were related one to each other in the series. Following relations were found statistically significant:

÷ $\alpha = 8.77 \cdot \gamma - 68.3$ ($r = 0.994$);

÷ $\beta = -0.14 - 144 \cdot \gamma^{-2.57}$ ($r > 0.999$);

New values for location were obtained from maximization of MLE for $LP3(x; 8.77 \cdot \gamma - 68.3, -0.14 - 144 \cdot \gamma^{-2.57}, \gamma)$. Agreements were measured again using C-S, A-D and K-S statistics for the

new estimations of α, β and γ. The following table gives the new locations, 3-parametrical and uniparametrical Log-P-3 MLE scores and p values measuring agreements in these two cases.

SS	nr.Obs	MLE		p_{K-S}		p_{A-D}		p_{C-S}		$\gamma_{uniparametrical}$
TT	1379	150.1	146.3	0.46	0.09	0.47	0.17	0.27	0.12	17.171
TD	1429	324.0	323.9	0.98	0.98	0.75	0.74	0.97	0.77	16.011
TP	1318	192.9	192.4	0.29	0.30	0.24	0.19	0.16	0.10	12.758
DT	996	-328.5	-335.3	0.39	0.47	0.41	0.52	0.3	0.55	11.640
DD	1084	-72.80	-72.80	0.81	0.88	0.66	0.66	0.62	0.47	36.364
DP	851	-387.4	-390.5	0.69	0.14	0.47	0.15	0.65	0.21	33.160
PT	1463	401.1	401.3	0.58	0.68	0.43	0.46	0.25	0.36	15.347
PD	1474	317.7	316.8	0.13	0.08	0.33	0.24	0.56	0.44	16.216
PP	1353	140.4	140.2	0.99	0.90	0.75	0.64	0.97	0.80	17.180
MLE, p_{K-S}, p_{A-D}, p_{C-S}: first column for three-parametrical, second for uniparametrical										

The analysis results given in the above table give no statistical reason to reject the hypothesis that *the distribution law of relative moments of evolution is a uniparametrical degeneration of log-Pearson of type III distribution and the location parameter is a characteristic of selection and survival strategy chousen.*

Using the values for $\gamma_{uniparametrical}$, mean, mode, median, standard deviation, skewness and kurtosis excess calculated using the obtained probability density functions a principal component analysis of these values were conducted using Statistica software.

The figure below depicts this analysis. The figure reveals relatives between PP & TT and TD & PD strategies.

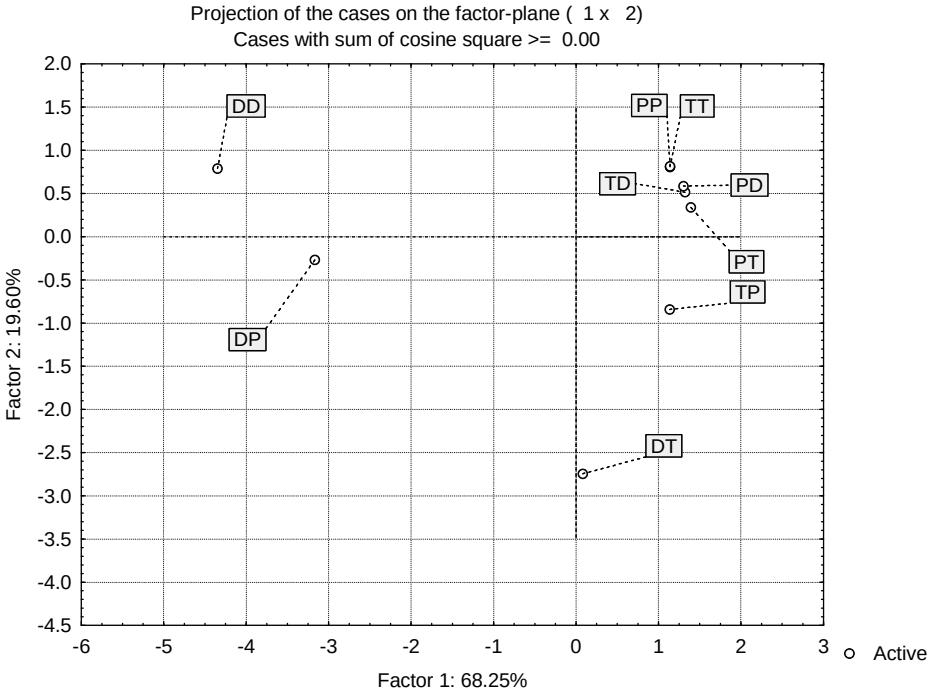


Figure: First two (principal) factors in values of γ, μ, $\hat{\mu}$, $\tilde{\mu}$, σ, γ_1, and γ_2

The distribution law for number of evolutions

Ten statistical experiments were conducted, one with all data together and 9 - one for every strategy separately. Number of evolutions in time frame from generation 0 (initial) to generation 20000 (end of the simulation) in every independent run (46 independent runs) were the observable. Sample of the observable has 46 observations for every strategy separately and 414 observations in all together. EasyFit software was used to conduct the experiment of agreement with the alternatives of distribution. Joining the results of ranks by statistic for all 10 experiments, Fisher-Tippett (Generalized Extreme Values) distribution were recorded with a rank of 284, followed at distance

by the rest of alternatives (over 420 is the rank of the next alternative). Hypothesis of distribution by Fisher-Tippett law was formulated for the number of evolutions. Very good agreements were observed between observations of number of evolutions and Fisher-Tippett distribution law (following table).

Agreement between Fisher-Tippett distribution and number of evolutions in independent runs of GA

Strategy	nr.Obs	K-S	P_{K-S}	A-D	P_{A-D}	C-S/df	P_{C-S}
TT	46	0.0924	0.7931	0.4183	0.6028	5.17/5	0.3956
TD	46	0.1199	0.4859	0.5976	0.4877	3.57/4	0.4671
TP	46	0.0454	0.9999	0.0818	0.8972	0.96/5	0.9661
DT	46	0.0632	0.9873	0.2303	0.7527	1.27/5	0.9381
DD	46	0.0615	0.9906	0.215	0.7665	0.72/5	0.9816
DP	46	0.0954	0.7612	0.2766	0.7127	3.76/4	0.4389
PT	46	0.0712	0.9608	0.2052	0.7754	4.23/5	0.5171
PD	46	0.0634	0.9869	0.1693	0.8090	0.99/5	0.9632
PP	46	0.0665	0.9787	0.2428	0.7417	0.69/5	0.9835
All	414	0.0342	0.7066	0.307	0.6875	7.14/8	0.5218

The agreement of Fisher-Tippett and observed data is very good. There are no rejections of the null hypothesis at risks being in error from 1% to 20%. More, the average agreement measured by K-S statistic is 86.5%, 72.3% when A-D is used, and 71.7% for C-S statistic. Lowest agreement is for TT when C-S is the measure of (39.6%), and for TD when K-S and A-D has lowest p values (48.6% and 48.8 respectively). MLE estimations for shape (k), scale (β) and location (λ) parameters of Fisher-Tippett distributions as well as their common statistics are given in the following table.

Statistics of Fisher-Tippett distributions givingthe number of evolutions to optimum

Strategy	F-T(α; β; γ) distribution	μ	$\hat{\mu}$	$\tilde{\mu}$	σ	γ_1	γ_2
TT	F-T(-0.0771; 8.0028; 26.929)	31.0	28	29.8	9.38	0.739	0.849
TD	F-T(-0.19367; 8.9378; 28.367)	32.1	30	31.5	9.44	0.276	-0.095
TP	F-T(0.04267; 8.7648; 24.208)	29.7	24	27.4	11.93	-1.420	3.975
DT	F-T(-0.0309; 7.0811; 18.775)	22.7	19	21.4	8.74	0.966	1.635
DD	F-T(-0.30349; 9.3813; 21.38)	24.6	25	24.6	9.26	-0.079	-0.289
DP	F-T(-0.27344; 8.0192; 16.622)	19.5	19	19.4	8.05	0.013	-0.280
PT	F-T(-0.15998; 8.6245; 29.02)	32.8	31	32.1	9.35	0.398	0.074
PD	F-T(-0.12837; 9.3279; 28.721)	33.0	30	32.1	10.39	0.520	0.299
PP	F-T(-0.24824; 9.8865; 26.7)	30.4	29	30.2	10.07	0.093	-0.249
All	F-T(-0.16044; 9.6882; 24.161)	28.4	26	27.6	10.50	0.396	0.072
μ: Mean; $\hat{\mu}$: Mode; $\tilde{\mu}$: Median; σ: Standard deviation; γ_1: Asimetry; γ_2: Kurtosis exces							

Results in the table above are close one to each other. In fact was not rejected the hypothesis that all numbers of evolutions come from same population (`All` entry in last two tables). Last table shows that one strategy - TP - has an extreme value of type II (Fréchet) distribution ($\alpha > 0$), all others being of type III (Weibull, $\alpha < 0$).

A variance calculation using the data from the table above ($\sigma_\Sigma^2 = (\sigma_{TT}^2 + ... + \sigma_{PP}^2)/9 = 9.68^2$) allow separation of total variance ($\sigma_{All}^2 = 10.5^2$) in variance inside strategies (9.68^2) and variance between strategies (4.07^2).

Main conclusions

÷ 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 & CDF of the determitation coefficient were obtained; probabilities from CDF to obtain 99% from the optimum in 1000 generations are as follows: 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 creates 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 occurences from 46 runs of above listed, etc), these populations were proof to be gouped and the groups were shown 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 uniparametrical degeneration of log-Pearson of type III curve, and two pairs of relatives (for relative moments of evolution) were found in strategies (PP & TT and TD & 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 using number of evolutions the variance between strategies were found significalty smaller (4.07^2) than the variance inside strategies (9.68^2).

Reprezentative papers published

- On about what Can Be Done and what Cannot Be Done with Genetic Algorithms in Phylogenetic Tree and Gene Sequence Analyses. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Bulletin UASVM, Horticulture* 65(1):63-70, 2008.
- Hard Problems in Gene Sequence Analysis: Classical Approaches and Suitability of Genetic Algorithms. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Biotechnology & Biotechnological Equipment* 23(2):1275-1280, 2009.
- Classical Approaches of Genetic Algorithms and their Suitability. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Asian Journal of Chemistry* 22(3):2275-2284, 2010.
- Distribution Fitting 1. Parameters Estimation under Assumption of Agreement between Observation and Model. Lorentz JÄNTSCHI, *Bulletin UASVM, Horticulture* 66(2):684-690, 2009. ArXiv electronic library permanent link (July 16, 2009): http://arxiv.org/abs/0907.2829 (Subject: Statistics - Methodology).
- Distribution Fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ. *Bulletin UASVM, Horticulture* 66(2):691-697, 2009. ArXiv electronic library permanent link (July 16, 2009): http://arxiv.org/abs/0907.2832 (Subject: Statistics - Methodology).
- Distribution Fitting 3. Analysis under Normality Assumption. Sorana D. BOLBOACĂ, Lorentz JÄNTSCHI. *Bulletin UASVM, Horticulture* 66(2):698-705, 2009.
- Distribution Fitting 4. Benford test on a sample of observed genotypes number from running of a genetic algorithm. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Carmen E. STOENOIU, Mihaela IANCU, Monica M. MARTA, Elena M. PICĂ, Monica ŞTEFU, Adriana F. SESTRAŞ, Marcel M. DUDA, Radu E. SESTRAŞ, Ştefan ȚIGAN, Ioan ABRUDAN, Mugur C. BĂLAN. *Bulletin UASVM, Agriculture* 66(1):82-88, 2009.
- Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Journal of Molecular Modeling* 16(2):377-386, 2010, DOI: 10.1007/s00894-009-0540-z.
- A Study of Genetic Algorithm Evolution on the Lipophilicity of Polychlorinated Biphenyls. Lorentz JÄNTSCHI, Sorana D. BOLBOACĂ, Radu E. SESTRAŞ. *Chemistry and Biodiversity*, 2010, DOI: 10.1002/cbdv.200900356.
- A genetic algorithm for structure-activity relationships: software implementation. Lorentz JÄNTSCHI. ArXiv electronic library permanent link (June 26, 2009): http://arxiv.org/abs/0906.4846 (Subject: Neural and Evolutionary Computing).