MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **71** (2014) 149-172

ISSN 0340 - 6253

Structure-Activity Relationships from Natural Evolution

Sorana D. Bolboacă[1,*], Daniela D. Roşca[2], and Lorentz Jäntschi[3]

[1] *Department of Medical Informatics and Biostatistics, "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, str. Louis Pasteur 6, 400349 Cluj-Napoca, Romania*
sbolboaca@umfcluj.ro

[2] *Department of Mathematics, Technical University of Cluj-Napoca, str. Memorandumului 28, 400114 Cluj-Napoca, Romania*
Daniela.Rosca@math.utcluj.ro

[3] *Department of Physics and Chemistry, Technical University of Cluj-Napoca, bdv. Muncii 103-105, 400641 Cluj-Napoca, Romania*
lorentz.jantschi@gmail.com

(Received April 15, 2013)

Abstract

Structure-activity relationships emulate the adaptation of chemical compounds to the biological environment. When a family of descriptors derived from a skeleton using different mathematical operations and physical properties is involved, the search space for structure-activity relationships is constructed in a natural way. A genetic algorithm implementing different selection and survival strategies, an unexplored issue, was designed and it is presented. A comparison of evolutionary strategies was conducted on a series of 206 polychlorinated biphenyls with known values of octan-1-ol/H_2O partition coefficients, on which a Molecular Descriptors Family (MDF) was generated as the search space. The obtained results showed that the implemented genetic algorithm proved to be a reliable method of finding optimal multiple-linear regression models that are able to explain relationships between structure and activity. The results showed that different tournament selection and proportional survival provide the solution closest to the one obtained by complete search. Furthermore, the results revealed that, in general, every pair of survival and selection strategies pushes evolution on significantly different paths and may form the basis of phylogeny analysis.

[*] Corresponding author

1. Introduction

Quantitative Structure Property/Activity Relationships (QSPR/QSAR) have many applications in drug design and discovery [1, 2]. One of the first methods used to explain the relation between the structure of compounds and their property/activity is the Multiple Linear Regression (MLR). This method is still a widely used approach in SPR/SAR studies, due to its form and accessible interpretable expressions [3, 4].

A crucial and difficult problem in SPR/SAR model development is the selection of the most relevant set of descriptors used as variables in MLR models. The description of the relationship between the structure of the compounds and their property/activity is also a difficult problem, since it involves the following issues: a). optimization - applied to the SPR/SAR model in order to maximize its estimation and prediction ability; b) classification - use of the SPR/SAR model in order to classify compounds into classes of activities/properties; c) decision - use of the SAR/SPR model in order to make a decision regarding the synthesis of a new compound for which the model predicts a better activity/property.

The difficult problem in structure-activity/property relationships could be stated as follows: Find the best structure-activity/property relationship that can describe the activity/property of the compounds (biochemical information) depending on their structure (structural information), when structural and biochemical information is available [5].

Usually, structural information is obtained from the molecular topology and geometry, and the biochemical information is obtained from an experiment.

The combination of Genetic Algorithm (GA) and MLR is used in QSAR/QSPR studies [6, 7] due to their capabilities of obtaining predictable models quickly. The differences in evolution, when different strategies are used for the selection of the progenitors and for the survival during generations of the sampled genetic material, are still unexplored. Structure-activity relationships emulate the adaptation of chemical compounds to the biological environment. When a family of descriptors derived from a skeleton using different mathematical operations and physical properties is involved, the search space for structure-activity relationships is constructed in a natural way.

Our goal was to compare the evolutions arising from a contingency of selection and survival strategies. For this, we have designed a GA, we have implemented and run it in order to obtain SAR. More precisely, we have solved the following difficult problems: *How to identify the relationship between the biochemical structures and the measured activity/properties of a set of compounds, when pools (families) of structure descriptors are available? Which evolutionary strategy is the best choice in order to obtain the relationship (which strategy provides the nearest optimum?).*

2. Methods

2.1. Genetic algorithm implementation

The problem of finding a link between the structure of compounds and their activity or property was first translated into genetic terms. In this research we used one family of descriptors (Molecular Descriptors Family (MDF) [8]), in order to define the portability of the program that implemented the genetic algorithm, but the approach is suitable to any family of descriptors (such as Fragmental Properties Index Family (FPIF) [9]; Molecular Descriptors Family on Vertices (MDFV) [10]; Structural Atomic Property Family (SAPF) [11]).

Every gene (one of the values from the Gene column in Table 1) encodes an operator which is used to construct the chromosome of a molecular descriptor. For example the gene sequence of the MDF family is $D_MA_PI_DI_MF_CS_ML_O$ as presented in Table 1.

Table 1: Search space using MDF family of molecular descriptors

Gene	Genome																												
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																							

MDF = Molecular Descriptors Family: ▪ D_M = distance operator: t = topologic distance; g = geometric distance. ▪ A_P = atomic property: C = cardinality; H = number of hydrogen atoms adjacent to the investigated atom; M = relative atomic mass; E = atomic electronegativity. ▪ G = group electronegativity; Q = atomic partial charge, semi-empirical extended Hückel model. ▪ I_D = interaction descriptor: D = d; d = 1/d; O = p_1; o = $1/p_1$; P = $p_1 \cdot p_2$; p = $1/p_1 \cdot p_2$; Q = $(p_1p_2)^{1/2}$; q = $1/(p_1 \cdot p_2)^{1/2}$; J = $p_1 \cdot d$; j = $1/p_1 \cdot d$; K = $p_1 \cdot p_2 \cdot d$; k = $1/p_1 \cdot p_2 \cdot d$; L = $d \cdot (p_1 \cdot p_2)^{1/2}$; l = $1/d \cdot (p_1p_2)^{1/2}$; V = p_1/d; E = p_1/d^2; W = p_1^2/d; w = $p_1 \cdot p_2/d$; F = p_1^2/d^2; f = $p_1 \cdot p_2/d^2$; S = p_1^2/d^3; s = $p_1 \cdot p_2/d^3$; T = p_1^2/d^4; t = $p_1 \cdot p_2/d^4$. ▪ I_M = overlapping interactions: r, R = models with sporadic and distant interactions; m, M = models with frequent and distant interactions; d, D = models with frequent and closed interactions. ▪ F_C = algorithm of molecular fragmentation applied on atomic pairs: m = fragmentation in minimal fragments; M = fragmentation in maximal fragments. D = fragmentation based on distances (Szeged criterion) [12]; P = fragmentation based on paths (Cluj criterion - [13]). ▪ S_M = global overlapping of fragments interaction: m = minimum value (group of values); M = maximum value (group of values); n = lowest absolute value (group values); N = highest absolute value (group of values); S = sum (group of means); A = arithmetic mean according to the number of fragment properties (group of means); a = arithmetic mean according to the number of atoms (group of means); B = (group of means); b = arithmetic mean according to the number of bonds (group of means); P = multiplication (geometric group); G = geometric mean according to the number of fragment properties (geometric group); g = geometric mean according to the number of fragments (geometric group); F = geometric mean according to the number of atoms (geometric group); f = geometric mean according to the number of bonds (geometric group); s = harmonic sum (harmonic group); H = harmonic mean according to the number of fragments property (harmonic group); h = harmonic mean according to the number of fragments (harmonic group); I = harmonic mean according to the number of atoms (harmonic group); i = harmonic mean according to the number of bonds (harmonic group). ▪ L_O = linearization operator: I = identity; i = inverse; A = absolute value; a = inverse of absolute value; L = logarithm; l = logarithm of absolute value.

Every descriptor in a family is a genotype (a possible set of values for every gene of a chromosome; e.g., *tCDrmmI* for *MDF*). The set of all genotypes represent the genetic material. The set of all possible combinations of values from the Genome column presented in Table 1 for MDF is:

$$\{t, g\} \cdot \{C, H, M, E, G, Q\} \cdot \{D, d, O, o, P, p, Q, q, J, j, K, k, L, l, V, E, W, w, F, f, S, s, T, t\} \cdot \{r, R, m, M, d, D\} \cdot\cdot \{m, M, D, P\} \cdot \{m, M, n, N, S, A, a, B, b, P, G, g, F, f, s, H, h, O, I, i\} \cdot \{I, i, A, a, L, l\}$$

The number of encoded values of the genes varies from two (for example for the gene encoding the metric type - topological or geometrical distance - D_M for *MDF*) to twenty-four (the I_D interaction descriptor of the *MDF* family). The size of the genetic material is of 787,968 for *MDF* ($2(D_M)\cdot 6(A_P)\cdot 24(I_D)\cdot 6(I_M)\cdot 4(F_C)\cdot 19(S_M)\cdot 6(L_O)$). The GA was used for searching the MDF descriptor space whereas the MLR (multiple linear regression) was used for fitness evaluation.

One of the following types of multiple linear regressions represents a possible solution and was searched on the molecular descriptors space:

$$b_0 + b_1X_1 + ... + b_kX_k = \hat{Y} \sim Y \quad (1)$$

$$b_1X_1 + ... + b_kX_k = \hat{Y} \sim Y \quad (2)$$

where Y is the array of the observed activity/property, $X_1, ... , X_k$ are descriptors drawn from a family, b_i, $i = 0, ..., k$ are the parameters of the model which have to be obtained under the assumption of least squares errors from a certain number M of observations, and Y is the activity/property estimated by the MLR equation (1) or (2).

We use the following notations:

- $k = |X|$ is the number of independent variables;
- $m = |Y| = |X_1| = ... = |X_k|$ is the number of experimental observations;
- $|b| = k + 1$ or $|b| = k$ is the number of unknown parameters of the multiple linear regression model (11) or (12), respectively.

The following assumptions were made in the multiple linear regression analysis:

- The measurement error of Y is both randomly and normally distributed;
- The values of the descriptors $X_1, ... ,X_k$ are normally distributed and are not affected by errors.

The calculation of the regression parameters b_i, $i \leq k$ from equation (1) or (2) is always risky. The statistical significance and the associated confidence intervals of regression parameters can be obtained using Student's t distribution - see [14,15].

If equation (1) or (2) has unique solution then $|b| \le m - 1$. However, this condition is not sufficient. The parameters (b_i), $i < k$ have statistical significance if $|b| \le m - 6$.

If b_0 from equation (1) is not statistically significant, then equation (2) is used as an alternative to (1). The absence of statistically significant coefficients b_i for $1 \le i \le k$ in equations (1) and (2) should reject the hypothesis that there is a linear relation between X_i and Y.

Let S denote the search space and let N be the total number of descriptors. Then its size is

$$|S| = \prod_{j=1}^{k} \frac{N - j + 1}{j} = \binom{N}{k} \quad (3)$$

Formula (3) expresses the number of all possible selections of k descriptors from a total of *N*. The value of |S| could be doubled if the search is conducted by both (1) and (2).

We can show that this search defines an NP-hard problem (a problem whose solution obtained by the best known algorithm requires an execution time that increases exponentially with the size of the input data).

The design of the genetic algorithm implies the random or deterministic initialization of a sample *p* of chromosomes from the genetic material. For example, a subset of the genetic material of the molecular descriptors family, such as, {*tCDrmmI*, *gHdRMMi*, *gMddMMi*} is a sample of size 3 for *MDF*. The descriptors $X_1, \ldots, X_p$ enter the evolutionary process in the cultivar. The evolutionary process is a complex genetic process that implies selection, crossover and mutation, while the cultivar is regarded as a memory or virtual space in which the genotypes are transformed into phenotypes by applying the operators defined by the gene values for the entire set of molecules; the phenotype associated with the genotype is thus an array of numerical values, one for each compound.

The genetic algorithm, regarded here as an algorithm that uses instructions to describe the evolutionary process applied to the sample, operates on a sample for which the content is modified in every generation. A generation is an iteration of the genetic algorithm. Every set of k distinct descriptors is a point in the search space and is a possible solution of regression equation defined by (1), or if (1) fails of (2). Our genetic algorithm implements the following operations:

- **Crossover** of two genotypes involves the choice (random or deterministic) of a contiguous sequence, which must be crossed over from the gene array. The values of the sequences are exchanged and two descendants are obtained.
- **Mutation** of a genotype implies a change in the value of a gene from a chromosome with other values from the list of possible values for the gene.

- **Selection** is the implicit operation that is required by mutation and crossover. Selection acts based on a selection score, F_S i.e. a numerical value that is associated with the individual and calculated from the fitness of the phenotype into its cultivar. At least part of the descendants should be viable descriptors (phenotypic viability refers to the potential use in regressions). A descriptor was considered to be viable if it had real and finite non-identical values for all of the molecules in the dataset. Other supplementary conditions imposed for phenotypic viability are a reasonable variability with the coefficient of variation, a reasonable departure from normality with Jarque-Bera test [16], and a reasonable power of explanation with its determination coefficient).
- **Survival** replaces some individuals from the sample with viable descendants. This process was applied based on a survival score, V_S, a numerical value associated with an individual, based on the genotype and on the phenotype. On the genotype it measures the similarity of a genotype with all of the other genotypes of the sample, for the purpose of maintaining diversity in the genetic material, while on the phenotype, it measure the similarity of the phenotype with all the other individuals from the cultivar, in order to preserve the diversity of the traits.
- **Evolutionary objective** is measured by an objective function, where the determination coefficient was used and the objective was to maximize it.

Not all of the individuals were included in the next generation; the individuals that did not survive were withdrawn. The number of the replaced individuals was equal to the number of viable descendants. This strategy was applied to maintain the same number of individuals in the cultivar. Selection and survival were applied based on selection and survival scores and were they implemented via selection and survival strategies.

The strategy is a method of extracting an individual from the sample using scores. Three approaches were applied (proportional, deterministic, and tournament) to the scores (see Table 2). The values of the scores were normalized from [min., max.] to [0, 1]. The values were updated in every generation during the entire evolutionary process. Score functions (f_i in Table 2) had different expressions for: evolution (evolution objective scores, Figure 1), selection (selection scores, Figure 1) and survival (survival scores, Figure 1).

Table 2: Evolutionary strategies (scores function fi = Fitness(Chromosome i))

Method		
Proportional	$p_i = f_i/\Sigma_i f_i$	Likelihood proportional to the score (using the p_i probability to extract)
Deterministic	$i \mid f_i$ = max. or min.	Extraction of the strongest or of the weakest individual (elitism)
Tournament	(f_i,f_j) = max. or min.	te for extraction

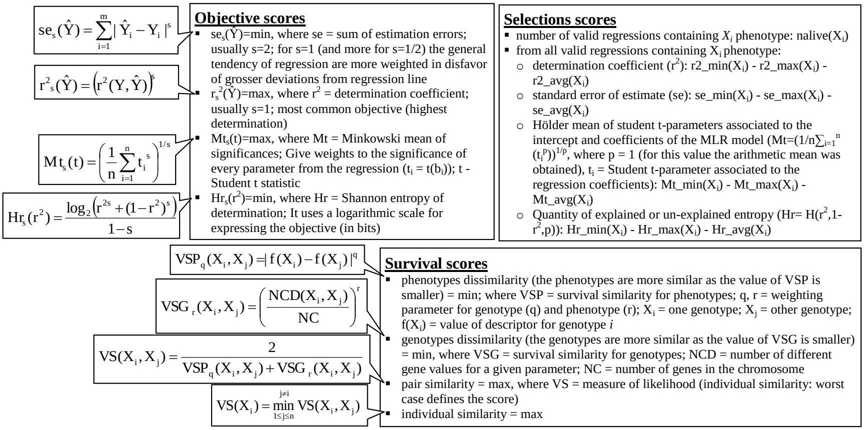

Figure 1: Objective, selection and survival scores for multiple linear regressions (used with eq.(1) or with eq.(2) when b_0 not statistically significant)

Our genetic algorithm (see Figure 2) generates randomly a sample of genotypes of a given size p, maintained constant during the evolution, $k < p < N$, in order to solve the NP-hard problem of multiple linear regressions, given in the algorithm 1.

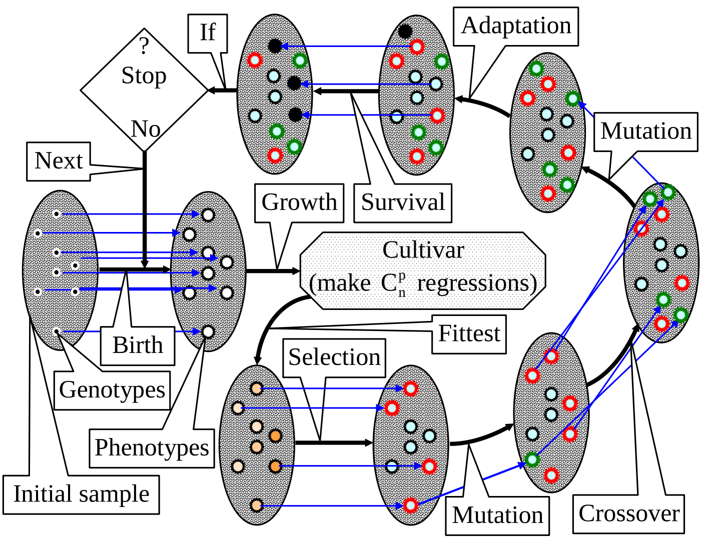

Figure 2: The genetic algorithm: evolution

Algorithm 1 *The GA−MLR−QSAR algorithm*

repeat

- *Obtain phenotypes from genotypes;*
- *Compute multiple linear regressions of type (1) and of type (2) if necessary; keep the best model found and mark the phenotypes, which act as descriptors in the model of the survivors; keep the regression scores;*
- *Obtain objective scores of the individuals from regression scores;*
- *Obtain selection scores of the individuals,* F_S;
- *Extract pairs of genotypes from a sample of size l (sample given) applying the s selection strategy on the selection scores;*
- *Mutate every 2l genotypes (parents) with a low probability pp;*
- *Crossover the l pairs of genotypes and obtain 2l new descendants;*
- *Mutate every 2l genotypes (children) with a low probability cp;*
- *Obtain a viable (adapted to the environment) subset of children of size* $v \leq 2l$;
- *Obtain survival scores of the remaining individuals (genotype and phenotype),* V_S;
- *Remove individuals from the sample applying the survival strategy v on the survival scores and replace them with a children subset;*

until *the imposed number of iterations (set at 20,000) was exhausted.*

The proposed genetic algorithm was implemented as a Windows-based FreePascal application with MySQL connectivity for fetching the data and was run as a stand-alone program.

2.2. Genetic algorithm assessment

The developed and implemented *GA−MLR−QSAR* was assessed on a sample of 206 polychlorinated biphenyls (*PCBs*) using the *MDF* descriptors family. The measured property was *octan-1-ol*/H_2O partition coefficients [17]. The HyperChem program (Hypercube, Inc., Gainesville, FL, USA) was used to draw the structures of PCBs. The partial charges of the compounds were calculated using the semiempirical extended Huckel model (single point approach [18]), and the geometry was optimized using the Austin method [19]. The following statistics were applied to test the normality of the experimental data [20]: Kolmogorov-Smirnov, Anderson-Darling, Chi-Square, Wilks-Shapiro, $Z_{skewness}$, $Z_{kurtosis}$, and Jarque-Bera tests. According to these statistics, the experimental data proved to be normally distributed [20]. The obtained descriptors were statistical analyzed in order to avoid potential overlapping and redundancy. The following descriptors were withdrawn from further MLR analysis:

- descriptors with identical names and/or values,
- descriptors with a Jarque-Bera value greater than the critical value for the experimental activity [21],
- highly inter-correlated descriptors.

For testing the *GA−MLR−QSAR* program, an experiment containing all possible combinations of selection and survival strategies was designed and run on five dual core processor-based machines. The results are presented in Table 3.

In order to avoid the overwriting of the files from one program to another, a random number was added automatically by the program to the name of the output file, as shown in Table 4. The following parameters were assigned to assess the implemented genetic algorithm:

- Search space: Molecular Descriptors Family on PCBs, already available in the MDF database from the previous investigation [17], http://l.academicdirect.org/Chemistry/SARs/MDF−SARs/.
- Initial sample: 12 descriptors randomly chosen from the pool of MDF descriptors.
- Genotype adaptation: minimum of absolute deviation relative to the deviation of measured activity (a ratio 0.1 was taken); maximum of ratio between Jarque-Bera values for the descriptor and the measured activity (1 was taken); and minimum value of the determination coefficient between estimated and experimental data (0.1 was taken).
- Number of independent variables in the MLR model (number of descriptors): 4.
- Evolution strategy: all possible pairs of survival and selection strategies (e.g., PP, PT, PD, TP, TT, TD, DP, DT, DD, where P = Proportional, T = Tournament, and D = Deterministic).
- Probability of parent/child mutation: set at 0.05.

Table 3: Experimental design for GA-MLR assessment: selection and survival strategies

Selection \ Survival	Proportional (P)	Deterministic (D)	Tournament (T)
Proportional (P)	P&P: 4044	P&D: 2441	P&T: 9878
Deterministic (D)	D&P: 5108	D&D: 6369	D&T: 6690
Tournament (T)	T&P: 5828	T&D: 4872	T&T: 1758

P = Proportional; D = Deterministic; T = Tournament;
Experimental design:
http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/PCB_XXXX_cfg.txt (were XXXX is the number corresponding to the selection-survival strategy: for example, XXXX = 4044 for PP evolution strategy);
Evolution records:
http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival/PCB_XXXX_evo.txt

- Two genes were implied in the mutation.
- Generations: The identified solutions were stored in the results files. The program continued to adapt, until the imposed maximum number of 20,000 generations.

- Optimization criterion: maximization of the determination coefficient obtained from *GA−MLR*.

The Chi-Square statistic [22, 23, 24] was used for testing the homogeneity of the populations' genotypes, which were obtained by different selection and survival strategies. The frequency of the genotypes without accounting the last gene of the MDF family was used as both an adaptation and a variability measure of the genetic material produced by the selection and survival strategies. In order to avoid a random bias, we have performed 46 runs for every pair of selection and survival strategies.

2.3. MLR evaluations

In order to identify the best model for every survival-selection strategy, we have used the following criteria [25]:

- Model assessment. Highest explanation of the observed variance (expressed as highest values of significant correlation coefficients between the observed and estimated activity), lowest standard error of estimate s_{est}, highest Fisher value (and lowest associated *p*-value) as well as significant coefficients of the regression model (highest *t*-value, lowest associated *p*-value).
- Internal validation. Cross-validation leave-one-out analysis (*cv-loo*) [26] was applied to test the performances of the identified *GA−MLR−QSAR* models. A *QSAR* model was considered reliable if a small difference between the determination coefficient r^2 and the cross-validation leave-one-out score $r^2_{cv\text{-}loo}$ was identified (*difference*<0.2, $r^2_{cv\text{-}loo}$>0.6). It was proved that leave-one-out analysis overestimates the predictive power of a model [27]
- Information criteria: seven information criteria [10, 28] were applied to the models given in (4)-(13), in order to compare the information stored by the models. The following criteria were used: Akaike information criteria (*AIC*); AIC based on the determination coefficient (AIC_{R2}); McQuarrie and Tsai corrected AIC (*AICu*); Bayesian Information Criterion (*BIC*); Amemiya Prediction Criterion (*APC*); Hannan-Quinn Criterion (*HQC*); and Kubinyi function (*FIT*). The best model is the one with smallest *AIC*, *BIC*, *APC* and *HQC* and highest *FIT*. The comparisons of the models were conducted on correlation coefficients using Steiger's formula [29].

3. Results and Discussion

3.1. Genetic algorithm

The developed *GA−MLR−QSAR* was successfully implemented. The *GA−MLR−QSAR* program was realized implementing the following algorithms:

Algorithm 2 *The algorithm for Selection scores* ***(FS)***

- *Compute all possible regressions between phenotypes and store those with valid selection scores;*
- *Compute the selection scores of the phenotypes from all of their occurrences in regressions;*
- *Compute the selection scores of the genotypes from all of their occurrences in phenotypes;*
- *Normalize the scores between generations whenever specified;*
- *Round the obtained values to the defined number of significant digits;*
- *Build ranks of the scores;*
- *Replace the scores with ranks if configured to do so;*
- *Sort out the scores;*
- *Outputs*: **FS** - *array of selection scores*; **FSD** - *array of distinct selection scores*; **FSC** - *occurrences of every distinct selection score*.

Algorithm 3 *Proportional strategy* ***(P)***

- *Set* **Selected−Genotypes** *to Empty*;
- *For every selection from* 1 *to N_Sel* (*N_Sel - number of selections to be performed*):
    - *Compute the sum of unselected genotype scores to FS_Sum*;
    - *Randomly generate a number FS_Freq between 0 and FS_Sum (inclusive)*;
    - *Find first index Group from* **FSD** *for which* $\mathrm{FS\_Freq} \le \sum_{i<\mathrm{Group}} \mathrm{FSD}_i \cdot \mathrm{FSC}_i$
    - *Randomly generate a number FSD_Next between 1 and* FSC_i;
    - *Push into* **Selected−Genotypes** *the FSD_Next value (not selected yet) of* FSD_{Group} *from* **FS** *and decrease* FSC_{Group} *with one.*

Algorithm 4 *Deterministic strategy* ***(D)***

- *Set* **Selected−Genotypes** *to* ∅*, Already_Selected to 0, Group to sample size*;
- *While Already_Selected +* $FSC_{Group} \le N\_Sel$ *assign the indices from* **FS** *equal to FSDGroup into* **Selected−Genotypes** *and decrease Group by one if possible, or otherwise, increase by one*;

- *While Already_Selected ≤ N_Sel (full groups are exhausted; only a part of the group will be selected)*;
  - *Randomly generate a number FSD_Next between* 1 *and* FSC_i;
  - *Add to* **Selected−Genotypes** *the FSD_Next value (not selected yet) of* FSD_{Group} *from* ***FS*** *and decrease* FSC_{Group} *with one.*

Algorithm 5 *Tournament strategy **(T)***

- *Let N_Gen be the number of genotypes from the sample;*
- *Randomly generate a permutation of {1 ... N_Gen} into* **Selected−Genotypes**;
- *For every i_Sel from 2 to N_Sel (first N_Sel competes in tournament)*
  - *If* $FS_{i\_Sel} \le FS_{i\_Sel-1}$ *then*

    * *If* $FS_{i\_Sel} = FS_{i\_Sel-1}$ *then if random selection between 0 and 1 generates 0, then continue (for iteration);*

    * *Exchange in* **FS** *the values from i_Sel and i_Sel − 1*;
- *If N_Sel < N_Gen then (last selected did not participate in tournament and there are still elements with which to compete in sample)*
  - *Generate randomly a number i_Sel between N_Sel + 1 and N_Gen;*
  - *If* $FS_{N\_Sel} \le FS_{i\_Sel}$ *then*

    * *If* $FS_{N\_Sel} = FS_{i\_Sel}$ *then if random selection between 0 and 1; when 0 then stop (tournament completed);*

    * *Exchange in FS the values from i_Sel and N_Sel.*

The same calculations used in the selection scores (F_S) were also applied to survival scores (V_S) - see Figure 1. Proportional survival strategy uses the same procedure on V_S as the proportional selection on F_S. Deterministic survival strategy uses the same procedure on V_S as deterministic selection on F_S. Tournament survival strategy uses the same procedure on V_S as tournament selection on F_S.

The evolutionary program which implements the genetic algorithm was built to work with any family of molecular descriptors and was parameterized through a series of configuration files. The program uses a configuration file to connect with the database in which molecular descriptors are stored. The c_galg.cfg configuration file specifies the security protocols required to connect to the database. The c_galg.cfg configuration file contains the definition of the genetic topology of the descriptors' family. The values of the parameters that define the evolution of the genetic algorithm were stored in the c_galg.cfg configuration file.

3.2. GA-MLR-QSAR on PCB data set

The summary of the results obtained on 46 runs on the investigated sample of PCBs was obtained by the processing of *_evo.txt files (Table 4).

The genotypes' adaptation capacity could be assessed by analyzing the frequency of genotype occurrences in the sample. This procedure also measures the variability of the genetic material induced by the selection and survival method. Tables 5 to 14 present the results obtained by checking the homogeneity hypotheses regarding the number of genotypes found in the evolution of generations. In these tables on the rows we have selection strategy; on the columns we have survival strategy.

The tables contain the observed numbers; while the expected numbers, according to the homogeneity hypothesis, are given between parentheses. The analysis of the results presented in Tables 5-14 revealed the following:

- The populations of the number of distinct genotypes, when the observations were drawn with proportional and deterministic selections, and all types of survival strategies were inhomogeneous (probability from Chi-Square distribution <5%, see Table 5).
- The populations of the number of distinct genotypes, when all of the survival strategies were applied were inhomogeneous for tournament and deterministic selection strategies (probability from Chi-Square distribution <5%, see Table 5).
- The populations of the total number of genotypes when the observations were drawn from different selection and survival strategies proved to be inhomogeneous (see Table 6).
- The populations of the genotypes that provided valid regressions when the observations were drawn from different selection and survival strategies proved to be inhomogeneous (see Table 7).
- The populations of the number of distinct genotypes from the top 23 proved to be non-homogenous when the deterministic selection strategy and all the survival strategies were applied. For all of the other possibilities, the alternative hypotheses could not be rejected (see Table 8).
- The populations of the total number of genotypes from the top 23 proved to be inhomogeneous when the observations were drawn using different selection and survival strategies (see Table 9).

Table 4: The most frequent genotypes found in the generations that led to evolution (improvement of the objective function) following 46 independent runs

Selection strategy

Proportional

VS	Gen	Num	Occ	Par
P	T23	13	406	389
	mMdlHg	1	46	43
	MDMKHt	1	40	39
	nDRLHt	1	40	39
	iPDKCg	1	39	39
	ADDJCg	1	35	35
	mDdjGg	1	31	30
	bDDDGg	1	28	19
	bDDJCg	1	27	27
	sDdLHg	1	25	25
	BDDDGg	1	24	22
	bDMLEg	1	24	24
	bDMLGg	1	24	24
	MMDPMt	1	23	23
	Tot	6760	16788	15902
D	T23	13	378	371
	iPMDHg	1	39	37
	bPRjCg	1	38	38
	IPMDEg	1	37	36
	mMdoHt	1	30	29
	IPRKCg	1	29	29
	MDRLHt	1	29	29
	MMdlHg	1	29	29
	MDmWHg	1	26	26
	BPRjCg	1	26	25
	NDRlHt	1	25	25
	iPMDCg	1	24	23
	bmrVCt	1	23	23
	IPMDCg	1	23	22
	Tot	8070	18240	17797
T	T23	6	214	207
	MMdlHg	1	47	47
	mMdlHg	1	46	43
	sPDLEg	1	38	38
	AMdwGg	1	29	29
	IPMDHg	1	29	27
	mMdqGt	1	25	23
	Tot	7466	16599	15739

Deterministic

VS	Gen	Num	Occ	Par
P	T23	3	89	72
	MDRLHt	1	31	31
	ImrWCg	1	30	19
	ImrWHg	1	28	22
	Total	3922	10764	9742
D	T23	32	893	893
	gmdKHg	1	48	48
	iPDDGg	1	43	43
	bmRkHg	1	37	37
	gMdEQg	1	34	34
	sDRDGg	1	34	34
	HDmLQt	1	33	33
	MDMKHt	1	33	33
	mMdLMt	1	30	30
	MMmwCg	1	29	29
	bmdFEt	1	29	29
	hDDJCg	1	27	27
	hDDpCg	1	27	27
	hPmEMg	1	27	27
	sPmJMt	1	27	27
	NmdlQg	1	26	26
	SMMFEg	1	26	26
	bMddEg	1	26	26
	sPRDHt	1	26	26
	BDrsGt	1	25	25
	hDMKEg	1	25	25
	smdoQg	1	25	25
	AMMpHt	1	24	24
	GPmVCg	1	24	24
	SMMjEt	1	24	24
	BPMkHg	1	23	23
	GmmlQt	1	23	23
	bPmjMg	1	23	23
	hDDDHg	1	23	23
	hMdWGt	1	23	23
	hPmSEg	1	23	23
	hmddCt	1	23	23
	imMtGg	1	23	23
	Tot	4385	13560	13316
T	T23	5	152	152
	NDRkHt	1	37	37
	sDDEMg	1	30	30
	hMrkGg	1	29	29
	MDDKHt	1	28	28
	sMrLCg	1	28	28
	Tot	4965	12504	11572

Tournament

VS	Gen	Num	Occ	Par
P	T23	13	419	405
	sPDJEg	1	64	64
	mMdlHg	1	44	42
	MMdlHg	1	40	40
	MDdjEg	1	32	30
	sDMDMg	1	29	28
	mMdqGt	1	29	23
	sDDKCg	1	28	28
	sPDLEg	1	28	28
	aDDKEg	1	27	27
	sDRKCg	1	26	26
	sPRKGg	1	25	22
	sDMLGg	1	24	24
	MDRLHt	1	23	23
	Tot	6537	16368	15317
D	T23	21	714	687
	MDRLHt	1	88	87
	IPMJCg	1	46	45
	IPMDEg	1	42	38
	sDRJEg	1	41	39
	iPMKCg	1	36	36
	iPDJCg	1	35	33
	sPDLEg	1	34	34
	mDRlHt	1	33	33
	nDRLHt	1	32	31
	sDMLCg	1	31	29
	iPDDGg	1	31	28
	iPDDEg	1	29	27
	mDRkHt	1	28	28
	IPRKCg	1	27	26
	IPDJCg	1	27	25
	iPDKCg	1	27	25
	bPmkEt	1	26	26
	sDDJEg	1	26	26
	MDDKHt	1	26	22
	IPDKCg	1	25	25
	sDDLHg	1	24	24
	Tot	7964	17700	17331
T	T23	8	217	213
	IDRwHt	1	34	34
	mMdlHg	1	28	28
	nMRSEt	1	28	27
	mPRDHt	1	27	26
	MDRLHt	1	26	26
	smmLCt	1	26	24
	AMDEQt	1	24	24
	IDRwGt	1	24	24
	Tot	7529	17100	16151

VS = Survival strategy; P = Proportional; T = Tournament; D = Deterministic; Gen = Genotypes; Num = Number (of distinct genotypes); Occ = Occurrences (of the genotypes); Par = Participations in valid regressions (of the genotypes); T23 = Top of the genotypes that occur more than or equal to 23 times; Tot = total number of all genotypes.

Table 5: Populations of distinct observed numbers of genotypes from total (expected numbers of genotypes provided in round brackets)

χ^2	P: Obs.(Exp.)	T: Obs. (Exp.)	D: Obs. (Exp.)	Σ	Unexplained squared error ($p_{\chi2}(x^2 > X^2,2)^*$)	
P	6760 (6665)	7466 (7726)	8070 (7904)	22296	$X^2(P,\cdot) = 13.6$ (1‰)	$X^2(\cdot,P) = 2.25$ (32%)
T	6537 (6586)	7529 (7634)	7964 (7810)	22030	$X^2(T,\cdot) = 4.85$ (9%)	$X^2(\cdot,T) = 39.3$ ($3\cdot10^{-9}$)
D	3922 (3968)	4965 (4599)	4385 (4705)	13272	$X^2(D,\cdot) = 51.4$ ($7\cdot10^{-12}$)	$X^2(\cdot,D) = 28.3$ ($7\cdot10^{-7}$)
Σ	17219	19960	20419	57598	$X^2(\cdot,\cdot) = 69.9$ $p_{\chi2}(x^2 > X^2,4) = 2\cdot10^{-14}$	

P = Proportional; T = Tournament; D = Deterministic; Obs. = Observed frequency; Exp. = Expected frequency; ∑ = sum; * Probability from Chi-Square distribution; X^2 = Chi-Square value; (pχ2(·,·)) = its associated probability to be observed

Table 6: Populations of observed numbers of genotypes (expected numbers provided in round brackets)

χ^2	P	T	D	Σ	Unexplained squared error ($p_{\chi2}(x^2 > X^2,2)^*$)	
P	16788 (16240)	16599 (17084)	18240 (18303)	51627	$X^2(P,\cdot) = 32.5$ ($9\cdot10^{-8}$)	$X^2(\cdot,P) = 81.3$ ($2\cdot10^{-18}$)
T	16368 (16095)	17100 (16932)	17700 (18140)	51168	$X^2(T,\cdot) = 17.0$ ($2\cdot10^{-4}$)	$X^2(\cdot,T) = 23.7$ ($7\cdot10^{-6}$)
D	10764 (11585)	12504 (12187)	13560 (13056)	36828	$X^2(D,\cdot) = 85.9$ ($2\cdot10^{-19}$)	$X^2(\cdot,D) = 30.3$ ($3\cdot10^{-7}$)
Σ	43920	46203	49500	139623	$X^2(\cdot,\cdot) = 135$ $p_{\chi2}(x^2 > X^2,4) = 3\cdot10^{-28}$	

P = Proportional; T = Tournament; D = Deterministic; ∑ = sum; * Probability from Chi-Square distribution

Table 7: Populations of observed number of genotypes that provided valid regressions from total (expected number of genotypes provided in round brackets)

χ^2	P	T	D	Σ	Unexplained squared error ($p_{\chi2}(x^2 > X^2,2)^*$)	
P	15902 (15241)	15739 (16172)	17797 (18025)	49438	$X^2(P,\cdot) = 43.1$ ($4\cdot10^{-10}$)	$X^2(\cdot,P) = 115$ ($9\cdot10^{-26}$)
T	15317 (15044)	16151 (15963)	17331 (17792)	48799	$X^2(T,\cdot) = 19.1$ ($7\cdot10^{-5}$)	$X^2(\cdot,T) = 19.1$ ($7\cdot10^{-5}$)
D	9742 (10676)	11572 (11328)	13316 (12626)	34630	$X^2(D,\cdot) = 125$ ($8\cdot10^{-28}$)	$X^2(\cdot,D) = 52.5$ ($4\cdot10^{-12}$)
Σ	40961	43462	48444	132867	$X^2(\cdot,\cdot) = 187$ $p_{\chi2}(x^2 > X^2,4) = 2\cdot10^{-39}$	

P = Proportional; T = Tournament; D = Deterministic; Obs. = Observed frequency; Exp. = Expected frequency; ∑ = sum; * Probability from Chi-Square distribution

Table 8: Populations of distinct observed numbers of genotypes from the top 23 (expected values provided in round brackets)

χ^2	P	T	D	Σ	Unexplained squared error ($p_{\chi2}(x^2 > X^2,2)$)	
P	13 (8)	6 (5)	13 (19)	32	$X^2(P,\cdot) = 5.22$ (7.4%)	$X^2(\cdot,P) = 8.39$ (1.5%)
T	13 (11)	8 (7)	21 (24)	42	$X^2(T,\cdot) = 0.88$ (64%)	$X^2(\cdot,T) = 0.91$ (63%)
D	3 (10)	5 (7)	32 (23)	40	$X^2(D,\cdot) = 8.99$ (1.1%)	$X^2(\cdot,D) = 5.79$ (5.5%)
Σ	29	19	66	114	$X^2(\cdot,\cdot) = 15.1$; $p_{\chi2}(x^2 > X^2,4) = 4.5$‰	

P = Proportional; T = Tournament; D = Deterministic; ∑ = sum; * Probability from Chi-Square distribution

- The populations of the genotypes from the top 23 that provided valid regressions, when the observations were drawn from different selection and survival strategies proved to be inhomogeneous (see Table 10).

Table 9: Populations of observed numbers of genotypes from the top 23 (expected numbers provided in round brackets)

χ^2	P	T	D	Σ	Unexplained squared error ($p_{\chi 2}(x^2 > X^2,2)$)	
P	406 (262)	214 (167)	378 (569)	998	$X^2(P,\cdot) = 156\ (10^{-34})$	$X^2(\cdot,P) = 238\ (2\cdot 10^{-52})$
T	419 (354)	217 (226)	714 (770)	1350	$X^2(T,\cdot) = 16.4$ (0.3‰)	$X^2(\cdot,T) = 21.2\ (3\cdot 10^{-5})$
D	89 (298)	152 (190)	893 (646)	1134	$X^2(D,\cdot) = 249\ (10^{-54})$	$X^2(\cdot,D) = 163\ (5\cdot 10^{-36})$
Σ	914	583	1985	3482	$X^2(\cdot,\cdot) = 421;\ p_{\chi 2}(x^2 > X^2,4) = 6\cdot 10^{-90}$	

P = Proportional; T = Tournament; D = Deterministic; ∑ = sum; * Probability from Chi-Square distribution

Table 10: Populations of observed genotypes that provided valid regressions from the top 23 (expected number of genotypes are provided in round brackets)

χ^2	P	T	D	Σ	Unexplained squared error ($p_{\chi 2}(x^2 > X^2,2)$)	
P	389 (247)	207 (163)	371 (557)	967	$X^2(P,\cdot) = 156\ (2\cdot 10^{-34})$	$X^2(\cdot,P) = 256\ (2\cdot 10^{-56})$
T	405 (333)	213 (220)	687 (751)	1305	$X^2(T,\cdot) = 21.2\ (2\cdot 10^{-5})$	$X^2(\cdot,T) = 19.3\ (6\cdot 10^{-5})$
D	72 (285)	152 (189)	893 (643)	1117	$X^2(D,\cdot) = 264\ (6\cdot 10^{-58})$	$X^2(\cdot,D) = 165\ (2\cdot 10^{-36})$
Σ	866	572	1951	3389	$X^2(\cdot,\cdot) = 441;\ p_{\chi 2}(x^2 > X^2,4) = 5\cdot 10^{-94}$	

P = Proportional; T = Tournament; D = Deterministic; Obs. = Observed frequency; Exp. = Expected frequency; ∑ = sum; * Probability from Chi-Square distribution

3.3. Model analysis

For each strategy pair, the equations of the most accurate best models are as follows:

$$\hat{Y}_{PP} = 40.90(\pm 9.08) + \mathit{lsDMLGg}\cdot 12.85(\pm 2.95) + \mathit{IBDmKGg}\cdot(5.21\cdot 10^{-4})\ (\pm 7.75\cdot 10^{-5}) + \mathit{IMDRLHt}\cdot(-2.06\cdot 10^{-2})(\pm 6.39\cdot 10^{-3}) + \mathit{IsDRLEg}\cdot(-176.68)(\pm 40.39) \tag{4}$$

$$\hat{Y}_{PD} = 26.33(\pm 4.59) + \mathit{iNDRlHt}\cdot(-1.86\cdot 10^{-2})(\pm 6.18\cdot 10^{-3}) + \mathit{IsPDJEg}\cdot(-50.14)(\pm 11.06) + \mathit{lSPRlEg}\cdot(-6.26)(\pm 1.25) + \mathit{ISDRKHt}\cdot(-5.87\cdot 10^{-5})(\pm 7.08\cdot 10^{-6}) \tag{5}$$

$$\hat{Y}_{PT} = 11.15(\pm 1.90) + \mathit{IHDMDHt}\cdot(-5.61\cdot 10^{-2})(\pm 6.70\cdot 10^{-3}) + \mathit{IiPDLCg}\cdot(-9.07)(\pm 2.15) + \mathit{imDRlHt}\cdot(1.93\cdot 10^{-2})(\pm 6.27\cdot 10^{-3}) + \mathit{iIPMDHg}\cdot(-1.97)(\pm 0.41) \tag{6}$$

$$\hat{Y}_{DP} = 24.50(\pm 4.80) + \mathit{ISDRkEg}\cdot(-4.58)(\pm 1.08) + \mathit{iSDRlGg}\cdot(-113.19)(\pm 26.73) + \mathit{InDRLHt}\cdot(2.16\cdot 10^{-2})(\pm 6.59\cdot 10^{-3}) + \mathit{iIDrkEg}\cdot(5.63\cdot 10^{-4})(\pm 7.88\cdot 10^{-5}) \tag{7}$$

$$\hat{Y}_{DD} = 4.24(\pm 0.47) + \mathit{LhDrjQg}\cdot(-0.40)(\pm 0.26) + \mathit{InDRLHt}\cdot 0.02(\pm 6.15\cdot 10^{-3}) + \mathit{iADRkGg}\cdot 0.06(\pm 7.05\cdot 10^{-3}) + \mathit{IiDDKGg}\cdot(-0.50)(\pm 0.08) \tag{8}$$

$$\hat{Y}_{DT} = 2.78(\pm 0.61) + \mathit{iADREMg}\cdot(-31.29)(\pm 4.33) + \mathit{IHDMLEg}\cdot 0.20(\pm 0.03) + \mathit{IHDDKEg}\cdot(-1.85\cdot 10^{-2})(\pm 1.11\cdot 10^{-2}) + \mathit{iNDRkHt}\cdot(-1.66\cdot 10^{-3})(\pm 6.54\cdot 10^{-4}) \tag{9}$$

$$\hat{Y}_{TP} = 21.50(\pm 3.52) + \mathit{liPRLCg}\cdot 9.74(\pm 1.78) + \mathit{IIPDKCg}\cdot(-14.83)(\pm 3.08) + \tag{10}$$

$$iaPDFEt \cdot 0.42(\pm 0.05) + InDRLHt \cdot (1.78 \cdot 10^{-2})(\pm 6.07 \cdot 10^{-3})$$

$$\hat{Y}_{TD} = 33.37(\pm 6.29) + IhDDJCt \cdot (-0.06)(\pm 7.04 \cdot 10^{-3}) + IsPDLEg \cdot (-59.20)(\pm 12.94) + IMDRLHt \cdot (-0.02)(\pm 6.15 \cdot 10^{-3}) + lsPRLCg \cdot 6.56(\pm 1.34) \quad (11)$$

$$\hat{Y}_{TT} = 27.74(\pm 5.38) + lsPRKEg \cdot 8.88(\pm 2.03) + IBDmKGg \cdot (8.22 \cdot 10^{-4})(\pm 9.99 \cdot 10^{-5}) + IsPRLGg \cdot (-204.95)(\pm 46.85) + IMDRLHt \cdot (-1.93 \cdot 10^{-2})(\pm 6.33 \cdot 10^{-3}) \quad (12)$$

Here Ŷ is the estimated *octan-1-ol/*H_2O partition coefficient and their indices come from the selection method (first letter) and from the survival method (second letter), with P = Proportional; T = Tournament, and D = Deterministic. The number associated with ± is the value to be extracted and added in order to obtain a 95% confidence interval associated with the regression coefficients and the variables *iADREMg*, *iADRkGg*, *iaPDFEt*, *IBDmKGg*, *IhDDJCt*, *IHDDKEg*, *IHDMDHt*, *IHDMLEg*, *IiDDKGg*, *iIDrkEg*, *IIPDKCg*, *IiPDLCg*, *iIPMDHg*, *IMDRLHt*, *imDRlHt*, *IMDRLHt*, *iNDRkHt*, *iNDRlHt*, *InDRLHt*, *ISDRkEg*, *ISDRKHt*, *IsDRLEg*, *iSDRlGg*, *IsPDJEg*, *IsPDLEg*, *IsPRLGg*, *LhDrjQg*, *liPRLCg*, *lsDMLGg*, *lsPRKEg*, *lsPRLCg*, and *lSPRlEg* are *MDF* descriptors, as independent variables. The median time needed per generation proved to be less than 0.1 seconds, and were obtained according to the MDF method [8].

In the present research the number of 20,000 generations was imposed, and thus the optimum solution was identified in less than 10 minutes. The equation of the best models obtained through a complete search is presented in [30]:

$$\hat{Y}_{SS} = 3.04(\pm 0.30) + IIDDKGg \cdot (-0.42)(\pm 0006) + IHDRKEg \cdot 0.04(\pm 2.09 \cdot 10^{-3}) + aHMmjQt \cdot 0.07(\pm 0.02) + aSMMjQq \cdot (-37.50)(\pm 10.10) \quad (13)$$

where SS states for systematic search and *IIDDKGg*, *IHDRKEg*, *aHMmjQt*, *aSMMjQq* are *MDF* descriptors. This equation is golden model for four-variable QSAR since any other than from this complete search for given data and given descriptors cannot be better.

The descriptive statistics for the models (4)-(13) are presented in Table 11.

Thee analysis of the GA−MLR models presented in Table 11 - Equations (4)-(12) we conclude that:

- All combinations of selection and survival strategies provided statistically significant models.
- The analysis of the *GA−MLR−QSAR* models (4)-(12) in terms of the descriptor's contribution to the property of PCBs leads to the data given in Table 12.

Table 12 shows that:

- The top-3 survival-selection strategies, according to the correlation coefficient, are: TP (r^2 = 0.9066), TD (r^2 = 0.9060), and PD (r^2 = 0.9058).

Table 11: MLR models: GA-MLR search vs. complete search (sample size of 206 PCBs)

Param	Eq(4)	Eq(5)	Eq(6)	Eq(7)	Eq(8)	Eq(9)	Eq(10)	Eq(11)	Eq(12)	Eq(13)
R	0.9511 [a]	0.9517 [b]	0.9516 [c]	0.9505 [d]	0.9504 [e]	0.9501 [f]	0.9521 [g]	0.9519 [h]	0.9512 [i]	0.9575 [j]
r^2	0.9045	0.9058	0.9056	0.9034	0.9032	0.9027	0.9066	0.9060	0.9047	0.9168
r^2_{adj}	0.9026	0.9039	0.9037	0.9015	0.9013	0.9008	0.9047	0.9042	0.9028	0.9151
s_{est}	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.24
F_{est}	476‡	483‡	482‡	470‡	469‡	466‡	488‡	485‡	477‡	554‡
t_{int}	9.54‡	11.32‡	11.60‡	10.06‡	17.95‡	8.94‡	12.04‡	10.47‡	10.16‡	19.72‡
t_{X1}	8.59‡	-5.92‡	-16.51‡	-8.37‡	-3.04†	-14.26‡	10.78‡	-16.38‡	8.65‡	-14.80‡
t_{X2}	13.26‡	-8.94‡	-8.31‡	-8.35‡	5.33‡	11.93‡	-9.48‡	-9.02‡	16.23‡	41.73‡
t_{X3}	-6.35‡	-9.88‡	6.07‡	6.47‡	16.30‡	-3.29†	16.23‡	-5.76‡	-8.63‡	6.64‡
t_{X4}	-8.63‡	-16.35‡	-9.46‡	14.08‡	-12.76‡	-5.02‡	5.80‡	9.63‡	-5.99‡	-7.32‡
$r^2_{cv\text{-}loo}$	0.8977	0.8985	0.8977	0.8967	0.8963	0.8956	0.8994	0.8986	0.8975	0.9093
$s_{cv\text{-}loo}$	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.27	0.25
F_{pred}	441‡	445‡	441‡	436‡	434‡	431‡	449‡	445‡	440‡	504‡

X_1, X_2, X_3, and X_4 = structural descriptors (MDF) used as independent variables; r = correlation coefficient, a-j = 95%CI = 95% confidence interval of correlation coefficient; r^2 = determination coefficient; r^2_{adj} = adjusted determination coefficient; s_{est} = standard error of estimate; F_{est} = F-value of estimate; t = t-value; int = intercept; $r^2_{cv\text{-}loo}$ = cross-validation leave-one-out square correlation coefficient; F_{pred}= F-value of predicted; $s_{cv\text{-}loo}$ = standard error of predicted; ‡ $p < 0.0001$; † $p < 0.01$; a = [0.9360; 0.9626]; b = [0.9368; 0.9630]; c = [0.9367; 0.9630]; d = [0.9353; 0.9621]; e = [0.9351; 0.962]; f = [0.9347; 0.9618]; g = [0.9373; 0.9633]; h = [0.9371; 0.9632]; i = [0.9362; 0.9627]; j = [0.9443; 0.9675]

Table 12: Descriptor contribution to the observed property of PCBs

	Eq(4)	Eq(5)	Eq(6)	Eq(7)	Eq(8)	Eq(9)	Eq(10)	Eq(11)	Eq(12)
r^2	90.45	90.58	90.56	90.34	90.32	90.27	90.66	90.6	90.47
IntVia	g-g-t-g	t-g-g-t	t-g-t-g	g-g-t-t	g-t-g-g	t-g-g-t	g-g-t-t	t-g-t-g	g-g-g-t
DAP	G-G-H-E	H-E-E-H	H-C-H-H	E-G-H-E	Q-H-G-G	M-E-E-H	C-C-E-H	C-E-H-C	E-G-G-H
OvrInt	M-M-R-R	R-D-R-R	M-D-R-D	R-R-R-r	r-R-R-D	R-M-D-R	R-D-D-R	D-D-R-R	R-m-R-R
SPS	l-I-I-I	i-I-l-I	I-I-i-i	I-i-I-i	L-I-i-I	i-I-I-i	l-I-i-I	I-I-I-l	l-I-I-I

r^2 - QSAR's coefficient of determination (%);
IntVia = Interaction Via - the 7th letter in the descriptor name: Space (geometry - g), Bonds (topology - t); DAP = Dominant Atomic Property - the 6th letter in descriptor name: Group electronegativity (G), Number of hydrogen atoms adjacent to the investigated atom (H), Atomic electronegativity (E), Cardinality (C), Atomic partial charge (Q), Relative atomic mass (M); OvrInt = Overlapping Interaction - the 4th letter in descriptor name: Frequent and distant interactions (M, m), Sporadic and distant interactions (r, R); SPS = Structure on Property Scale - 1st letter in descriptor name: Identity (I), Logarithm of absolute value (l), Inverse (i), Logarithm (L).

- The top-3 survival-selection strategies, according to the results obtained in leave one-out analysis, are: TP ($r^2_{cv\text{-}loo} = 0.8994$), TD ($r^2_{cv\text{-}loo} = 0.8986$), and PD ($r^2_{cv\text{-}loo} = 0.8985$).
- The top-3 survival-selection strategies, according to the smallest difference between determination coefficient and leave-one-out scores), are: PP ($r^2 - r^2_{cv\text{-}loo} = 0.0068$); DP ($r^2 - r^2_{cv\text{-}loo} = 0.0068$); DD ($r^2 - r^2_{cv\text{-}loo} = 0.0069$), and DT ($r^2 - r^2_{cv\text{-}loo} = 0.0071$).
- The squared cross-validation leave-one-out correlation coefficient proved to be, for each evolutionary strategy, greater than 0.6 [31], and the difference from the determination

coefficient smaller than 0.02. This scenario sustained the reliability of all *GA−MLR−QSAR* models.

The models presented in (4)-(13) were used to predict the *octan-1-ol/*H_2O partition coefficient of three PCBs: 2,3-Dichlorobiphenyl, 3,4'- Dichlorobiphenyl, and 2,2',3,4,4',5-Hexachlorobiphenyl. All values predicted by QSAR models were in-between 4.151 and 9.603 with one exception, represented by eq(4) where proportional selection and proportional survival strategy were used (Table 13). The equation of the most accurate model obtained when proportional selection and survival strategy (eq(4)) provided provided values of 2,3-Dichlorobiphenyl and 3,4'- Dichlorobiphenyl lower than the minimum value in the sample (equal with 4.151). These results suggest that the *GA−MLR* model that used proportional selection and tournament survival strategies is not reliable.

Several information criteria were used to compare the information stored in the *GA−MLR−QSAR* models obtained by pairs of investigated selection-survival strategies, including also the QSAR model obtained by a complete search (Table 14).

Table 13: Predicted values by applying formulas (4)-(13)

Eq	2,3-Dichlorobiphenyl	3,4'- Dichlorobiphenyl	2,2',3,4,4',5-Hexachlorobiphenyl
4	1.9302	2.2518	4.1696
5	4.9165	5.1385	7.1225
6	4.8829	5.4007	7.1958
7	5.0174	5.2201	7.1513
8	4.6834	5.1199	6.8793
9	4.6586	5.0298	6.9328
10	4.9062	5.1712	7.1042
11	4.7944	5.1898	7.0391
12	4.8818	5.4524	7.1502
13	4.4329	4.8505	6.3831

Table 14: Results of information criterion analysis applied on obtained MLR models

IC	Eq(4)	Eq(5)	Eq(6)	Eq(7)	Eq(8)	Eq(9)	Eq(10)	Eq(11)	Eq(12)	Eq(13)
AIC	-550.92	-553.65	-553.20	-548.64	-548.17	-547.08	-555.43	-554.25	-551.33	-579.33
$w_{i\text{-}AIC}$	$6.78\cdot10^{-7}$	$2.66\cdot10^{-6}$	$2.12\cdot10^{-6}$	$2.17\cdot10^{-7}$	$1.71\cdot10^{-7}$	$9.96\cdot10^{-8}$	$6.46\cdot10^{-6}$	$3.59\cdot10^{-6}$	$8.36\cdot10^{-7}$	$1.00\cdot10^{0}$
AIC_{R^2}	2.32	2.31	2.31	2.33	2.34	2.34	2.30	2.31	2.32	2.19
$w_{i\text{-}AICR^2}$	0.0992	0.0998	0.0997	0.0986	0.0985	0.0983	0.1003	0.1000	0.0993	0.1063
AIC_u	-1.64	-1.65	-1.65	-1.63	-1.63	-1.62	-1.66	-1.66	-1.64	-1.78
$w_{i\text{-}AICu}$	0.0992	0.0998	0.0997	0.0986	0.0985	0.0983	0.1003	0.1000	0.0993	0.1063
BIC	-529.52	-532.25	-531.80	-527.24	-526.77	-525.68	-534.02	-532.85	-529.93	-557.92
APC	0.0689	0.0679	0.0681	0.0696	0.0698	0.0701	0.0674	0.0677	0.0687	0.0600
HQC	-544.49	-547.22	-546.77	-542.21	-541.74	-540.65	-549.00	-547.82	-544.91	-572.90
FIT	8.20	8.32	8.30	8.10	8.08	8.03	8.40	8.35	8.22	9.54

IC = information criterion; AIC = Akaike information criteria; AIC_{R2} = AIC based on the determination coefficient; AIC_u = McQuarrie and Tsai corrected AIC; BIC = Bayesian Information Criterion; APC = Amemiya Prediction Criterion; HQC = Hannan-Quinn Criterion; FIT = Kubinyi function; w_i = Akaike weights for model *i*.

The analysis of the results presented in Table 14 revealed the following:

- According to the Akaikes information criteria and the AIC weight, the best model is the model that resulted from the systematic search (13). The model presented in (10) is the best model according to the Akaikes information criteria, when only the *GA−MLR* models are compared.
- According to the Akaikes weights (AIC based on the coefficient of determination and AIC corrected by McQuarrie and Tsai), the *GA−MLR* models presented in (9) is the best model. Moreover, all models have smaller values of these weights compared to the systematic search. Note that the weights identified the models with the smallest relative distance from the "truth".
- According to the Bayesian Information Criterion, the Amemiya Prediction Criterion, the Hannan-Quinn Criterion, and the Kubinyi function, the model that provides most information is the model obtained through a systematic search. The model from (10) is the best model, when only the GA−MLR models are compared.

The analysis of correlation coefficients of the *GA−MLR* models and the model obtained through the systematic search revealed the following:

- The greatest value is obtained by a systematic search.
- The *GA−MLR−QSAR* model with the highest correlation coefficient is (10).
- With two exceptions, (8) and (9), the correlation coefficients of the *GA−MLR−QSAR* models do not have a statistically significant difference ($p \geq 0.0591$) compared to the correlation coefficient of the model obtained through a systematic search, at a significance level of 5%, by Steiger's Z test:

  $Z_{(13)-(4)}(p) = 1.50276\ (0.0665)$, $Z_{(13)-(5)}(p) = 1.34603\ (0.0891)$,
  $Z_{(13)-(6)}(p) = 1.36491\ (0.0861)$, $Z_{(13)-(7)}(p) = 1.56277\ (0.0591)$,
  $Z_{(13)-(8)}(p) = 1.74524\ (0.0405)$, $Z_{(13)-(9)}(p) = 1.79056\ (0.0367)$,
  $Z_{(13)-(10)}(p) = 1.2725\ (0.1016)$, $Z_{(13)-(11)}(p) = 1.32485(0.0926)$,
  $Z_{(13)-(12)}(p) = 1.45678\ (0.0726)$.

- The smallest difference between two correlation coefficients is 0.00536 and it was obtained for the model presented by (10) compared with a systematic search.

In this study, we used GA for searching the MDF descriptors space and the MLR for fitness evaluation. Several guidelines that comprise how to validate a QSAR model have been previously published [32, 33]. To predict of the outcome is just one of the aim of linear regression analysis, beside identification of the strength of the linear association between outcome and factors of interest or to identify those factors that affect the outcome [34]. Beside recommendation of assessment the model on an external data-set [32, 33], several parameters have been reported as useful in evaluation of predictive power a QSAR model (such as predictive square correlation coefficients in training, test sets and external sets [35, 36, 37], external predictive ability [38, 39], predictive power by Fisher's approach [10]). Furthermore, a series of classification methods could be useful whenever appropriate [28, 41]. The validation of the GA-MLR models was beyond the aim of this study since it has been previously proved [30]. Current research in our laboratory is on implementation of a GA-MLR able to identify the best performing model with highest performances both in training and test sets as well as in external sets.

4. Conclusions

The proposed genetic algorithm for multiple linear regressions with families of descriptors for structure-property/activity relationships was successfully implemented and proved its efficiency in *QSAR* models identification. Different selection and survival strategies created different partitions of the entire population of genotypes, since different pathways of evolution can be followed under the pressure of various environmental factors. Moreover, the resulting models proved to have different estimation and prediction abilities, and some $GA-MLR$ models were revealed not to be significantly different from the golden *QSAR* model obtained through a complete search. This result shows that, even if the evolution follows different pathways, it is likely to reach the same stages of development. The $GA-MLR-QSAR$ model obtained with tournament selection and proportional survival proved to be the closest to the model obtained by complete search. Moreover, tournament selection and proportional survival seem to be the natural way of evolution since it proved to be the most effective and since the nature always evolve to maximize the chances of adaptation.

References

[1] J. C. Dearden, M. T. D. Cronin, Quantitative structure-activity relationships (QSAR) in drug design, in: H. J. Smith, H. Williams (Eds.), *Introduction to the Principles of Drug Design and Action*, CRC Press, Boca Raton, 2006, pp. 185–209.

[2] O. Nicolotti, I. Giangreco, A. Introcaso, F. Leonetti, A. Stefanachi, A. Carotti, Strategies of multi-objective optimization in drug discovery and development, *Expert Opin. Drug Discov.* **6** (2011) 871–884.

[3] B. Buszewski, M. Michel, Quantitative structure-retention relationship studies as an analytical tool in the determination and modeling of pesticide residues in plant organisms, *J. AOAC Int.* **93** (2010) 1703–1714.

[4] M. T. H. Khan, I. Sylte, Predictive QSAR modeling for the successful predictions of the ADMET properties of candidate drug molecules, *Curr. Drug Discov. Technol.* **4** (2007) 141–149.

[5] L. Jäntschi, S. D. Bolboacă, M. V. Diudea, Chromatographic retention times of polychlorinated biphenyls: from structural information to property characterization, *Int. J. Mol. Sci.* **8** (2007) 1125–1157.

[6] M. H. Fatemi, E. Baher, A novel quantitative structure-activity relationship model for prediction of biomagnification factor of some organochlorine pollutants, *Mol. Divers.* **13** (2009) 343–352.

[7] L. Saghaie, M. Shahlaei, A. Fassihi, A. Madadkar-Sobhani, M. B. Gholivand, A. Pourhossein, QSAR analysis for some diaryl-substituted pyrazoles as CCR2 inhibitors by GA-stepwise MLR, *Chem. Biol. Drug Des.* **77** (2011) 75–85.

[8] L. Jäntschi, MDF - A new QSAR/QSPR molecular descriptors family, *Leonardo J. Sci.* **3** (2004) 68–85.

[9] L. Jäntschi, G. Katona, M. V. Diudea, Modeling molecular properties by Cluj indices, *MATCH Commun. Math. Comput. Chem.* **41** (2000) 151–188.

[10] S. D. Bolboacă, L. Jäntschi, Comparison of QSAR performances on carboquinone derivatives, *Sci. World J.* **9** (2009) 1148–1166.

[11] R. Sestraş, L. Jäntschi, S. Bolboacă, Poisson parameters of antimicrobial activity: A quantitative structure-activity approach, *Int. J. Mol. Sci.* **13** (2012) 5207–5229.

[12] P. V. Khadikar, N. V. Deshpande, P. P. Kale, A. Dobrynin, I. Gutman, G. J. Domotor, The Szeged index and an analogy with the Wiener index, *J. Chem. Inf. Comput. Sci.* **35** (1995) 547–550.

[13] M. V. Diudea, Cluj matrix invariants, *J. Chem. Inf. Comput. Sci.* **37** (1997) 300–305.

[14] R. A. Fisher, The goodness of fit of regression formulae and the distribution of regression coefficients, *J. R. Stat. Soc. C* **85** (1922) 597–612.

[15] W. S. Gosset, The probable error of a mean, *Biometrika* **6** (1908) 1–25.

[16] C. M. Jarque, A. K. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence, *Econ. Lett.* **7** (1981) 313–318.

[17] L. Jäntschi, S. D. Bolboacă, Molecular descriptors family on structure activity relationships 6. Octanol-water partition coefficient of polychlorinated biphenyls, *Leonardo El. J. Pract. Technol.* **5** (2006) 71–86.

[18] R. Hoffmann, An extended Hückel theory. I. Hydrocarbons, *J. Chem. Phys.* **39** (1963) 1397–1412.

[19] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy, J. J. P. Stewart, Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model, *J. Am. Chem. Soc.* **107** (1985) 3902–3909.

[20] L. Jäntschi, S. D. Bolboacă, Distribution fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics, *Bull. UASVM Hort.* **66** (2009) 691–697.

[21] C. M. Jarque, A. K. Bera, Efficient tests for normality, homoscedasticity and serial independence of regression residuals, *Econ. Lett.* **6** (1980) 255–259.

[22] R. A. Fisher, On the interpretation of $\chi 2$ from contingency tables, and the calculation of P, *J. R. Stat. Soc.* **85** (1922) 87–94.

[23] R. A. Fisher, The mathematical distributions used in the common tests of significance, *Econometrica* **3** (1935) 353–365.

[24] K. Pearson, 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** (1900) 157–175.

[25] S. D. Bolboacă, L. Jäntschi, Modelling the property of compounds from structure: statistical methods for models validation, *Environ. Chem. Lett.* **6** (2008) 175–181.

[26] R. D. Cramer III, J. D. Bunce, D. E. Patterson, I. E. Frank, Crossvalidation, bootstrapping, and partial least squares compared with multiple regression in conventional QSAR studies, *Quant. Struct.-Act. Relat.* **7** (1988) 18–25; erratum **7** (1988) 91.

[27] A. Golbraikh, A. Tropsha, Beware of q2!, *J. Mol. Graphics Mod.* **20** (2002) 269–276.

[28] S. D. Bolboacă, L. Jäntschi, predictivity approach for quantitative structure-property models. Application for blood-brain barrier permeation of diverse drug-like compounds, *Int. J. Mol. Sci.* **12** (2011) 4348–4364.

[29] J. H. Steiger, Tests for comparing elements of a correlation matrix, *Psychol. Bull.* **87** (1980) 245–251.

[30] L. Jäntschi, S. D. Bolboacă, R. E. Sestraş, Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls, *J. Mol. Model.* **16** (2010) 377–386.

[31] A. Tropsha, Best practices for QSAR model development, validation, and exploitation, *Mol. Inf.* **29** (2010) 476−488.

[32] J. S. Jaworska, M. Comber, C. Auer, C. J. Van Leeuwen, Summary of a workshop on regulatory acceptance of (Q)SARs for human health and environmental endpoints, *Environ. Health Persp.* 111 (2003) 1358−1360.

[33] Guidance Document on the Validation of (Quantitative) Structure-Activity Relationship [(Q)SAR] Models. [online] [Accessed 31 May 2012]. ENV/JM/MONO (OECD Environment Health and Safety Publications) 2007;2. Available from: URL:

http://www.oecd.org/officialdocuments/displaydocumentpdf/?cote=env/jm/mono(2007)2&doclanguage=en

[34] Y. M. Chan, Biostatistics 201: Linear regression analysis, *Singapore Med. J.* **45** (2004) 55−64.

[35] L. M. Shi, H. Fang, W. Tong, J. Wu, R. Perkins, R. M. Blair, W. S. Branham, S. L. Dial, C. L. Moland, D. M. Sheehan, QSAR models using a large diverse set of estrogens, *J. Chem. Inf. Comput. Sci.* **41** (2001) 186–195.

[36] S. D. Bolboacă, L. Jäntschi, The effect of leverage and/or influential on structure-activity relationships, *Comb. Chem. High Throughput Screen.* **16** (2013) 288–297.

[37] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, *Molecules* **17** (2012) 4791–4810.

[38] N. Chirico, P. Gramatica, Real external predictivity of QSAR models: How to evaluate It? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, *J. Chem. Inf. Model.* **51** (2011) 2320–2335.

[39] N. Chirico, P. Gramatica, Real external predictivity of QSAR models. Part 2. New intercomparable thresholds for different validation criteria and the need for scatter plot inspection, *J. Chem. Inf. Model.* **52** (2012) 2044–2058.

[40] R. A. Fisher, The goodness of fit of regression formulae, and the distribution of regression coefficients, *J. R. Stat. Soc.* **85** (1922) 597–612.

[41] K. Mansouri, T. Ringsted, D. Ballabio, R. Todeschini, V. Consonni, Quantitative structure–activity relationship models for ready biodegradability of chemicals, *J. Chem. Inf. Model.* **53** (2013) 867–878.