J Mol Model (2010) 16:377–386
DOI 10.1007/s00894-009-0540-z

ORIGINAL PAPER

Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls

Lorentz Jäntschi · Sorana D. Bolboacă · Radu E. Sestraş

Received: 25 March 2009 / Accepted: 13 May 2009 / Published online: 17 July 2009

Abstract A genetic algorithm was developed and assessed in order to select pairs of proper structural descriptors able to estimate and predict octanol-water partition coefficients of polychlorinated biphenyls (PCBs). The molecular descriptors family was calculated for a sample of 206 PCBs. The problem of searching for the proper descriptors in order to identify structure-activity relationships was translated in genetic terms. The following parameters were imposed in the genetic algorithm (GA) search: sample size − 12, number of variables in multivariate linear regression − 4, imposed adaptation requirements − 3 criteria, maximum number of generations − 50,000, selection strategy − tournament, probability of parent/child mutation − 0.05, number of genes implied in the mutation − 2, optimization parameter - determination coefficient, optimization score - minimum in the sample, and optimization objective - maximum. The highest determination coefficient was obtained in the generation 17,277. Twenty-one evolutions were studied until the optimum solution was obtained. The model identified by the implemented genetic algorithm proved not to be statistically different from the model identified through complete search ($Z_{Steiger}$=1.37, p=0.0861). According to this GA model, the relationship between the structure of PCBs and octanol-water partition coefficients was of geometric and topological nature as previously revealed by the complete search. The genetic algorithm proved its ability to identify two pairs of molecular descriptors able to characterize the relationship between the structure of PCBs and the octanol-water partition coefficient.

Keywords Genetic algorithm (GA) · Molecular descriptors family (MDF) · Multivariate linear regression (MLR) · Polychlorinated biphenyls (PCBs)

Electronic supplementary material The online version of this article (doi:10.1007/s00894-009-0540-z) contains supplementary material, which is available to authorized users.

L. Jäntschi (✉)
Technical University of Cluj-Napoca,
103-105 Muncii Bvd,
400641 Cluj-Napoca, Romania
e-mail: lori@academicdirect.org

S. D. Bolboacă
Department of Medical Informatics and Biostatistics,
"Iuliu Haţieganu" University of Medicine and Pharmacy
Cluj-Napoca,
6 Louis Pasteur,
400349 Cluj-Napoca, Romania

R. E. Sestraş
University of Agricultural Sciences and Veterinary Medicine
Cluj-Napoca,
3-5 Mănăştur,
400372 Cluj-Napoca, Romania

Introduction

The methods of structure-property/activity relationships (QSP/AR), which proved their utility in different research fields, were introduced over 40 years ago [1]. Various approaches to the generation and calculation of descriptors were developed and proved their abilities to estimate and predict different properties/activities on different classes of chemically active compounds: comparative molecular field analysis (CoMFA) [2] and its variant comparative molecular similarity indices analysis (CoMSIA) [3]); weighted holistic invariant molecular (WHIM) [4] and its variant molecular surface weighted holistic invariant molecular (MS-WHIM) [5]);minimal topological distance (MTD) and its variant minimum steric difference (MSD) [6]); molecular descriptor family (MDF) [7-9]; S-SAR (Spectral Structure-Activity Rela-

tionship) [10]); fragment-based two dimensional QSAR (FB-QSAP) [11]; multiple field three dimensional QSAR (MF-3D-QSAP) [12] & three-dimensional holographic vector of atomic interaction field (3D-HoVAIF) [13]; four dimensional QSAR and five dimensional QSAR (4D-QSAR & 5D-QSAR) [14, 15] (*etc*.

Different strategies are used to select the descriptors with the highest ability to explain the property/activity of interest (e.g., principal component analysis [16], factor analysis [17], stepwise multiple linear/non-linear regression [18-20]). Genetic algorithms [21-23], self-organizing-maps [24], machine learning and artificial intelligence [25, 26] were also proved to be useful techniques for selecting descriptors.

Polychlorinated biphenyls (PCBs) are chlorinated congeners known as stable organic industrial chemicals widely used as insulating, hydraulic and lubricating fluids, heat exchanger fluids as well as additives in adhesive inks and paints [27]. SAR analyses had previously been performed on PCBs (octanol/soil-water partition coefficient [28, 29], bio concentration factors [30], biodegradation rate [31], liquid vapour pressure [32], retention time [33], gas-particle partitioning in the atmosphere [34]) due to their persistence in the environment [35], bioaccumulation [36, 37], and toxicities or health effects [38].

The objective of the paper was to assess the ability of our genetic algorithm to select two pairs of structural descriptors with the highest contribution to the octanol-water partition coefficient on a sample of polychlorinated biphenyls.

Materials and methods

Polychlorinated biphenyls and SAR-MDF

The octanol-water partition coefficient of polychlorinated biphenyls (PCBs, see generic structure in Fig. 1) had previously been modeled by using complex information obtained from the structure of compounds using the MDF approach [39]. The MDF approach was integrated into a series of home-made software in order to generate and calculate molecular descriptors [40].

The octanol-water partition coefficient expressed in logarithmic scale ($\log K_{ow}$, $K_{ow}=C_{ow}/C_{wo}$, where C_{ow} is the concentration of the solute in octanol saturated with water, and C_{wo} is the concentration of the solute in water saturated with octanol [41]) was taken from a previously reported research [42] (see supplementary Table 1). The sample size of the investigated PCBs comprised 206 compounds out of the total number of 209 PCBs, due to the non availability or unsteadiness of the measured octanol-water partition coefficients (two dichlorobiphenyls and one hexachlorobiphenyls).

The molecular structure of the PCBs was drawn by using HyperChem.[1] The semi-empirical extended Hückel model was applied in order to calculate the partial charges [43]. The semi-empirical single-point Austin method (AM1) was applied in order to optimize compound geometry [44] (step 1). The optimized molecules, which store information on the topology, geometry and charge distribution of the PCBs, represented the primary data for generating the MDF (step 2). The MDF was generated and comprised a number of 787,968 members resulted from the multiplication of all the case sensible operators used to generate them ($2 \times 6 \times 24 \times 6 \times 4 \times 19 \times 6$) (step 3, see Table 1). A selection algorithm able to identify valid descriptors was applied and the descriptors containing little or redundant information were eliminated (step 3). The descriptors whose values were identical or nearly identical were also removed (step 3). SAR models with two pairs of descriptors were identified through complete search (all possible solutions, combination of the 787,968 MDF members taken 4 – $1.61 \cdot 10^{22}$) by using MDF members to explain the octanol-water partition coefficient of PCBs (step 4). Valid SAR models (in terms of model significance, significance of the model coefficients, *etc*.) underwent cross-validation analysis (step 5).

The highest performance SAR model obtained through complete search is presented in Eq. (1):

$$
\begin{aligned}
&\widehat{Y}_{CS} = 3.04(\pm 0.30) + IIDDKGg \cdot (-0.42)(\pm 0.06) + IHDRKEg \cdot 0.04\left(\pm 2.09 \cdot 10^{-3}\right) + \\
&\qquad aHMmjQt \cdot 0.07(\pm 0.02) + aSMMjQg \cdot (-37.50)(\pm 10.10) \\
&r[95\%CI] = 0.9575[0.9443 - 0.9675]; r^2 = 0.9168; n = 206; \\
&r^2_{adj} = 0.9151; s_{est} = 0.24; F_{est}(p) = 554\left(2.75 \cdot 10^{-107}\right); \\
&t_{\text{int}}(p) = 19.72\left(7.27 \cdot 10^{-49}\right); t_{X1}(p) = -14.80\left(5.09 \cdot 10^{-34}\right); t_{X2}(p) = 41.73\left(5.97 \cdot 10^{-101}\right); \\
&t_{X3}(p) = 6.64\left(2.89 \cdot 10^{-10}\right); t_{X4}(p) = -7.32\left(5.86 \cdot 10^{-32}\right); \\
&r^2_{cv-loo} = 0.9093; F_{pred}(p) = 504\left(8.08 \cdot 10^{-105}\right); s_{cv-loo} = 0.25; r^2 - r^2_{cv-loo} = 0.0075
\end{aligned}
\tag{1}
$$

where $\widehat{Y}_{CS}$=octanol-water partition coefficient by Eq. (1), CS = complete search; IIDDKGg (X1), IHDRKEg (X2),

[1] http://hyper.com/products/

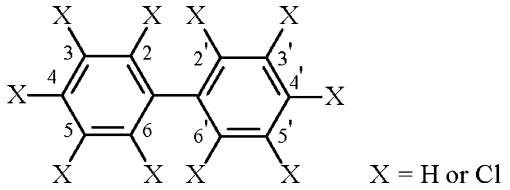

Fig. 1 Generic structure of PCBs

aSMMjQg (X3) and aHMmjQt (X4) = MDF members; r = correlation coefficient, 95%CI=95% confidence interval of correlation coefficient; r^2 = determination coefficient; n = sample size; r^2_{adj} = adjusted determination coefficient; s_{est} = standard error of estimate; $F_{est}(p)$ = F value of estimate (significance); t = t-value; int = intercept; r^2_{cv-loo} = cross − validation leave − one − out squared correlation coefficient; $F_{pred}(p)$ = F value of predicted (significance); $s_{cv\text{-}loo}$ = standard error of predicted.

Table 1 Genetic representation of the MDF genotype

Gene	Encodes ...	Values
d	Distance operator	*g* = geometric distance
		t = topological distance
p	Atomic property used to construct the phenotype	*M* = relative atomic mass
		Q = atomic partial charge, semi-empirical Extended Hückel model, single point approach
		C = cardinality, trivial atomic property - its value for any atom is equal with 1
		E = atomic electronegativity - relative electronegativity value on the Sanderson scale
		G = group electronegativity - value obtained by calculating the geometric mean of the electronegativity associated to the atoms group adjacent to the investigated atom
		H = number of hydrogen atoms adjacent to the investigated atom
I	Interaction descriptor	could take one of the following twenty-two values: $D(d)$, $d(1/d)$, $O(p_1)$, $o(1/p_1)$, $P(p_1p_2)$, $p(1/p_1p_2)$, $Q(\sqrt{p_1p_2})$, $q(1/\sqrt{p_1p_2})$, $J(p_1d)$, $j(1/p_1d)$, $K(p_1p_2d)$, $k(1/p_1p_2d)$, $L(d\sqrt{p_1p_2})$, $l(1/d\sqrt{p_1p_2})$, $V(p_1/d)$, $E(p_1/d_2)$, $W(p_1^2/d)$, $w(p_1p_2/d)$, $F(p_1^2/d^2)$, $f(p_1p_2/d^2)$, $S(p_1^2/d^3)$, $s(p_1p_2/d^3)$, $T(p_1^2/d^4)$, $t(p_1p_2/d^4)$; where *d* = distance operator and *p* = atomic property
O	Overlapping interactions	Six values were implemented, two for the models with sporadic and distant interactions (*R* and *r*), two for the models with frequent and distant interactions (*M* and *m*, respectively), and two for the models with frequent and closed interactions (*D* and *d*)
f	Algorithm of molecular fragmentation on atom pairs	*P* = fragmentation based on paths
		D = fragmentation based on distances
		M = fragmentation in maximal fragments
		m = fragmentation in minimal fragments - trivial fragments with one atom
M	Global overlapping of fragment interaction	▪ group of values: *m* = minimum value, *M* = maximum value, *n* = lowest absolute value, *N* = highest absolute value;
		▪ group of mean: *S* = sum, *A* = arithmetic mean according to the number of fragment properties, *a* = arithmetic mean according to the number of fragments, *B* = arithmetic mean according to the number of atoms, *b* = arithmetic mean according to the number of bonds;
		▪ geometric group: *P* = multiplication, *G* = geometric mean according to the number of fragment properties, *g* = geometric mean according to the number of fragments, *F* = geometric mean according to the number of atoms, f = geometric mean according to the number of bonds;
		▪ harmonic group: *s* = harmonic sum, *H* = harmonic mean according to the number of fragment properties, *h* = harmonic mean according to the number of fragments, *I* = harmonic mean according to the number of atoms, i = harmonic mean according to the number of bonds
L	Linearization operators	*I* = identity
		i = inverse
		A = absolute value
		a = inverse of absolute value
		l = logarithm of absolute value
		L = logarithm.
		Obs.: One of these operators was applied to evaluate the fittest of each descriptor

These scores were calculated and displayed for each generation in order to analyze the GA.

The model identified through complete search used three molecular descriptors that referred to the geometry (IIDDKG**g**, IHDRKE**g**, and aSMMjQ**g**) and the topology of compounds (aHMmjQ**t**). In terms of atomic property, aHMmj**Q**t, and aSMMj**Q**g referred to partial charge, IIDDK**G**g to group electronegativity and IHDRK**E**g to atomic electronegativity.

Generic algorithm

A heuristic search based on a genetic algorithm was proposed for identifying multivariate linear regression models. The following key elements were taken into consideration when the genetic algorithm was implemented: genetic model (genotype-phenotype dualism) [45]; mapping (character-gene dualism) [46]; mutation (random [47], deliberate [48], environment [49]); and survival of the fittest [50].

Three criteria were taken into account when the heuristic algorithm was developed: speed (the solution had to be obtained in a short time), precision (the solution had to be as close as possible to the global optimum, see the Eq. (1)), and applicability domain (entry data sets from molecular descriptor family).

The structure-activity relationship that had to be solved was: which SAR best described the octanol-water partition coefficient of PCBs as a function of the structure based on the structural information of PCBs? The molecular topology of each compound was based on the chemical bonds between atoms and molecular geometry. Molecular topology was obtained by applying the approximate models of quantum and molecular physics and was regarded as structural information.

The following parameters were assigned in order to solve the problem:

- ***Search space:*** MDF of PCBs.
- ***Initial sample*****:** Each MDF member is characterized by seven genes (one for each letter in the descriptor's name, see Table 1). The volume of the initial sample was of 12 descriptors.
- ***Adaptation***: Genotypes were transformed into phenotypes by checking their values in the environment of the experimental (observed) data and by applying the linearization operator. Three criteria were used to adapt each genotype: minimum of the absolute variance (a ratio of measured data variance; 0.1 was used); maximum of the Jarque-Bera value (no higher than the Jarque-Bera value ratio on the measured data [51], 1.0 was used); and minimum value of the determination coefficient between estimated and experimental data (0.1 was used).
- ***Fittest and phenotyping***. The fittest score of an individual was defined as the minimum MLR determination coefficient in relation to all the other individuals in the sample. The following fittest scores were defined and calculated in order to characterize an individual:
- Sum of residuals in estimate: $se = \Sigma\left|\widehat{Y}i - Yi\right|p$ (objective: minimum); where p=1.
- Determination coefficient: $r^2 = \left(r^2\left(Y, \widehat{Y}\right)\right)^p$: (objective: maximum); where p = 2.
- Hölder mean of student t-parameters associated to the intercept and coefficients of the MLR model: $Mt = \left(1/n\sum_{i=1}^{n} t_i^p\right)^{1/p}$ (objective: maximum); where p=1 (for this value the arithmetic mean was obtained), t_i = Student t-parameter associated with the regression coefficients (including the intercept).
- Quantity of explained or un-explained entropy: Hr=H (r^2,1-r^2,p) (objective: minimum).
- ***Selection:*** Sample pairs of individuals were selected for reproduction. The binary tournament selection was used (two individuals competed for selection; the fittest criterion was applied). The selected individuals underwent genotype crossover and mutation. The equation used for selection was: $p_i=f_i/\Sigma_i f_i$ (where p_i = probability used for selection, f_i = fittest score).
- ***Crossover & mutation:*** Two genotypes were selected and crossed over (two point crossover = randomly selected two crossover points) and two offspring were obtained. A mutation (which implied the randomized selection of the gene that had to be mutated and the random mutation of the selected gene) was applied to the parents (with a low probability p_1) or to an offspring (with a low probability p_2). In our study, two genes were implied in mutation. The imposed probability of parent and/or child mutations was set to 0.05.
- ***Survival:*** Valid offspring replaced two individuals in the sample in the following order: dead individuals, parents, others (deterministic type). At the end of this step, an evolution cycle was completed and a new generation of the sample was created.
- ***Evolution:*** The identified solutions were stored in the database. The cycle continued to adapt until the imposed maximum number of generations (50,000 for this experiment) or an imposed value of the best (or worst) fittest score were obtained.

The GA-MLR objective was to obtain the four-member SAR model with the highest determination coefficient. The GA was implemented as a Windows based FreePascal application with MySQL connectivity for fetching the data.

The maximization of the minimum value of the determination coefficient obtained from genetic algorithm - multiple linear regression (GA-MLR) was the criterion used for optimization.

The partial F test [52] and Steiger's Z test [53] were applied in order to test if adding a new descriptor significantly improved the estimation of octanol-water partition coefficients of PCBs.

The GA-MLR model was also subject to principal component and classification analysis by using Statistica 8.0 software (significance level of 5%) in order to identify factors within the descriptors of the GA-MLR.

The GA-MLR SAR model was analyzed by applying two internal validation methods: leave-one-out and leave-many-out analyses (training versus test analysis) [54]. The leave-many-out analysis was performed by randomly splitting the PCBs sample into training and test sets.

The Fisher Z test was used to compare the GA-MLR correlation coefficient with the correlation coefficient identified through complete search (H_0 hypothesis: the correlation coefficient obtained in GA-MLR was not different from the correlation coefficient obtained in the complete search) [53]. The Z critical value for a significance level of 5% was equal to 1.96($Z_{calculated} \in (-\infty, 1.96] \cup [1.96, +\infty)$, then the H_0 is rejected).

Results

The improvement of the determination coefficient was observed in 21 generations out of the total number of 50,000 generations. The characteristics of the GA-MLR in the generations that underwent evolution expressed as: the generation in which an improved r^2 was obtained, the number of genotypes and phenotypes in cultivar; the number of valid regressions, the correlation coefficient and associated 95% confidence interval, the sum of residuals in estimate, the Hölder mean of t values, and the quantity of explained and un-explained entropy are presented in Table 2. The evolution of the GA-MLR in terms of determination coefficients is graphically represented in Fig. 2.

The GA-MLR model with the highest correlation coefficient was obtained at generation 17,277 (out of 50,000). The characteristics of the GA-MLR model are presented in Eq. (2):

$$\begin{aligned}
\hat{Y}_{GA-MLR} &= 11.15(\pm 1.895) - \text{iIPRJCg} \cdot 1.97(\pm 0.412) - \text{IiPDLCg} \cdot 9.07(\pm 2.152) - \\
&\quad \text{IMDRLHt} \cdot 0.02(\pm 0.006) - \text{IhDDJCt} \cdot 0.06(\pm 0.007) \\
r[95\%CI] &= 0.9516[0.9367 - 0.9630]; r^2 = 0.9056; n = 206; \\
r^2_{adj} &= 0.9037; s_{est} = 0.26; F_{est}(p) = 482\left(9.31 \cdot 10^{-102}\right); \\
t_{int}(p) &= 11.60\left(3.72 \cdot 10^{-24}\right); t_{X1}(p) = -9.46\left(8.65 \cdot 10^{-18}\right); t_{X2}(p) = -8.31\left(1.39 \cdot 10^{-14}\right); \\
t_{X3}(p) &= -6.07\left(6.12 \cdot 10^{-9}\right); t_{X4}(p) = -16.51\left(2.89 \cdot 10^{-39}\right); \\
r^2_{cv-loo} &= 0.8977; F_{pred}(p) = 441\left(1.79 \cdot 10^{-99}\right); s_{cv-loo} = 0.27; r^2 - r^2_{cv-loo} = 0.0079
\end{aligned} \tag{2}$$

where $\hat{Y}_{GA-MLR}$ = estimated octanol-water partition coefficient(expressed in logarithmic scale) by the GA-MLR model; iIPRJCg (X1), IiPDLCg (X2), IMDRLHt (X3), IhDDJCt (X4) = MDF members of the GA-MLR model.

We analyzed whether adding a new descriptor significantly improved the estimation of the octanol-water partition coefficient. The results are presented in the supplementary Table 2.

Four factors were identified when principal components and classification analysis of the descriptors used by Eq. (2) was performed. The contribution of each descriptor to the factors was:

- Factor 1: 0.254859 - iIPRJCg, 0.246527 - IiPDLCg, 0.245152 - IMDRLHt, 0.253462 - IhDDJCt;
- Factor 2: 0.204077 - iIPRJCg, 0.294446 - IiPDLCg, 0.297065 - IMDRLHt, 0.204411 - IhDDJCt;
- Factor 3: 0.166662 - iIPRJCg, 0.185968 - IiPDLCg, 0.319939 - IMDRLHt, 0.327431 - IhDDJCt;
- Factor 5: 0.374402 - iIPRJCg, 0.273059 - IiPDLCg, 0.137844 - IMDRLHt, 0.214695 - IhDDJCt.

Table 2 Summary of GA performances according to generation

Evol	Geno	Pheno	NAli	r[a] [95% CI]	se	Mt	Hr
0	12	18	733	0.9413 [0.9233–0.9550]	16.13	12.48	0.5119
23	12	22	1317	0.9438 [0.9266–0.9570]	15.45	7.92	0.4975
210	12	22	1962	0.9442 [0.9271–0.9572]	15.36	8.06	0.4955
234	12	23	2700	0.9447 [0.9278–0.9577]	15.21	8.36	0.4924
273	12	20	1197	0.9448 [0.9279–0.9578]	15.18	8.42	0.4917
478	12	19	1002	0.9450 [0.9281–0.9579]	15.15	8.47	0.4909
3179	12	22	2133	0.9450[0.9281–0.9579]	15.15	8.47	0.4909
3962	12	20	956	0.9450 [0.9281–0.9579]	15.15	8.47	0.4909
4662	12	21	1194	0.9463 [0.9298–0.9589]	14.80	10.30	0.4834
4808	12	16	451	0.9463 [0.9298–0.9589]	14.79	10.32	0.4833
6283	12	19	628	0.9463 [0.9298–0.9589]	14.78	10.23	0.4831
9229	12	16	401	0.9464 [0.9299–0.9589]	14.77	10.33	0.4827
10499	12	19	823	0.9465 [0.9301–0.9591]	14.73	10.30	0.4818
15605	12	16	341	0.9468 [0.9304–0.9592]	14.67	10.54	0.4805
15779	12	19	846	0.9471 [0.9308–0.9595]	14.58	9.82	0.4786
15802	12	20	918	0.9471 [0.9309–0.9595]	14.57	9.92	0.4784
16813	12	20	787	0.9475 [0.9314–0.9598]	14.47	10.29	0.4761
16857	12	20	1073	0.9482 [0.9323–0.9604]	14.27	11.08	0.4716
16861	11	18	656	0.9510 [0.9359–0.9625]	13.53	12.52	0.4549
17232	12	19	1255	0.9512 [0.9362–0.9627]	13.46	12.57	0.4535
17274	12	21	1889	0.9516 [0.9366–0.9629]	13.37	12.94	0.4514
17277	10	17	739	0.9516 [0.9367–0.9630]	13.36	12.99	0.4511

Evol = generation in which an improved r^2 was obtained;

Geno = number of genotypes in cultivar;

Pheno = number of alive phenotypes in cultivars (number of distinct phenotypes in cultivar);

NAli = number of valid regressions;

r = correlation coefficient; [95% CI]=95% confidence interval associated to the correlation coefficient;

[a] = correlation coefficients are different at 15 decimals;

se = sum of residuals in estimate;

Mt = Hölder mean of t values;

Hr = quantity of explained and un-explained entropy

The projection of the descriptors used by Eq. (2) on the plane of the two most important factors is presented in supplementary Fig. 1.

Training versus test analysis was applied as the internal validation method of the GA-MLR model by randomly splitting PCBs compounds as 154(training):52(test). Statistics of these models are presented in Eqs. (3) and (4):

$$\begin{aligned}\hat{Y}_{GA-Tr} &= 12.52(\pm 2.415) - \text{iIPRJCg} \cdot 2.30(\pm 0.524) - \text{IiPDLCg} \cdot 10.65(\pm 2.748) - \\ &\quad \text{IMDRLHt} \cdot 0.02(\pm 0.007) - \text{IhDDJCt} \cdot 0.06(\pm 0.008) \\ r_{tr}[95\%CI] &= 0.9514[0.9337 - 0.9644]; r_{tr}^2 = 0.9051; n_{tr} = 154; \\ r_{adj-tr}^2 &= 0.9026; s_{tr} = 0.25; F_{tr}(p) = 355\left(4.35 \cdot 10^{-75}\right); \\ t_{int}(p) &= 10.24\left(5.72 \cdot 10^{-19}\right); t_{X1}(p) = -8.67\left(6.86 \cdot 10^{-15}\right); t_{X2}(p) = -7.66\left(2.19 \cdot 10^{-12}\right); \\ t_{X3}(p) &= -5.36\left(3.16 \cdot 10^{-7}\right); t_{X4}(p) = -14.49\left(3.03 \cdot 10^{-30}\right).\end{aligned} \quad (3)$$

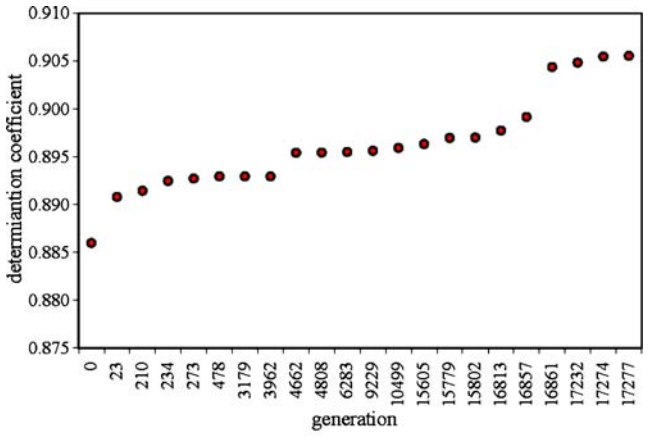

Fig. 2 Evolution of r^2: PCBs sample

$$r_{ts}[95\%CI] = 0.9583[0.9280 - 0.9759]; r_{ts}^2 = 0.9183; n_{ts} = 52; r_{adj-ts}^2 = 0.9113; s = 0.27; F_{ts}(p) = 132\left(6.27 \cdot 10^{-25}\right) \quad (4)$$

where $\widehat{Y}_{GA-Tr}$ = estimated octanol - water partition coefficient(expressed in logarithmic scale)in training set; $\widehat{Y}_{GA-Ts}$ = predicted octanol - water partition coefficient (expressed in logarithmic scale)in test set; s = standard error; F(p) = F-value (significance); X1 = iIPRJCg, X2 = IiPDLCg, X3 = IMDRLHt, X4 = IhDDJCt; int = intercept; t = t value; tr = training set; ts = test set.

Additional results obtained by randomly splitting PCBs sample in training and test sets, starting with 120 compounds in the training set, and applying an increment of two until the number of compounds in the training set was 180 are presented in supplementary Table 3.

The abilities of the models presented in Eqs. (3) and (4) is presented in Fig. 3.

The correlation coefficient identified by the GA-MLR method (see Eq. (2)) was not statistically different compared to the correlation coefficient of the SAR model obtained by complete search (see Eq. (1)) ($Z_{Steiger}$=1.37, df=203, p=0.0861). The GA-MLR model proved to be slightly better in terms of the difference between the observed and the estimated activity (the lowest value of residuals was obtained in 108 cases out of 206) compared to the SAR model reported in Eq. (1) (98 cases). The difference between the measured and estimated/predicted varied from −1.1999 to 1.3383 for the GA-MLR model and from −0.8173 to 0.9743 for the SAR model obtained by complete search.

Discussion

Our genetic algorithm was implemented and proved its ability to identify MDF members able to explain the relationship between the PCBs structure and the logarithm of octanol-water partition coefficients. We used the tournament method to select MDF descriptors. The ability of the GA-MLR model with four descriptors was successfully investigated and the implemented GA revealed its ability to identify the closest to optimum solution.

The genetic algorithm was analyzed in terms of both its characteristics and the GA-MLR model identified. The characterization of the implemented genetic algorithm revealed the following:

- The close to optimum solution was obtained in a shorter time (less than 0.1 seconds per generation) than the complete search for the best model (time dimension measured in days). The time needed to identify the closest to optimum solution directly depends on the number of selected generations. In the present research 50,000 generations were imposed and the optimum solution was identified in a few minutes.
- The obtained solution is close to the optimum solution (obtained by complete search). The absence of a significant difference between the correlation coefficient of the model from Eq. (1) and the GA-MLR model (Eq. (2)) supported this statement. The genetic algorithm obtained the best solution in the generation 17,277 (see Fig. 2). Twenty-one evolutions in terms of determination coefficients were observed.

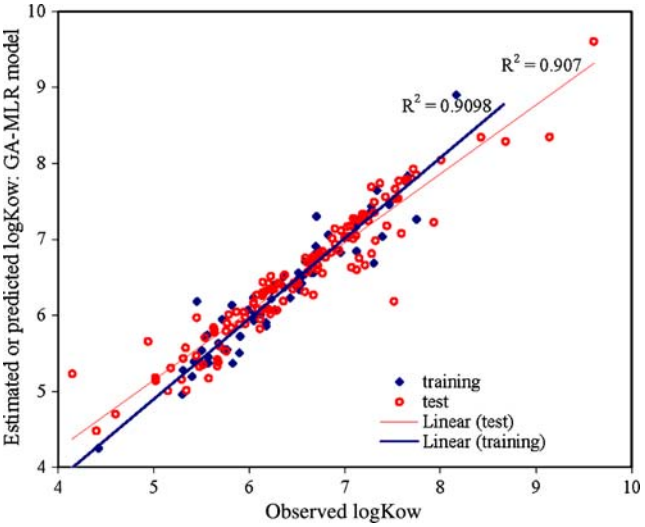

Fig. 3 GA-MLR model: training *vs.* test analysis (154 PCBs in training set)

The determination coefficient increased in comparison to the value obtained in previous generation from $9.99 \cdot 10^{-16}$ (generation 3179 compared to generation 478) to $5.22 \cdot 10^{-3}$ (generation 16,861 compared to generation 16,857) when the analysis was performed on the r^2 values with a precision of 15 decimals.

- As far as the applicability domain is concerned, the implemented GA proved its ability to identify the closest to the optimum as possible determination coefficient by searching in the sample of PCBs genotypes.

Computing efficiency of the GA algorithm could be analyzed in terms of combinations needed to be performed for identification of the closest to optimum solution. The number of investigated regressions on complete and GA search for models with 4 descriptors was of $6.46 \cdot 10^{17}$ (complete search) and $3.70 \cdot 10^{10}$ (GA algorithm) (see supplementary Appendix 1). The time needed to find the optimum (complete search) and closest to optimum model (GA algorithm) is proportional with complexity of calculation (in the above example the complexity was reduced by $1.75 \cdot 10^7$ times).

The characteristics of the GA-MLR performances were analyzed and the following were revealed:

- The number of genotypes equaled the imposed number (set in the present research at 12), with one exception (the generation when the highest determination coefficient was obtained, where the number of genotypes equaled 10).
- The number of distinct phenotypes in cultivar varied from 16 to 23 and, as expected, was directly related to the number of valid regressions.
- The correlation coefficients of the GA-MLR models in evolution were included into the 95% confidence interval of the correlation coefficient associated to the best GA-MLR model. This was expected since the difference between the minimum and maximum correlation coefficients was equal to 0.0103.
- The sum of residuals in estimate varied from 13.36 to 16.13. The minimum value was obtained when the GA-MLR model had the highest determination coefficient.
- The Hölder mean varied from 7.92 to 12.99; the maximum value was obtained when the GA-MLR model had the highest determination coefficient.
- The quantity of explained and un-explained entropy varied from 0.4511 to 0.5119; the minimum value was obtained when the GA-MLR model had the highest determination coefficient.

The analysis of the GA was also performed by analyzing the GA-MLR model. The GA-MLR model was statistically significant, each MDF descriptor contributed to the explanation of the octanol-water partition coefficient expressed in logarithmic scale (see Eq. (2)). The contribution of each descriptor from GA-MLR model was investigated by using the partial F test. Starting to the model with one descriptor, adding of a new descriptor proved to significantly improve the estimation of the model (see supplementary Table 2). In addition, adding of a new descriptor revealed to provide a model with a significantly higher correlation coefficient (see Steiger's Z test, supplementary Table 2). These results sustain the validity of the GA-MLR model. Besides the determination correlation coefficient and its adjusted value, the standard error of estimate sustained the estimation abilities of the GA-MLR model (see Eq. (2)). The analysis of the GA-MLR model showed that the octanol-water partition coefficient was of geometric (iIPRJC**g** and IiPDLC**g**) and topological (IMDRLH**t** and IhDDJC**t**) nature and depended on PCBs cardinality (iIPRJ**C**g, IiPDL**C**g, and IhDDJ**C**t) and number of hydrogen atoms (IMDRL**H**t) as atomic properties. As can be observed, when Eqs. (1) and (2) were compared, the closest to the optimum solution (Eq. (2)) preserved only the geometric and topological nature of the octanol-water partition coefficient in relation with the structure of PCBs.

The MDF descriptors identified by GA proved to be useful in characterization of relationships between octanol-water partition coefficient and structure of PCBs in principal components and classification analysis. Four factors were obtained; the first two factors were revealed to have a contribution of 98.51% (see supplementary Fig. 1).

The results obtained in leave-25%-out sustained the abilities of GA-MLR model. Both models presented in Eqs. (3) and (4) are statistically significant. All descriptors had significant contribution in explanation of the relationship between compounds structure and octanol-water partition coefficients (see t-values, Eqs. (3) and (4)). Moreover, the values of the determination coefficients obtained in training - Eq. (3) and test - Eq. (4) are comprised into the 95% confidence interval of determination coefficient obtained by GA-MLR model - Eq. (2).

The results obtained in training *vs*. test analysis (Eqs. (3) and (4), and additionally results presented in supplementary Table 3), the determination coefficient and the standard error of predicted (see Eq. (2)) supported the abilities of the model in prediction. The analysis of the GA-MLR model in training *vs*. test experiment when the number of compounds in training sets varied from 120 to 180 with an increment of 2 (supplementary Table 3) revealed the following:

- All models, in training as well as in test sets, were statistically significant (see supplementary Table 3). Moreover, the two decimal values of the intercept and the descriptors' coefficients of the models obtained in the training set belonged to the 95% confidence intervals, with one exception (the coefficient of the

IhDDJCt descriptor, 156 PCBs in the training test, which was under the lower boundary - see supplementary Table 3 and Eq. (2)).

- The correlation coefficients obtained in the training sets belonged to the 95% confidence interval of the GA-MLR model (see Eq. (2) and supplementary Table 3). The statement is also true for test sets, with three exceptions (numbers 28, 42, and 44 in test sets); in all three cases the correlation coefficient exceeded the 95% confidence interval of the GA-MLR model (see Eq. (2) and Table 3).
- The graphical representation in Fig. 3, where the training set comprises 154 PCBs, supported the estimation and prediction abilities of the GA-MLR model.

In our study, the search for the pairs of MDF descriptors able to explain the relationship between the structure of PCBs and the octanol-water partition coefficients (expressed in logarithmic scale) was translated into genetic terms and solved by developing, implementing and assessing a genetic algorithm. Although the method is not new in the SAR analysis of PCBs [55, 56, 57], we approached it differently by using genetic algorithms and implementing them on PCBs. The implemented GA identified the pairs of MDF descriptors able to characterize the relationships between the structure of the PCBs and the octanol-water partition coefficient.

Conclusions

The proposed genetic algorithm proved its abilities in terms of speed, precision and applicability domain on multivariate linear regression models on octanol-water partition coefficients of the investigated polychlorinated biphenyls.

The genetic algorithm obtained a close to optimum solution (not significantly different from the one obtained by complete search) in a very short time. Moreover, the relationship between the structure of PCBs and the octanol-water partition coefficient obtained by applying the GA proved to be of geometric and topological nature as previously identified by the complete search, proving thus the consistency of the solution proposed by the genetic algorithm.

Acknowledgments UEFISCSU Romania partially supported this research through project (ID-202/01.10.2007 & ID-206/01.10.2007).

References

1. Hansch C, Leo A (1979) Substituent constants for correlation analysis in chemistry and biology. Wiley, New York
2. Kaminski JJ (1994) Computer-assisted drug design and selection. Adv Drug Deliver Rev 14(2–3):331–337. doi:10.1016/0169-409X(94)90049-3
3. Barbosa F, Horvath D (2004) Molecular similarity and property similarity. Curr Top Med Chem 4(6):589–600. doi:10.2174/1568026043451186
4. Baumann K (1999) Uniform-length molecular descriptors for quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR): classification studies and similarity searching. Trends Analyt Chem 18(1):36–46. doi:10.1016/S0165-9936(98)00075-2
5. Bravi G, Gancia E, Mascagni P, Pegna M, Todeschini R, Zaliani A (1997) MS-WHIM, new 3D theoretical descriptors derived from molecular surface properties: A comparative 3D QSAR study in a series of steroids. J Comput Aid Mol Des 11(1):79–92. doi:10.1023/A:1008079512289
6. Ciubotariu D, Deretey E, Oprea TI, Sulea T, Simon Z, Kurunczi L, Chiriac A (2006) Multiconformational minimal steric difference. Structure-acetylcholinesterase hydrolysis rates relations for acetic acid Esters. QSAR Comb Sci 12(4):367–372. doi:10.1002/qsar.19930120404
7. Jäntschi L (2005) Molecular descriptors family on structure activity relationships 1. Review of the methodology. Leonardo Electron J Pract Technol 6:76–98 Available via: http://lejpt.academicdirect.org/A06/76_98.htm. Accessed 15 April 2009
8. Jäntschi L, Bolboacă SD (2007) Results from the use of molecular descriptors family on structure property/activity relationships. Int J Mol Sci 8(3):189–203. doi:10.3390/i8030189
9. Putz MV, Lacrămă AM (2007) Introducing Spectral Structure Activity Relationship (S-SAR) analysis. Application to ecotoxicology. Int J Mol Sci 8(5):363–469. doi:10.3390/i8050363
10. Du QS, Huang RB, Wei YT, Pang ZW, Du LQ, Chou KC (2009) Fragment-based quantitative structure-activity relationship (FB-QSAR) for fragment-based drug design. J Comput Chem 30 (2):295–304. doi:10.1002/jcc.21056
11. Du QS, Huang RB, Wei YT, Du LQ, Chou KC (2008) Multiple field three dimensional quantitative structure-activity relationship (MF-3D-QSAR). J Comput Chem 29(2):211–219. doi:10.1002/jcc.20776
12. Jing JH, Xiao SY, Li ZL (2008) Quantitative structure-activity relationship studies of fatty acids in ranunculus ternatus thunb using three-dimensional holographic vector of atomic interaction field. Fenxi Huaxue/ Chinese J Anal Chem 36(7):971–974
13. Vedani A, McMasters DR, Dobler M (2000) Multi-conformational ligand representation in 4D-QSAR: Reducing the bias associated with ligand alignment. Quant Struct-Act Relatsh 19(2):149–161. doi:10.1002/1521-3838(200004)19:2<149::AID-QSAR149>3.0.CO;2-9
14. Vedani A, Dobler M (2002) 5D-QSAR: The key for simulating induced fit? J Med Chem 45(11):2139–2149. doi:10.1021/jm011005p
15. Eriksson L, Johansson E, Lindgren F, Sjöström M, Wold S (2002) Megavariate analysis of hierarchial QSAR data. J Comput Aided Mol Des 16(10):711–726. doi:10.1023/A:1022450725545
16. Liang G, Chen G, Niu W, Li Z (2008) Factor analysis scales of generalized amino acid information as applied in predicting interactions between the human amphiphysin-1 SH3 domains and their peptide ligands. Chem Biol Drug Des 71(4):345–351. doi:10.1111/j.1747-0285.2008.00641.x
17. Tsygankova IG (2008) Variable selection in QSAR models for drug design. Curr Comput Aided Drug Des 4(2):132–142. doi:10.2174/157340908784533238
18. Khan MTH, Sylte I (2007) Predictive QSAR modeling for the successful predictions of the ADMET properties of candidate drug molecules. Curr Drug Discov Technol 4(3):141–149
19. Vighi M, Migliorati S, Monti GS (2009) Toxicity on the luminescent bacterium Vibrio fischeri (Beijerinck). I: QSAR equation for narcotics and polar narcotics. Ecotoxicol Environ Saf 72(1):154–161. doi:10.1016/j.ecoenv.2008.05.008

20. Lin W-Q, Jiang J-H, Wu H-L, Shen G-L, Yu R-Q (2006) Recent advances in chemometric methodologies for QSAR studies. Curr Comput Aided Drug Des 2(3):255–266. doi:10.2174/157340906778226418
21. Riahi S, Pourbasheer E, Dinarvand R, Ganjali MR, Norouzi P (2008) QSAR Study of 2-(1-Propylpiperidin-4-yl)-1H-Benzimidazole-4-Carboxamide as PARP inhibitors for treatment of cancer. Chem Biol Drug Des 72(6):575–584. doi:10.1111/j.1747-0285.2008.00739.x
22. Duchowicz PR, Castro EA (2008) Partial order theory applied to QSPR-QSAR studies. Comb Chem High Throughput Screen 11 (10):783–793. doi:10.2174/138620708786734316
23. Xiao Y-D, Harris R, Bayram E, Santago P II, Schmitt JD (2006) Supervised self-organizing maps in drug discovery. 2. Improvements in descriptor selection and model validation. J Chem Inf Model 46(1):137–144. doi:10.1021/ci0500841
24. Fox T, Kriegl JM (2006) Machine learning techniques for in silico modeling of drug metabolism. Curr Top Med Chem 6(15):1579–1591. doi:10.2174/156802606778108915
25. Du H, Wang J, Hu Z, Yao X, Zhang X (2008) Prediction of fungicidal activities of rice blast disease based on least-squares support vector machines and project pursuit regression. J Agric Food Chem 56(22):10785–10792. doi:10.1021/jf8022194
26. George CJ, Bennett GF, Simoneaux D, George WJ (1988) Polychlorinated biphenyls a toxicological review. J Hazard Mater 18(2):113–144. doi:10.1016/0304-3894(88)85018-0
27. Hansen BG, Paya-Perez AB, Rahman M, Larsen BR (1999) QSARs for K(ow) and K(oc) of PCB congeners: A critical examination of data, assumptions and statistical approaches. Chemosphere 39(13):2209–2228. doi:10.1016/S0045-6535(99)00145-9
28. Giri S, Roy DR, Van Damme S, Bultinck P, Subramanian V, Chattaraj PK (2008) An atom counting QSPR protocol. QSAR Comb Sci 27(2):208–230. doi:10.1002/qsar.200730109
29. Ivanciuc T, Ivanciuc O, Klein DJ (2006) Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quatitative super-structure/activity relationships (QSSAR). Mol Divers 10:133–145. doi:10.1007/s11030-005-9003-3
30. Jiang GX, Niu JF, Zhang SP, Zhang ZY, Xie B (2008) Prediction of biodegradation rate constants of hydroxylated polychlorinated biphenyls by fungal laccases from Trametes versicolor and Pleurotus ostreatus. Bull Environ Contam Toxicol 81(1):1–6. doi:10.1007/s00128-008-9433-6
31. Zeng X, Wang Z, Ge Z, Liu H (2007) Quantitative structure-property relationships for predicting subcooled liquid vapor pressure (PL) of 209 polychlorinated diphenyl ethers (PCDEs) by DFT and the position of Cl substitution (PCS) methods. Atmos Environ 41(17):3590–3603. doi:10.1016/j.atmosenv.2006.12.039
32. Jäntschi L, Bolboacă SD, Diudea MV (2007) Chromatographic retention times of polychlorinated biphenyls: From structural information to property characterization. Int J Mol Sci 8 (11):1125–1157. doi:0.3390/i8111125
33. Wei B, Xie S, Yu M, Wu L (2007) QSPR-based prediction of gas/particle partitioning of polychlorinated biphenyls in the atmosphere. Chemosphere 66(10):1807–1820. doi:10.1016/j.chemosphere.2006.09.029
34. Borja J, Taleon DM, Auresenia J, Gallardo S (2005) Polychlorinated biphenyls and their biodegradation. Process Biochem 40 (6):1999–2013. doi:10.1016/j.procbio.2004.08.006
35. Hertz-Picciotto I, Charles MJ, James RA, Keller JA, Willman E, Teplin S (2005) In utero polychlorinated biphenyl exposures in relation to fetal and early childhood growth. Epidemiology 16 (5):648–656. doi:10.1097/01.ede.0000173043.85834.f3
36. Bodin N, Le Loc'h F, Caisey X, Le Guellec A-M, Abarnou A, Loizeau V, Latrouite D (2008) Congener-specific accumulation and trophic transfer of polychlorinated biphenyls in spider crab food webs revealed by stable isotope analysis. Environ Pollut 151 (1):252–261. doi:10.1016/j.envpol.2007.01.051
37. Ruiz P, Faroon O, Moudgal CJ, Hansen H, De Rosa CT, Mumtaz M (2008) Prediction of the health effects of polychlorinated biphenyls (PCBs) and their metabolites using quantitative structure-activity relationship (QSAR). Toxicol Lett 181(1):53–65. doi:10.1016/j.toxlet.2008.06.870
38. Jäntschi L, Bolboacă SD (2006) Molecular Descriptors Family on Structure Activity Relationships 6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls. Leonardo Electron J Pract Technol 8:71–86 Available via: http://lejpt.utcluj.ro/A08/71_86.htm. Accessed 15 April 2009
39. Jäntschi L, Bolboacă SD (2007) Integrated Structural Investigations on Biological Active Compounds (Research Report; in Romanian). Available via: http://lori.academicdirect.org/research/grants/Raport_Cercetare_ET036_2007.pdf. Accessed 15 April 2009
40. Connor MS (1985) Comment on „fish/sediment concentration ratios for organic compounds". Environ Sci Technol 19(2):198–199. doi:10.1021/es00132a015
41. Eisler R, Belisle AA (1996) Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review. Contaminant Hazard Reviews 1–96. Available via: http://www.pwrc.usgs.gov/infobase/eisler/chr_31_planar_pcbs.pdf. Accessed 18 April 2009
42. Hoffmann R (1963) An extended Hückel theory. I. Hydrocarbons. J Chem Phys 39(6):1397–1412. doi:10.1063/1.1734456
43. Dewar MJS, Zoebisch EG, Healy EF, Stewart JJP (1985) Development and use of quantum mechanical molecular models. 76. AM1: a new general purpose quantum mechanical molecular model. J Am Chem Soc 107(13):3902–3909. doi:10.1021/ja00299a024
44. Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans R Soc Edin 52:399–433
45. Weismann A (1893) The germ-plasm: a theory of heredity. C. Scribner's Sons, New York
46. de Veies H (1902) The origin of species by mutation. Science 15 (384):721–729. doi:10.1126/science.15.384.721
47. Auerbach C, Robson JM, Carr JG (1947) The chemical production of mutations. Science 105(2723):243–247. doi:10.1126/science.105.2723.243
48. Cairns J, Overbaugh J, Miller S (1988) The origin of mutants. Nature 335(6186):142–145. doi:10.1038/335142a0
49. Darwin CR (1859) On the origin of species by means of natural selection. J Murray, London
50. Jarque CM, Bera AK (1980) Efficient tests for normality, homoscedasticity and serial independence of regression residuals. Econ Lett 6(3):255–259. doi:10.1016/0165-1765(80)90024-5
51. Kleinbaum DG, Kupper LL, Muller KE, Nizam A (2008) Applied regression analysis and multivariable methods, 4th edn. Duxbury (Thomson Higher Educatio), Canada, pp 141–146
52. Steiger JH (1980) Tests for comparing elements of a correlation matrix. Psychol Bull 87:245–251
53. Bolboacă SD, Jäntschi L (2008) Modelling the property of compounds from structure: statistical methods for models validation. Environ Chem Lett 6:175–181. doi:10.1007/s10311-007-0119-9
54. Daren Z (2001) QSPR studies of PCBs by the combination of genetic algorithms and PLS analysis. Comput Chem 25(2):197–204. doi:10.1016/S0097-8485(00)00081-4
55. Todeschini R, Consonni V, Mauri A, Pavan M (2004) Detecting "bad" regression models: Multicriteria fitness functions in regression analysis. Anal Chim Acta 515(1):199–208. doi:10.1016/j.aca.2003.12.010
56. Pavan M, Mauri A, Todeschini R (2004) Total ranking models by the genetic algorithm variable subset selection (GA-VSS) approach for environmental priority settings. Anal Bioanal Chem 380:430–444. doi:10.1007/s00216-004-2762-3