Bulletin UASVM Horticulture, 66(1)/2009
Print ISSN 1843-5254; Electronic ISSN 1843-5394

Analysis of the genotypes number in different selection and survival strategies

Lorentz JÄNTSCHI[1], Mugur C. BALAN[2],

[1] Faculty of Materials Science and Engineering, Technical University of Cluj-Napoca
Bd. Muncii 103-105, 400641Cluj-Napoca, Romania; lori@academicdirect.org
[2] Faculty of Mechanics, Technical University of Cluj-Napoca
Bd. Muncii 103-105, 400641Cluj-Napoca, Romania; mugur.balan@termo.utcluj.ro

Abstract. The paper presents the results of a new genetic algorithm applied on a set of polychlorinated biphenyls on searching for structure-activity relationships. Different possible genotypes, resulted by implementing different methods of selection and survival were observed during evolution. The occurrences of the genotypes in the sample allow appreciations concerning their adaptation capacity and in the same time are representing a measure of the sample genetic material variability, induced by the selection and survival methods. Based on the analysis of the results, important and fundamental conclusions were extracted.

Keywords: genetic algorithm, selection strategy, survival strategy, genotypes number

INTRODUCTION

The family of descriptors chosen to evaluate the performances of the developed genetic algorithm is Molecular Descriptors Family (MDF), because of the following considerations:

÷ It is completely developed and proposed by the first author (Jäntschi 2004);
÷ The generating, storing and interrogation system is modern, based on client-server applications with possible parallel processing (Jäntschi 2004);
÷ The method is stable, being revised and fully documented (Jäntschi 2005);
÷ The working tools and the results are available on-line (Jäntschi 2007);
÷ The efficiency of the method was proven in the prediction of physical, chemical and biological properties of more than 50 investigated chemical compounds (Jäntschi and Bolboacă 2007).

The set of molecules chosen for investigation is the series of polychlorinated biphenyls (PCBs), a series formed by 209 compounds. Its study is of high importance for the impact on the ecosystem.

The measured activity is represented by the coefficient of partition octanol/water (K_{ow}), representing the fractions between the concentrations of a chemical compound between octanol and water, at a certain temperature. It is an adimensional parameter, frequently expressed on a logarithmic scale ($\log(K_{ow})$). This phisico-chemical property is used in many studies concerning the environment such as (U.S. Geological Survey, 2008). The values of the measured activity are based on the study of (Eisler and Belisle, 1996) which is a synthesis of many other results reported in the study of PCBs by many authors. The previous study (Jäntschi and Bolboacă, 2006), indicated that it can be obtained a linear multiple regression equation in 4 variables, to explain the measured activity ($\log(K_{ow})$) in percent of 91%, even if that equation is not respecting all the imposed phenotypic viability (variability, deviation from normality and reasonable determination).

The original developed genetic algorithm is described on (Jäntschi, 2009).

The paper analyses the number of genotypes observed during evolution in a series of 46 independent runs of the genetic algorithm for three selection strategies (proportional, tournament, deterministic) and three survival strategies (proportional, tournament, deterministic).

MATERIALS AND METHODS

Together with the choice of the MDF family of descriptors, the set of 206 PCBs molecules and the measured activity ($\log(K_{ow})$), the software of evolution, required a set of working parameters.

Thus, the execution configuration file of the genetic algorithm, count 30 parameters of which 18 are ordinals and 12 contain values of finite and defined lists. Only the last 12 parameters, give a total possible number of combinations to investigate, of 134784. For each of this combinations exist practically an infinity of configuration possibility for the rest of 18 parameters. The large number of possible combinations is discouraging from the point of view of a systematic analysis, but it is also suggesting an extremely large number of possible states of the algorithm, ensuring its variability.

For the algorithm evaluation, it is of both theoretical and practical importance, to compare the performances obtained for two major parameters of the evolution process:

÷ Method giving the individuals for evolution process (selection);

÷ Method giving the individuals to be substituted by descendents (survival).

The influence of the selection method and the survival method on the evolution process is representing the main objective of the study.

Table 1 is presenting the scheme of the experimental design, from the selection and survival methods point of view.

Table 1. Modalities of selection and survival: experimental design

Selection vs. Survival	Proportional (P)	Deterministic (D)	Tournament (T)
Proportional (P)	P:P	P:D	P:T
Deterministic (D)	D:P	D:D	D:T
Tournament (T)	T:P	T:D	T:T

Computers of the P6 (Dual P5) generation, were used for the executions of the software, in the period of January-February 2009. The results were stored in imposed format data files, available for download at the address:

http://l.academicdirec.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/

Table 2 is presenting the configuration and the results data files, considering the experimental design from Table 1.

Table 2. Configuration and results data files

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

The software was executed for 46 times, for each pair of selection and survival method, obtaining different possible "evolutions". Each execution of the software is representing one experiment.

The frequency of the genotypes apparition in the sample during the evolutions, allow appreciations considering their capacity of adaptation and a measure of the variability of the sample's genetic material, induced by the selection and survival methods. In the analysis are considered both the genotypes that appeared more than 23 times (the half of the experiments number) and the total number of genotypes.

The main goal of the research is to test if the number of distinct genotypes obtained, the total number of genotypes obtained, the total number of apparition and the number of participants in regressions, are independent or not, from the selection and survival methods point of view.

In order to reach this objective, the χ2 (Pearson's chi-square) test was applied to each category of the mentioned results. This test is able to check the agreement between observation and hypothesis, independence and homogeneity (Chernoff and Lehmann 1954), (Plackett 1983) and (Fisher 1923).

RESULTS AND DISCUSSION

The information concerning the frequency of genotypes apparition during the evolutions, obtained as result of software execution is presented in Table 3.

Table 3. Frequency of genotypes apparition

Selection	Survival	Reference	NDG	NGA	NPR
Proportional	Proportional	Top 23	13	406	389
		Total	6760	16788	15902
Proportional	Deterministic	Top 23	13	378	371
		Total	8070	18240	17797
Proportional	Tournament	Top 23	6	214	207
		Total	7466	16599	15739
Deterministic	Proportional	Top 23	3	89	72
		Total	3922	10764	9742
Deterministic	Deterministic	Top 23	32	893	893
		Total	4385	13560	13316
Deterministic	Tournament	Top 23	5	152	152
		Total	4965	12504	11572
Tournament	Proportional	Top 23	13	419	405
		Total	6537	16368	15317
Tournament	Deterministic	Top 23	21	714	687
		Total	7964	17700	17331
Tournament	Tournament	Top 23	8	217	213
		Total	7529	17100	16151
NDG: Number of distinct genotypes; NGA: Number of genotypes apparitions; NPR: Number of participants in regressions					

To each category of results was applied the χ^2 test. The resulted observations, presented in table 3, are compared with estimations, presented in brackets, calculated as indicated in (Fisher 1923). Thus was obtained a complex contingence tables, presented as follows, where X^2 and p from χ^2 are also calculated according to (Fisher 1923).

In Table 4 are given the analysis of the hypotheses concerning the contingence between survival and selection methods table for the numbers of genotypes given in Table 3.

Table 4. Genotypes number: Analysis of independence on selection and survival methods

χ^2	P	T	D	Σ	Question and their answer
P	6760 (6665)	7466 (7726)	8070 (7904)	22296	Is the NDG - Total independent of selection and survival methods? - NO $X^2 \cong 70$; $p_{\chi^2}(X^2,4) \cong 2\cdot10^{-14}$
T	6537 (6586)	7529 (7634)	7964 (7810)	22030	
D	3922 (3968)	4965 (4599)	4385 (4705)	13272	
Σ	17219	19960	20419	57598	
P	16788 (16240)	16599 (17084)	18240 (18303)	51627	Is the NGA - Total independent of selection and survival methods? - NO $X^2 \cong 135$; $p_{\chi^2}(X^2,4) \cong 3\cdot10^{-28}$
T	16368 (16095)	17100 (16932)	17700 (18140)	51168	
D	10764 (11585)	12504 (12187)	13560 (13056)	36828	
Σ	43920	46203	49500	139623	
P	15902 (15241)	15739 (16172)	17797 (18025)	49438	Is the NPR - Total independent of selection and survival methods? - NO $X^2 \cong 187$; $p_{\chi^2}(X^2,4) \cong 2\cdot10^{-39}$
T	15317 (15044)	16151 (15963)	17331 (17792)	48799	
D	9742 (10676)	11572 (11328)	13316 (12626)	34630	
Σ	40961	43462	48444	132867	
P	13 (8)	6 (5)	13 (19)	32	Is the NDG - Top 23 independent of selection and survival methods? - NO $X^2(4) \cong 14.6$; $p \cong 6\cdot10^{-3}$
T	13 (11)	8 (7)	21 (24)	42	
D	3 (10)	5 (7)	32 (23)	40	
Σ	29	19	66	114	
P	406 (262)	214 (167)	378 (569)	998	Is the NGA - Top 23 independent of selection and survival methods? - NO $X^2(4) \cong 420$; $p \cong 1\cdot10^{-89}$
T	419 (354)	217 (226)	714 (770)	1350	
D	89 (298)	152 (190)	893 (646)	1134	
Σ	914	583	1985	3482	
P	389 (247)	207 (163)	371 (557)	967	Is the NPR - Top 23 independent of selection and survival methods? - NO $X^2(4) \cong 440$; $p \cong 6\cdot10^{-94}$
T	405 (333)	213 (220)	687 (751)	1305	
D	72 (285)	152 (189)	893 (643)	1117	
Σ	866	572	1951	3389	

The answer at the question "Is there any link between the three series of genotypes numbers?" can be obtained searching on linear relationships. There is an association between number of distinct genotypes (NDG), number of genotypes apparitions (NGA) and number of participants in regressions (NPR). This association is expected, since existing genotypes (NDG) has a given number of apparitions (NGA) which participate in regressions (NPR). The associations were depicted in Figures 1 to 3.

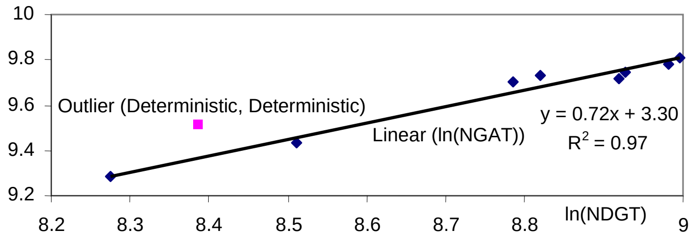

Figure 1. The linear relationship between number of occurrences and number of genotypes

As it can be observed, in all three cases, the deterministic selection and survival is an outlier from the linear relationship between logarithms of the observed genotypes numbers. The fact that deterministic selection and survival is an outlier can be statistically proof.

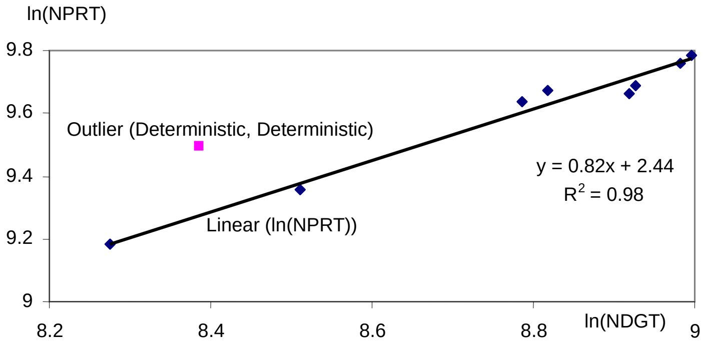

Figure 2. The linear relationship between regression participants and number of genotypes

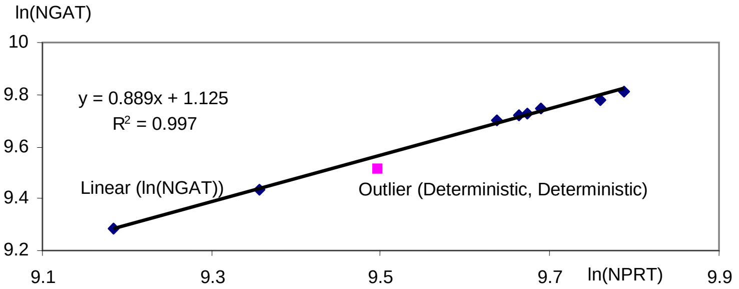

Figure 3. The linear relationship between number of occurrences and regression participants

Student t test (Student, 1908) can be used in order to proof that deterministic selection and survival is an outlier of the regression line (Fisher, 1922). Table 5 give this analysis.

Table 5. Outliers of the regression lines

Figure	Difference	Stdev	t value	Probability
1	0.2212	0.03107	20.14	$2\cdot10^{-7}$
2	0.1469	0.03104	13.38	$3\cdot10^{-6}$
3	0.0494	0.01093	12.79	$4\cdot10^{-6}$
Difference	between natural logarithm of deterministic selection and survival strategy observed and predicted by regression line;			
Stdev	standard deviation of the error of estimate;			
t value	√8·|Difference|/Stdev (known mean of error being 0)			
Probability	to observe a such departure from 0 of the observation error			

Table 5 shows that the observed departures from the regression line of the deterministic selection and survival has, in all three cases, probabilities to be observed below 0.01‰. A possible explanation can be given for this dissimilarity of the deterministic type strategies. Thus, when both selection and survival are deterministic conducted, in the cultivar are constantly kept the best found genotypes and the variability of the genetic material are dramatically reduced.

The confidence interval of the mean and of the standard deviation individually (Table 6, confidence interval for the mean) or simultaneously (Table 7, both confidence intervals, for the mean and for the standard deviation) can be involved to distinguish between different selection and survival strategies based on their observed genotypes number under the assumption that the sampling distribution normalizes the sample around the associated statistic parameter of the population.

Table 6. Statistically significant deviations (at α=5%) in the no. of genotypes (means)

Genotypes	Observable	Mean	CI(95%,Mean)		Out of the range*
Top 23	NDG	12.7	6	20	(T,D), (D,D) > CIU; (D,P), (D,T) < CIL
Top 23	NGA	387	182	592	
Top 23	NPR	377	173	580	
Total	NDG	6400	5183	7617	(D,·) < CIL; (P,D), (T,D) > CIU
Total	NGA	15514	13522	17505	(D,P), (D,T) < CIL; (P,D), (T,D) > CIU
Total	NPR	14763	12697	16829	(D,P), (D,T) < CIL; (P,D), (T,D) > CIU

* (Sel,Srv) - Observation considering selection method *Sel* and survival method *Srv*
(Sel,·) - Observation considering selection method *Sel* and any survival method
(·,Srv) - Observation considering survival method *Srv* and any selection method
CIL, CIU - Lower and Upper limits of the confidence interval of 95%

Table 7. Statistically significant deviations (at α=5%) in no. of genotypes (means and deviations)

Ref.	Obs.	Group: Mean; Std	CI(95%,Mean)		Dev	CI(95%,Dev)		Outside of the CI
Top23	NDG	(P,·): 10.7; 4.0	6	20	9.10	6	17	Dev(P,·)
Top23	NGA	(P,·): 333; 104	182	592	266	180	510	Dev(P,·)
Top23	NPR	(P,·): 322; 100	173	580	265	179	508	Dev(P,·)
Total	NDG	(P,·): 7432; 656	5183	7617	1583	1069	3033	Dev(P,·)
Total	NGA	(P,·): 17209; 898	13522	17505	2591	1750	4963	Dev(P,·)
Total	NPR	(P,·): 16479; 1144	12697	16829	2687	1815	5148	Dev(P,·)
Top23	NDG	(T,·): 14.0; 6.6	6	20	9.10	6	17	-
Top23	NGA	(T,·): 450; 250	182	592	266	180	510	-
Top23	NPR	(T,·): 435; 238	173	580	265	179	508	-
Total	NDG	(T,·): 7343; 731	5183	7617	1583	1069	3033	Dev(T,·)
Total	NGA	(T,·): 17056; 667	13522	17505	2591	1750	4963	Dev(T,·)
Total	NPR	(T,·): 16266; 1012	12697	16829	2687	1815	5148	Dev(T,·)
Top23	NDG	(D,·): 13.3; 16.2	6	20	9.10	6	17	-
Top23	NGA	(D,·): 378; 447	182	592	266	180	510	-
Top23	NPR	(D,·): 372; 453	173	580	265	179	508	-
Total	NDG	(D,·): 4424; 523	5183	7617	1583	1069	3033	Mean(D,·); Dev(D,·)
Total	NGA	(D,·): 12276; 1412	13522	17505	2591	1750	4963	Mean(D,·); Dev(D,·)
Total	NPR	(D,·): 11543; 1787	12697	16829	2687	1815	5148	Mean(D,·); Dev(D,·)
Top23	NDG	(·,P): 9.7; 5.8	6	20	9.10	6	17	Dev(·,P)
Top23	NGA	(·,P): 305; 187	182	592	266	180	510	-
Top23	NPR	(·,P): 289; 188	173	580	265	179	508	-
Total	NDG	(·,P): 5740; 1578	5183	7617	1583	1069	3033	-
Total	NGA	(·,P): 14640; 3363	13522	17505	2591	1750	4963	-
Total	NPR	(·,P): 13654; 3400	12697	16829	2687	1815	5148	-
Top23	NDG	(·,T): 6.3; 1.5	6	20	9.10	6	17	Dev(·,T)
Top23	NGA	(·,T): 194; 37	182	592	266	180	510	Dev(·,T)
Top23	NPR	(·,T): 191; 34	173	580	265	179	508	Dev(·,T)
Total	NDG	(·,T): 6653; 1462	5183	7617	1583	1069	3033	-
Total	NGA	(·,T): 15401; 2521	13522	17505	2591	1750	4963	-
Total	NPR	(·,T): 14487; 2533	12697	16829	2687	1815	5148	-
Top23	NDG	(·,D): 22.0; 9.5	6	20	9.10	6	17	Mean(·,D)
Top23	NGA	(·,D): 662; 261	182	592	266	180	510	Mean(·,D)
Top23	NPR	(·,D): 650; 263	173	580	265	179	508	Mean(·,D)
Total	NDG	(·,D): 6806; 2098	5183	7617	1583	1069	3033	-
Total	NGA	(·,D): 16500; 2560	13522	17505	2591	1750	4963	-
Total	NPR	(·,D): 16148; 2464	12697	16829	2687	1815	5148	-

Analysing Table 6, the following conclusions can be extracted:

- For the deterministic (D) selection:
    - For any survival method, the total number of genotypes is decreasing statistically significant;
    - For the tournament (T) or proportional (P) survival, all the observed parameters (Top 23 and Total, Distinct, Apparitions, Participations) are decreasing;
    - For the Deterministic (D) survival the most frequent genotypes for all parameters (Distinct, Apparitions, Participations) are increasing statistically significant;
- For the deterministic (D) survival:
    - For the tournament (T) or proportional (P) selection, the total number of genotypes is increasing for all the parameters (Distinct, Apparitions, Participations);

The results from table 7 are indicating that:

- The deterministic survival (D) is significantly increasing from statistic point of view, the group of the most frequent genotypes (Top 23) from the generations producing evolutions, while the deterministic selection (D) is significantly decreasing from statistic point of view, the total number of genotypes from the generations producing evolutions;
- Practically each selection method is defining a genotypic population in the generations that are producing evolution. A analysis of variance sustain this result; thus for any studied parameter of the genotypes number (Distinct, Apparitions, Participants in regressions, for Total or Top 23), the total variance is significantly greater than the variance on a given strategy; taking for illustration the number of distinct genotypes, the variances are (entries in Table 7):
    - Total variance: 1583^2 with the confidence interval of 95%: $[1069^2, 3033^2]$;
    - Variance of the population produced by the proportional selection (P): $656^2 < 1069^2$;
    - Variance of the population produced by the tournament selection (T): $731^2 < 1069^2$;
    - Variance of the population produced by the deterministic selection (P): $523^2 < 1069^2$;
- A different conclusion can be extracted, concerning the survival method, for which is produced population segregation. Only the deterministic survival (D) is creating a population with an average number of genotypes significantly statistic, higher than the proportional (P) and tournament (T) survival methods.

CONCLUSIONS

It can be remarked that the confidence in the dependence of the genotypes number on selection and survival strategy is increasing in the order: Number of distinct genotypes (NDG), number of genotypes occurrences (NGA), Number of genotypes participants in regressions (NPR). The number of observations in not increasing in the same indicated order.

Figures 1 to 3 describes the mechanistic of the determination between genotypes presence (NDG), their frequency (NGA), and their phenotypic association (NPR), on which deterministic selection and survival strategy makes a clear distinction to the rest of strategies.

The following major conclusion can be extracted:

- For the deterministic selection, for any survival method, the total number of genotypes is decreasing statistically significant.
- For the deterministic selection, for the tournament and proportional survival, all the observed parameters (Distinct, Apparitions, Participations) are decreasing.
- For the deterministic selection and the deterministic survival, the number of the number of the most frequent genotypes for all parameters (Distinct, Apparitions, Participations) is increasing statistically significant.

÷ For the deterministic survival and for the tournament or proportional selection, the number of genotypes is increasing for all the parameters (Distinct, Apparitions, Participations).

÷ The deterministic survival is significantly increasing from statistic point of view, the group of the most frequent genotypes from the generations producing evolutions, while the deterministic selection is significantly decreasing from statistic point of view, the total number of genotypes from the generations producing evolutions.

÷ Each selection method is defining a genotypic population in the generations that are producing evolution

REFERENCES

Eisler, R. and A. Belisle (1996). Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review. Contaminant Hazard Reviews. Biological Report 31.

Chernoff, H and E. L. Lehmann (1954). "The use of maximum likelihood estimates in $\chi 2$ tests for goodness-of-fit". The Annals of Mathematical Statistics 25: 579-586.

Plackett, R.L. (1983). "Karl Pearson and the Chi-Squared Test". International Statistical Review 51 (1): 59-72.

Fisher, R. A. (1922). The Goodness of Fit of Regression Formulae and the Distribution of Regression Coefficients. Journal of the Royal Statistical Society 85:597-612.

Fisher, R. A. (1923). Studies in Crop Variation. II. The Manurial Response of Different Potato Varieties. Journal of Agricultural Science 13:311-320.

Jäntschi, L. (2004). MDF - A New QSAR/QSPR Molecular Descriptors Family. Leonardo Journal of Sciences 3(4):68-85.

Jäntschi, L. (2004). Delphi Client - Server Implementation of Multiple Linear Regression Findings: a QSAR/QSPR Application. Applied Medical Informatics 15(3-4):48-55.

Jäntschi, L. (2005). Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. Leonardo Electronic Journal of Practices and Technologies 4(6):76-98.

Jäntschi, L. (2007). http://l.academicdirect.org/Chemistry/SARs/MDF_SARs/

Jäntschi, L., Bolboacă, S. (2007). Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships, International Journal of Molecular Sciences 8(3):189-203.

Jäntschi, L. (2009). A genetic algorithm for structure-activity relationships: software implementation. Manuscript:
http://arxiv.org/abs/0906.4846 (abstract), http://arxiv.org/pdf/0906.4846 (PDF).

Student (Gosset, W. S.). (1908). The probable error of a mean. Biometrika 6(1):1-25.

U.S. Geological Survey (2008). Octanol-Water Partition Coefficient (K_{OW}). U.S. Department of the Interior. http://toxics.usgs.gov/definitions/kow.html (Page Last Modified: Thursday, 13-Mar-2008 13:25:59 EDT).