Chem Biol Drug Des 2013; **82:** 603–611

Research Article

Supervised Evolution: Research Concerning the Number of Evolutions that Occur Under Certain Constraints

Lorentz Jäntschi[1] and Sorana D. Bolboacă[2,*]

[1]Department of Physics & Chemistry, Technical University of Cluj-Napoca, 103-105 Muncii Boulevard, Cluj-Napoca, 400641, Romania
[2]Department of Medical Informatics and Biostatistics, "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 6 Louis Pasteur, Cluj-Napoca, 400349, Romania
**Corresponding author: Sorana D. Bolboacă, sbolboaca@umfcluj.ro*

It is known that evolution may lead to a new species while adaptation may lead to a new variety. In this manuscript, we present an analysis of the number of evolutions (defined by improvement of the score associated with an objective function of a genetic algorithm) in an experiment supervised by a genetic algorithm, experiment conducted on octan-1-ol/H20 partition coefficient of polychlorinated biphenyls. The numbers of evolutions resulted from 9 implemented evolution strategies were investigated. Evolutions arisen from the first 20 000 generations coming from 46 independent runs were recorded. A distribution analysis has been conducted for each evolution strategy. Without exception, the Weibull distribution fits well with the number of evolutions at a significance level of 5% for any evolution strategy. Furthermore, the Weibull distribution could not be rejected when different merged samples were investigated.

Key words: genetic algorithm, number of evolutions, supervised evolution

Received 5 April 2013, revised 28 May 2013 and accepted for publication 17 June 2013

Evolution is a process which refers to populations not to individuals and could be seen as the process that results in heritable changes spread over generations (1). The model of selection and fitness introduced by Darwin (1) inspired the evolutionary processes implemented in genetic algorithms (2,3). Numerical simulations of genetically explicit evolutionary processes could be considered as a valuable tool for study of evolution (4–7).

Studies dealing with use of genetic algorithms in structure–activity relationship modeling [methods able to establish functional links between the structure of chemical compounds and the associated physical–chemical properties (SPRs) or biological activities (SARs) (8)] are reported in literature. Generally, the algorithm effectiveness (speed required to achieve the imposed objective function), mutation, and cross-operators have been investigated and reported (9–11), besides identification of the optimum solution (12,13). Furthermore, some studies related to genetic algorithms were conducted regarding the correlation coefficient (14), evolution (15), validation of number of genotypes present in the generations when an evolution occurred (16), and assessment of the distribution law for the relative moments of evolution (17).

The aim of present research was to analyze the behavior of number of evolutions expressed as an increase in determination coefficient of linear regression model applied to the structure–activity relationship for the octan-1-ol/H_2O partition coefficient of polychlorinated biphenyls under imposed selection and survival strategies of a genetic algorithm. Exploiting the advantage of molecular descriptors family as a huge population of molecular descriptors, population of which structure is naturally constructed of genetic type, the experiment of evolution was conducted keeping constant all parameters excepting strategy of selection and strategy of survival to isolate the variability induced by the strategy. Proportional, tournament, and deterministic were used as selection and survival strategies, generating nine pairs of strategies for evolution.

Methods and Materials

An evolution experiment supervised by a genetic algorithm (16) using the relationships between MDF [molecular descriptors family (18)] descriptors (as genetic material) and the octan-1-ol/H_2O partition coefficient of PCBs [Polychlorinated Biphenyls (19)] was conducted to achieve the aim of the research. The workflow of the study is presented in Figure 1 while details about genetic algorithm implementation could be found in (16). Three strategies (p = proportional – the chance is proportional with the value of score function; T = tournament – two genotypes randomly drawn are the candidates for selection and the one with the highest value of the score function is chosen; and

 doi: 10.1111/cbdd.12190

D = deterministic – each time the genotypes with the higher scores are selected) were used as both selection and survival functions, resulting in nine selection-survival strategies investigated (TT, TD, TP, DT, DD, DP, PT, PD, and PP). Three objective functions supervise the evolution: (i) minimum of the determination coefficient from regressions of a certain genotype from the sample as score for selection (r^2_{min}); (ii) $[|r^2_{min}(X_j) - r^2_{min}(X_k)| + NCD(X_j, X_k)/NC]\cdot 0.5$ as score for survival (where X_j, X_k are genotypes, *NCD* is the number of distinct genes, *NC* is the total number of genes (6 in this experiment); and (iii) the determination coefficient (r^2) of the best regression with genotypes from the sample of a generation as score for evolution.

The first letter in the selection-survival strategy refers to the selection, while the second refers to the survival. The selection score was defined as the highest value from the minimum of adaptation, defined as a linear expression of phenotypes association. The objective of this experiment was to maximize the determination coefficient of the regression models (evolution score).

The initial genetic sample, represented by MDF descriptors, comprised 12 genotypes (descriptors). Four genotypes were selected from the initial sample for crossover and mutation. The mutation appears with a probability of 5% before and after crossover. As survival strategy, four genotypes resulted after crossover and/or mutation were selected for replacement in each generation.

The numbers of evolutions on the first 20 000 generations obtained in 46 independent runs are presented in Table 1. The evolution occurred whenever an improvement of the evolution score was obtained (defined in this experiment as improvement of determination coefficient).

The following steps were applied to analyze the number of evolutions:

- Chi-square test on selection-survival contingency to identify if selection and/or survival strategies contribute to the pathway of the number of evolutions (factored). The proposed algorithm runs if the hypothesis of association could not be rejected ($Obs_{c,r} = a_c \cdot b_r \sim Est_{c,r}$, and $\Sigma_r Obs \cdot \Sigma_c Obs / \Sigma_{c,r} Obs \sim Est_{c,r} \sim (a_c \cdot b_r)$, where *Obs* = observed value, *Est* = estimated value, *a, b* = factors in the contingency (in this experiment one factor is represented by selection strategy and the other factor is represented by survival strategy); *c* = refers the column, *r* = refers the row, $\Sigma_c Obs$ = sum of the c^{th} column; $\Sigma_r Obs$ = sum of the r^{th} row; $\Sigma_{c,r} Obs$= sample size in the contingency table (20).

- ANOVA test to identify the sources of variance (selection strategy and/or survival strategy) in the number of evolutions (21).

- Distribution analysis on the number of evolutions. A previous analysis identified that in 9 of 10 cases, the number

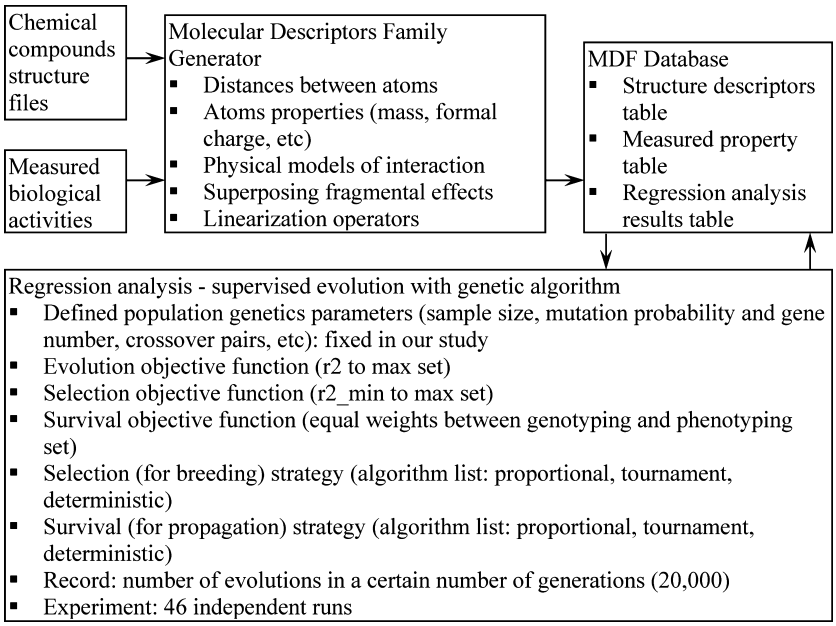

Figure 1: Schematic of the experimental workflow. The geometry of chemical compounds of PCBs was used as input to the computation of MDF descriptors as independent variable (data stored in the MDF database). The linearity between dependent variable represented by octan-1-ol/H_20 partition coefficient and MDF descriptors was of interest in this study. The supervised evolution of the regression analysis was conducted with genetic algorithm in a constrained environment with 12 genotypes in the sample. Two pairs of two genotypes were selected for crossover and mutation (probability of 5%), and descendants competed for survival in each generation. The algorithm runs 46 times for 20 000 generations, and the number of evolutions defined as improvement of the determination coefficient was recorded and evaluated.

Table 1: Number of evolutions (increase in the determination coefficient) on the first 20,000 generations obtained in 46 independent runs according to selection-survival strategy

Run	TT	TD	TP	DT	DD	DP	PT	PD	PP	Run	TT	TD	TP	DT	DD	DP	PT	PD	PP
1	18	27	13	13	18	8	53	23	38	24	32	21	27	11	47	20	14	25	29
2	28	28	41	14	20	26	21	37	20	25	22	36	48	18	26	15	21	41	27
3	27	33	29	31	25	14	30	31	39	26	18	30	29	22	21	25	26	47	18
4	28	19	17	18	37	27	17	28	37	27	38	29	30	20	33	14	32	33	37
5	29	43	14	33	28	23	21	37	24	28	19	39	35	5	19	25	34	33	19
6	21	50	18	26	9	26	28	25	29	29	51	39	19	23	26	35	27	57	33
7	20	39	37	35	26	21	29	40	29	30	38	52	44	21	19	11	45	29	27
8	24	39	21	24	24	18	36	24	21	31	24	30	21	22	27	12	35	23	15
9	38	35	45	13	34	22	31	29	41	32	19	35	21	36	33	7	33	41	13
10	19	40	23	15	37	20	18	39	62	33	37	21	24	16	34	26	27	20	22
11	37	20	16	28	12	25	32	34	34	34	33	23	31	44	7	31	28	43	14
12	26	36	25	28	27	25	34	39	34	35	28	43	65	16	32	14	33	30	31
13	36	27	39	22	7	15	40	50	34	36	37	40	23	15	15	15	58	29	33
14	49	21	26	16	33	6	23	19	44	37	31	17	32	24	29	28	34	36	24
15	37	45	40	20	33	5	41	21	28	38	26	41	33	25	14	23	45	21	39
16	22	25	37	13	21	11	35	51	29	39	28	24	18	44	19	30	36	34	41
17	57	42	28	34	30	32	27	29	47	40	27	47	51	25	23	11	24	27	9
18	25	34	24	19	28	29	46	50	20	41	32	21	19	16	35	22	27	30	26
19	30	20	24	24	12	17	37	15	22	42	20	23	57	22	23	15	27	36	25
20	24	31	29	24	23	7	43	46	36	43	48	24	22	14	20	24	26	45	41
21	32	26	39	18	21	19	37	19	46	44	39	40	27	37	9	11	35	30	35
22	38	26	32	19	30	16	42	18	34	45	35	33	31	28	25	35	39	27	32
23	38	21	15	33	43	15	48	48	35	46	40	40	25	18	16	21	34	31	26

P, proportional; T, tournament; and D, deterministic.

of evolutions follows the Weibull distribution (22). In this research, the maximum likelihood estimation (MLE) was applied to obtain population's parameter for the Weibull distribution.

• Verify the linearity between parameters of Weibull distribution. This step was introduced as dependence between parameters of probability distribution function was previously identified (17).

• Optimize the parameters and test the agreement between the new models and observation using Kolmogorov-Smirnov (23,24), Anderson-Darling (25), and Chi-squared (26) statistics, whenever linearity exists. Evaluate the overall agreement of linearity with F-C-S (27).

• Compare selection strategies, survival strategies, and selection-survival strategies in terms of differences between speciation produced by a certain strategy and speciation produced by another strategy. For any two Weibull distributions associated with the number of evolutions, DWeibull [Eqn (1)] represents the positive difference of probability associated with the *j* strategy relative to the *k* strategy, while PDWeibull [Eqn (2)] represents the difference of the probability.

$$\mathrm{DWeibull}_{j,k}(x) = \begin{cases} 0, \mathrm{Weibull}(x;\alpha_j, a\cdot\alpha_j + b) \le \mathrm{Weibull}\ (x;\ \alpha_k,\ a\cdot\alpha_k + b) \\ \mathrm{Weibull}\ (x;\ \alpha_j,\ a\cdot\alpha_j + b) - \mathrm{Weibull}\ (x;\ \alpha_k,\ a\cdot\alpha_k + b), \\ \quad \text{otherwise} \end{cases} \tag{1}$$

$$\mathrm{PDWeibull}_{j,k} = \int_0^\infty \mathrm{DWeibull}_{j,k}(x) \tag{2}$$

where DWeibull = positive difference of the *j* evolution strategy probability relative to *k* evolution strategy; PDWeibull = positive difference of probability density function of *j* evolution strategy relative to *k* evolution strategy; *x* = random variable – number of evolutions; α = shape; $a\cdot\alpha+b$ = estimated scale of Weibull distribution (under the hypothesis of linearity between shape and scale).

Similarly, the mean difference between evolutions when *j* strategy is compared with *k* strategy could be calculated using the formula presented in Eqn (3).

$$\mathrm{MDWeibull}_{j,k} = \int_0^\infty x\cdot\mathrm{DWeibull}_{j,k}(x) \tag{3}$$

where MDWeibull = mean difference of number of evolutions with *j* and *k* strategies; *x* = random variable – number of evolutions.

Identify the properties associated with the distribution of number of evolutions. Three properties were of interest: skewness, excess kurtosis, and information entropy. The Shannon's information entropy (28) associated with obtained probability distribution function (Weibull distribution) was calculated using the formula presented in Eqn (4).

$$H = \gamma\left(1 - \frac{1}{\alpha}\right) + \ln\left(\frac{\beta}{\alpha}\right) + 1 \quad (4)$$

where γ = Euler-Mascheroni constant (0.57721…) (29); α = shape; β = scale.

Results and Discussion

The first step in the analysis was to identify if there was any association between selection and survival strategies when the criterion was number of evolutions. The summary of the numbers of evolutions in the contingency of selection-survival strategies when the octan-1-ol/H_2O partition coefficient of PCBs was investigated is presented in Table 2.

The results presented in Table 2 revealed that the factorization hypothesis of number of evolutions as function of selection and survival strategies could not be rejected (p-value = 0.0651, Table 2). The ANOVA test was applied to identify the source of variation regarding the number of evolutions (selection and/or survival strategy) and the results are presented in Table 3.

As could be observed from results presented in Table 3, both selection and survival strategies proved to have a significant influence on the number of evolutions with a higher influence of the selection strategy. The contribution of the survival strategy to the number of evolutions was one tenth of the contribution of the selection strategy (see the values of the mean squares – Table 3).

As Table 2 revealed the existence of the dependency between selection and survival strategies, the presence of association was explored by applying the following steps:

• (start point) It is known from previous results that evolution depends on selection and survival strategies and that the numbers of evolutions follow the Weibull distribution (22). Starting with this information, the population's parameters for samples of paired selection or survival strategies as well as for merged samples (selection and survival) were calculated and are presented in Table 4. Thus, classical MLE estimates of the population characteristics [Weibull shape (α) and scale (β)] were calculated for each sample: $(\alpha_i, \beta_i)_{1\le i\le 16}$, given in Table 4 (two MLE parameters estimated for each sample).

Table 3: ANOVA test results for selection and survival strategies

Source of variation	SS	df	MS	*F*-value[a]	p-value
Selection strategy	366866.9	2	183433.4	140.192	0.000198
Survival strategy	37586.89	2	18793.44	14.3632	0.014939
Error	5233.778	4	1308.444		
Total	409687.6	8			

SS, sum of squares; df, degrees of freedom; MS, mean squares; *F*-value = Fisher's statistics; p-value, significance; Selection: T_: m (arithmetic mean)=1421, var(variance)=3090; D_: m = 1023, var=13843; P_: m = 1476, var=4477; P, proportional; T, tournament; and D, deterministic; Survival: _T: m = 1325, var=61972; _D: m = 1375, var=45525; _P: m = 1220, var = 78553.
[a]*F* critical value at 5% confidence level (equal with 6.944276).

• The linearity analysis between Weibull distribution parameters was conducted and the hypothesis of linearity between shape and scale could not be rejected at a significance level of 5% ($\beta' = 8.5 + 6.95\cdot\alpha$, $r^2 = 0.74$ – where β' = estimated scale, α = shape), and this linear association was assumed for the next steps.

• (end-point) Under the supposition of the linear dependence, new MLE estimates should be calculated; if (α_i, β_i) are maximum likelihood estimates for Weibull (α,β) from Obs_i under assumption of independence (for i from 1 to 16) then, under assumption of dependence ($\beta = a\cdot\alpha + b$), the new MLE estimates for a, b, and α_i (for i from 1 to 16) must come from:

$$\text{MLE}\Sigma = \sum_{i=1}^{16} \text{MLE}_i = \sum_{i=1}^{16}\sum_{j=1}^{n_i} \text{Weibull}(x; \alpha_i, a\cdot\alpha_i + b) \rightarrow \max.$$

where MLE = maximum likelihood estimation; n = sample size; α = shape parameter of the Weibull distribution; a = intercept of the shape; b = intercept of simple linear regression between shape and scale.

Table 2: Number of evolutions at the contingency of selection and survival strategies: observed values (Obs), expected values (Exp), and $\Sigma_{r,c}(\text{Obs-Exp})^2/\text{Exp}$

Obs	_T	_D	_P	Exp[a]	_T	_D	_P	χ^{2} [b]	_T	_D	_P
T_	1425	1475	1364	T_	1441.5	1495.5	1326.9	T_	0.2	0.3	1.0
D_	1042	1130	897	D_	1037.5	1076.4	955.1	D_	0.0	2.7	3.5
P_	1509	1520	1399	P_	1497.0	1553.1	1378.0	P_	0.1	0.7	0.3

T_, D_, P_ = selection strategy; _T, _D, _P = survival strategy; P, proportional; T, tournament, and D, deterministic; Obs, observed evolution; Exp, expected evolution.
[a]$(\text{Exp}=(\Sigma_r\text{Obs})(\Sigma_c\text{Obs})/(\Sigma_{r,c}\text{Obs})$, r = rows, c = column).
[b]χ^2(Chi-squared test) = $\Sigma_{r,c}(\text{Obs-Exp})^2/\text{Exp}$ = 8.8; $p^{\chi^2}(8.8,4)$=6.51%.

The shape and scale Weibull's parameters were optimized along with regression coefficients (we should call this combined MLE). The results of the agreement analysis conducted between new obtained models and observations are presented in Table 5.

Thus, for independence (Table 4), 16 × 2 = 32 parameters were estimated using MLE and, for dependence (Table 5), only 16 + 2 = 18. From the new estimates of the population parameters (α_i, $a{\cdot}\alpha_i + b$), statistics of the agreement between observation and the model were calculated (A-D, K-S, C-S in Table 5).

Table 4: Weibull population's parameter from samples of paired selection and survival strategies and from merged samples

No (i)	Strategy	n_i	Weibull$(x;\alpha,\beta)=\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}\cdot\exp\left(-\left(\frac{x}{\beta}\right)^{\alpha}\right)$
1	TT	46	Weibull (x; 4.0192, 33.524)
2	TD	46	Weibull (x; 3.9073, 34.947)
3	TP	46	Weibull (x; 3.2206, 32.182)
4	DT	46	Weibull (x; 2.8911, 25.01)
5	DD	46	Weibull (x; 2.6152, 27.349)
6	DP	46	Weibull (x; 2.4423, 21.771)
7	PT	46	Weibull (x; 4.0584, 35.527)
8	PD	46	Weibull (x; 3.7579, 35.967)
9	PP	46	Weibull (x; 3.2433, 33.27)
10	T_	138	Weibull (x; 3.7631, 33.889)
11	D_	138	Weibull (x; 2.7172, 24.853)
12	P_	138	Weibull (x; 3.7449, 35.27)
13	_T	138	Weibull (x; 3.2447, 31.937)
14	_D	138	Weibull (x; 3.1577, 33.249)
15	_P	138	Weibull (x; 2.6709, 29.544)
16	__	414	Weibull (x; 3.0274, 31.693)

Stra = strategy; n = sample size; __=∑(all observations from both complete and incomplete selection-survival strategies); – observation independent by strategy; x = random variable – number of evolutions; α = shape; β = scale; P, proportional; T, tournament; and D, deterministic.

The results of combined Fisher chi-square (F-C-S) test showed that the data follow the Weibull distribution (χ^2 = 30.32, df = 48, p = 0.9783). Probability distribution functions of investigated pairs of selection-survival strategies are presented in Figures 2 and 3.

The analysis of Figure 1 reveals relatively compact groups in populations that are not deterministic in selection. Besides this group, the subpopulations that comprise deterministic selection with all investigated survival strategies can be observed. Similarly, Figure 2 shows a compact group that excluded deterministic selection and proportional survival strategies, while the subpopulation that comprises the observations from all strategies is on a median position between the compact group and its outliers. It is important to appreciate the difference between speciation produced by a strategy against speciation produced by another strategy. A measure of this difference could be given by the difference in probabilities reported to the higher values; this approach was applied in this study to compare strategies. As could be seen from Figures 2 and 3, whenever such a difference exists it represents the difference of probability relative to the higher values.

Distinct properties of Weibull distribution for number of evolutions have been identified as functions of shape parameter, and the results are presented in Figures 3–5.

The linearity between shape and scale allowed to estimate the skewness and excess kurtosis for any selection and

Table 5: Statistics of agreements for optimized shape and scale Weibull parameters from combined MLE

	Weibull (x, α, β')		Kolmogorov-Smirnov		Anderson-Darling		Chi-Squared	
Stra	α	β'	Stat	p-value	Stat	p-value	Stat (df)	p-value
TT	3.46	34.42	0.11413	0.5486	0.75038	0.4329	1.3325 (3)	0.7214
TD	3.61	35.68	0.09594	0.7550	0.56791	0.5363	2.3273 (4)	0.6758
TP	3.25	32.69	0.09481	0.7673	0.82647	0.3954	2.9440 (5)	0.7086
DT	2.49	26.20	0.12812	0.4030	0.92734	0.3502	5.5181 (4)	0.2381
DD	2.69	27.96	0.08054	0.9031	0.33530	0.6985	2.6045 (5)	0.7607
DP	2.16	23.47	0.11331	0.5577	0.84605	0.3862	5.9434 (4)	0.2034
PT	3.67	36.23	0.10038	0.7052	0.40711	0.6445	4.0939 (4)	0.3935
PD	3.72	36.66	0.10689	0.6306	0.42194	0.6338	1.7556 (5)	0.8818
PP	3.37	33.73	0.06154	0.9906	0.18747	0.8214	0.1631 (5)	0.9995
T_	3.44	34.26	0.07783	0.3549	1.21300	0.2468	9.9073 (7)	0.1939
D_	2.43	25.74	0.07629	0.3790	0.97770	0.3294	5.0953 (7)	0.6483
P_	3.60	35.61	0.06371	0.6068	0.46426	0.6041	3.8776 (7)	0.7938
_T	3.18	32.08	0.07338	0.4270	0.51944	0.5671	9.5454 (7)	0.2158
_D	3.32	33.25	0.05651	0.7484	0.47152	0.5991	4.9757 (7)	0.6629
_P	2.84	29.15	0.04054	0.9703	0.46618	0.6027	5.3000 (7)	0.6234
__	3.12	31.56	0.04364	0.3983	0.90044	0.3618	6.6673 (8)	0.5729

Stra = strategy, Stat = statistics; p-value = probability; df = degrees of freedom; x = random variable – number of evolutions; α = shape; β' = estimated scale ($\beta' = 8.5 + 6.95{\cdot}\alpha$); P, proportional; T, tournament; and D, deterministic.
Probability of observation equal to 97.83%.

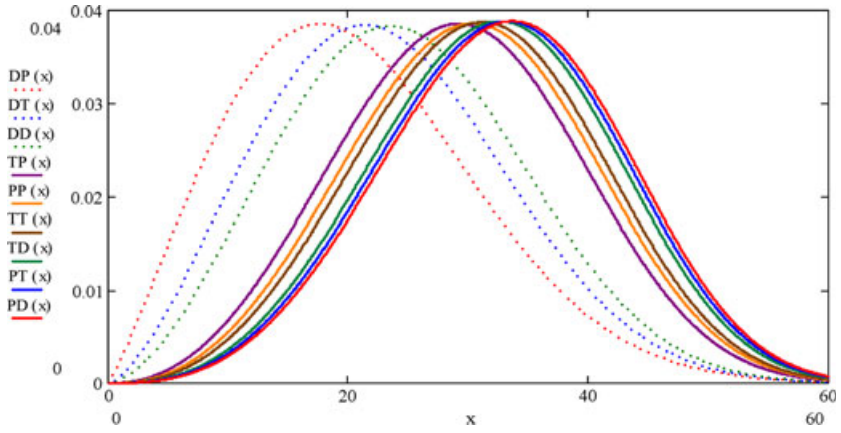

Figure 2: Weibull probability densities for evolution according to complete selection-survival strategies. The lines represent the Weibull probability distribution associated with the number of evolutions (defined as increase in determination coefficient of linear regression models) for all possible pairs of selection-survival strategy (where p = proportional, T = tournament, and D = deterministic). The selection-survival strategies that are deterministic in selection are quite different for all other selection-survival strategies. The deterministic selection strategy led to reaching the maximum number of evolutions earlier compared with other strategies. Other two groups of selection-survival strategies could be identified, the one resulted from the combination of tournament and proportional strategies (PP, TP, and TT) and the other form as a combination of all implemented strategies (TP, PT, and PD).

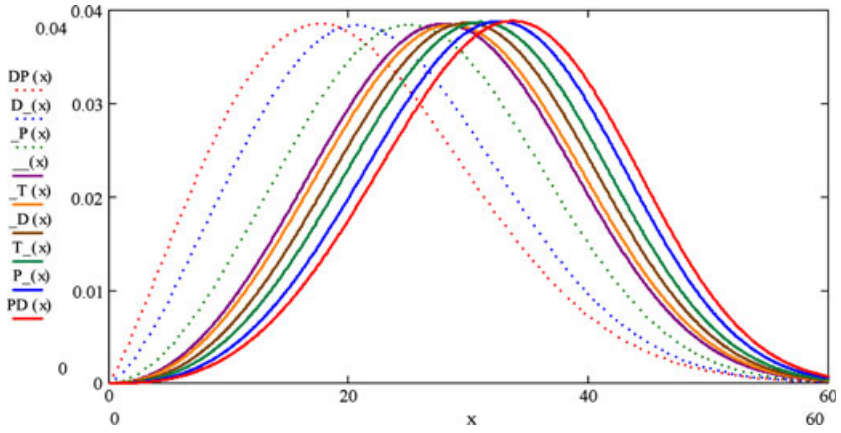

Figure 3: Weibull probability densities for evolution according to incomplete selection or survival strategies (where p = proportional, T = tournament, and D = deterministic) in context of extreme complete strategies (DP and PD). The behavior of probability distribution function associated with evolutions of incomplete strategies is presented in the boundaries of two extreme complete distribution functions represented by DP and PD. Two distinct behaviors could be observed near the left-hand boundary, one with deterministic selection, and the other one with proportional survival.

survival strategy as the dependencies had determination coefficients so closed to the perfect model (Figures 4 and 5). Furthermore, Figures 4 and 5 give an analytical relationship with a determination of 99% of exponential type (Figure 4) and respectively of rational type (Figure 5).

Exponential as well as rational functions are common in process modeling; it is not a surprise that we found them here. For example, the process of drug diffusion on the water proved exponential, (30) while rational function proved its usefulness in numerical simulation of stroke (31). Likewise, it also happens for entropy, when the entropy could be seen as a function of shape parameter of the Weibull distribution (Figure 6). The models for skewness and excess kurtosis as third-degree polynomial functions are not very useful because any process modeling proved up to now to follow this kind of functions.

In summary, the primary research aim has been achieved by identifying the behavior of number of evolutions under certain conditions and restrictions. It has been identified that the number of evolutions when the study was conducted on the octan-1-ol/H_2O partition coefficient of PCBs depends by both selection and survival strategy. The selection strategy proved to have a more significant influence in the number of evolutions supervised by genetic algorithms. The probability distribution function of the number of evolutions was already known as following a Weibull distribution. Furthermore, the scale and shape parameters of the Weibull distribution proved linearly related.

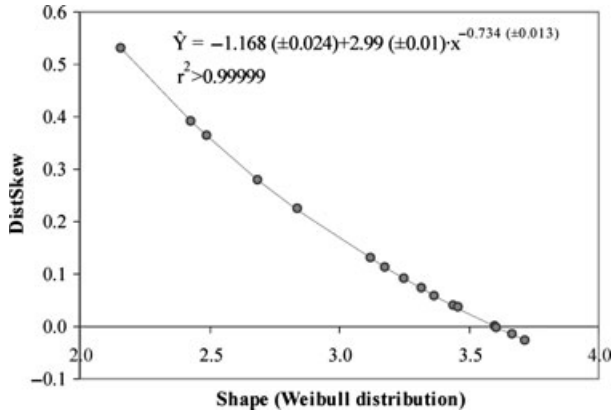

Figure 4: Dist skewness as function of shape. The exponential function with shape of the Weibull probability density function (*x*) as independent variable fits 99% of the skewness. The numerical values in the round brackets are the amount that must be subtracted and added to obtained the confidence level of the associated value.

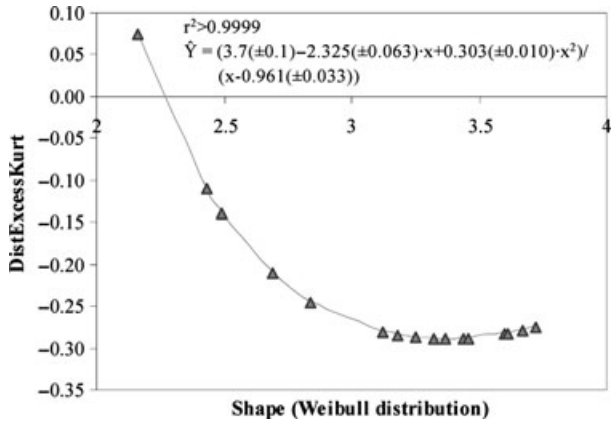

Figure 5: Dist excess kurt as function of shape. The exponential function with shape of the Weibull probability density function (*x*) as independent variable fits more than 99% of the excess kurtosis. The numerical values in the round brackets are the amount that must be subtract and add to obtain the confidence level of the associated value.

Based on this information, the population's parameters for full and partial selection-survival strategies, where full selection-survival strategy implied the presence of both strategies while partial selection-survival strategies implied the presence of selection or survival strategy, had been calculated. Distinct pattern of Weibull probability densities for evolution was identified for both full and partial selection-survival strategies: (i) relatively compact groups of populations that are not deterministic in selection for complete selection-survival strategies; (ii) distinct behavior with deterministic selection; (iii) distinct behavior with proportional selection. There are significant ramifications of these findings. The number of evolutions proved follows a natural process because the Weibull modeled natural processes (32) such as wind energy potential (33,34) and microbial survival (35). Moreover, their probability distribution function could be defined by one unknown parameter. All complete selection-survival strategies that comprise deterministic selection had distinct behavior compared with all other selection-survival strategies. This behavior consists on an earlier attainment of the maximum number of evolutions. Therefore, deterministic selection strategy is recommended when it is desirable to obtain the maximum number of evolutions in a shorter time. Moreover, two main properties of Weibull distribution for number of evolutions were identified: exponential function of skewness as function of shape and rational function of excess kurtosis as function of shape.

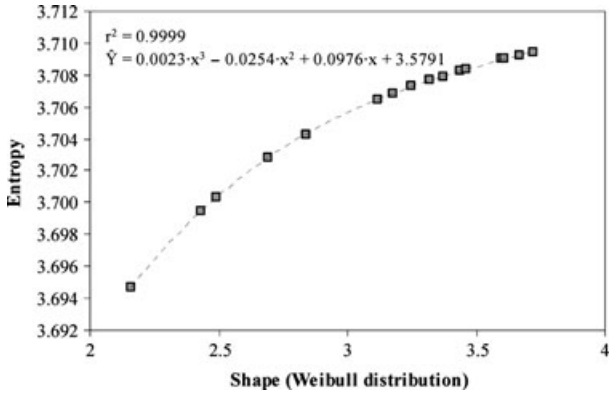

Figure 6: Entropy as function of shape. A third-degree polynomial function explains the entropy as function of shape parameter associated with Weibull probability density function (*x*), with a fit more than 99%.

This study showed distinct behavior in the number of evolutions in the QSAR on octan-1-ol/H_2O partition coefficient of PCBs according to selection and/or survival strategies. We hope that the results obtained in this study are extendable to other cases. The most important result pointed out in this study is related with the degrees of freedom of the supervised evolution. Surprisingly or not, our results revealed the existence of only one independent parameter that characterizes an evolution supervised in a certain strategy. The possibility of translating the experimental setting proposed here should be explored for certain pure biological processes of supervised evolution. Furthermore, ongoing studies in our laboratory aim to analyze whether the results obtained in this study could be generalized on number of evolutions for quantitative structure–activity relationships using the same conditions of genetic algorithm regardless the type of investigated compounds or regardless the sample size.

Conclusions

The number of evolutions expressed as increase in determination coefficient of linear regression model applied to the structure–activity relationship on the octan-1-ol/H_2O partition coefficient of polychlorinated biphenyls proved

significantly influenced by both genetic algorithm selection and survival strategy with a higher significant influence of selection strategy. Whenever reaching the maximum number of evolution in a shorter time is desired, a deterministic selection strategy is needed.

The analysis of distribution showed that the number of evolutions in a certain time frame of a system under constrained evolution comes from the Weibull distribution in which it is very likely (we found over 74% determination in the initial seek for linear association) to be only one independent statistical parameter. This parameter shapes the distribution and better supervising strategy applies when higher value is obtained.

From the pair of selection-survival strategies included in the study, we found that the highest value of the Weibull distribution shape parameter for the number of evolutions occurs when the proportional selection is associated with deterministic survival. On the opposite, the smallest number of evolutions is most likely to be observed when the evolution is supervised by deterministic selection and proportional survival.

Acknowledgments

The study was supported by POSDRU/89/1.5/S/62371 through a postdoctoral fellowship for L. Jäntschi. The funder had no role in study design, data collection, analysis and interpretation of data, in the writing of the report or in the decision to submit the article for publication.

References

1. Darwin C. (1859) On the Origin of Species by Means of Natural Selection. London, UK: Clowes and Sons.
2. Fraser A.S. (1957) Simulation of genetic systems by automatic digital computers. I. Introduction. Aust J Biol Sci;10:484–491.
3. Fraser A.S. (1957) Simulation of genetic systems by automatic digital computers. II. Effects on linkage on rates of advance under selection. Aust J Biol Sci;10:492–499.
4. Balloux F. (2001) EASYPOP (Version 1.7): A Computer program for the simulation of population genetics. J Hered;92:301–302.
5. Guillaume F., Rougemont J. (2006) Nemo: an evolutionary and population genetics programming framework. Bioinformatics;22:2556–2557.
6. Padhukasahasram B., Marjoram P., Wall J.D., Bustamante C.D., Nordborg M. (2008) Exploring population genetic models with recombination using efficient forward-time simulations. Genetics;178:2417–2427.
7. O'Fallon B. (2010) TreesimJ: a flexible forward time population genetic simulator. Bioinformatics;26: 2200–2201.
8. Hammett L.P. (1935) Some relations between reaction rates and equilibrium constants. Chem Rev;17:125–136.
9. Martin W.N., Spears W.M. (2001) Foundations of Genetic Algorithms 6. Calif: Morgan Kaufmann.
10. Prügel-Bennett A. (2001) The Mixing Rate of Different Crossover Operators. Foundations of Genetic Algorithms 6:261–274.
11. Spears W.M. (2000) The Equilibrium and Transient Behavior of Mutation and Recombination. Foundations of Genetic Algorithms 6:241–260.
12. Liu G., Yu Y.L., Tong B.G. (2012) Optimal energy-utilization ratio for long-distance cruising of a model fish. Phys Rev E 86: Article Number: 016308.
13. Rahmani H., Bonyadi M.R., Momeni A., Moghaddam M.E., Abbaspour M. (2011) Hardware design of a new genetic based disk scheduling method. Real-Time Syst;47:41–71.
14. Jäntschi L., Bolboacă S.D., Sestraş R.E. (2010) A study of genetic algorithm evolution on the lipophilicity of polychlorinated biphenyls. Chem Biodivers;7:1978–1989.
15. Jäntschi L., Bolboacă S.D., Sestraş R.E. (2010) Recording evolution supervised by a genetic algorithm for quantitative structure-activity relationship optimization. Appl Med Inform;26:89–100.
16. Jäntschi L. (2009) A genetic algorithm for structure-activity relationships: software implementation. Manuscript arXiv:0906.4846.
17. Jäntschi L., Bolboacă S.D., Sestraş R.E. (2012) A simulation study for the distribution law of relative moments of evolution. Complexity;17:52–63.
18. Jäntschi L. (2004) MDF – A new QSAR/QSPR molecular descriptors family. Leonardo J Sci;3:68–85.
19. Eisler R., Belisle A.A. (1996) Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review. Contaminant Hazard Reviews, 1-96.
20. Bolboacă S.D., Jäntschi L., Sestraş A.F., Sestraş R.E., Pamfil D.C. (2011) Pearson-fisher chi-square statistic revisited. Information;2:528–545.
21. Fisher R.A. (1973) Statistical Methods for Research Workers. New York: Hafner Publishing Company.
22. Jäntschi L. (2010) Genetic algorithms and their applications. Ph.D. Thesis; Prof. Dr. Radu E. Sestraş (supervisor), 172 pp.
23. Kolmogorov A. (1933) On the empirical determination of a distribution function. Giornale dell Instituto Italiano degli Attuari;4:83–91.
24. Smirnov N.V. (1939) On the estimation of the discrepancy between empirical curves of distribution for two independent samples. Bulletin of Moscow University 2: 3–16.
25. Anderson T.W., Darling D.A. (1952) Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Ann Math Stat;23:193–212.
26. Pearson K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag;50:157–175.
27. Fisher R.A. (1948) Combining independent tests of significance. Am Stat;2:30–31.
28. Shannon C.E. (1948) A mathematical theory of communication. Bell System Tech J;27:379–423 & 623-656.

29. Euler L. (1734) De Progressionibus harmonicus observations. Commentarii Academiæ Scientarum Imperialis Petropolitanæ;7:150–160.
30. Siepmann J., Siepmann F. (2008) Mathematical modeling of drug delivery. Int J Pharm;364:328–343.
31. Descombes S., Dumont T. (2008) Numerical simulation of a stroke: computational problems and methodology. Prog Biophys Mol Biol;97:40–53.
32. Weibull W. (1951) A statistical distribution function of wide applicability. J Appl Mech;18:293–297.
33. Mahmoudi H., Spahis N., Goosen M.F., Sablani S., Abdul-Wahab S.A., Ghaffour N., Drouiche N. (2009) Assessment of wind energy to power solar brackish water greenhouse desalination units: a case study from Algeria. Renew Sust Energ Rev;13:2149–2155.
34. Feijóo A., Villanueva D., Pazos J.L., Sobolewski R. (2011) Simulation of correlated wind speeds: a review. Renew Sust Energ Rev;15:2826–2832.
35. Cullen P.J., Tiwari B.K., O'Donnell C.P., Muthukumarappan K. (2009) Modelling approaches to ozone processing of liquid foods. Trends Food Sci Technol;20:125–136.