Bulletin UASVM Horticulture, 67(1)/2010
Print ISSN 1843-5254; Electronic ISSN 1843-5394

Average Trends over Millennia of Evolution Supervised by Genetic Algorithms. 1. Analysis of Genotypes

Lorentz JANTSCHI[1)], Sorana D. BOLBOACA[2)], Mircea V. DIUDEA[3)], Radu E. SESTRAS[4)]

[1)] Technical University of Cluj-Napoca, Department of Chemistry, 103-105 Muncii Bvd., 400641, Cluj-Napoca, Romania; lori@academicdirect.org
[2)] "Iuliu Hatieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur Street, 400349, Cluj-Napoca, Romania; sbolboaca@umfcluj.ro
[3)] Babes Bolyai University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos Street, 400028, Cluj-Napoca, Romania; diudea@chem.ubbcluj.ro
[4)] University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Manastur Street, 400372 Cluj-Napoca, Romania; rsestras@usamvcluj.ro

Abstract. A genetic algorithm (GA) had been developed and implemented in order to identify the optimal solution in term of determination coefficient and estimation power of a multiple linear regression approach of structure-activity relationships. The Molecular Descriptors Family for structure characterization of a sample of 206 polychlorinated biphenyls with measured octanol-water partition coefficients was used as case study. The research aimed to analyze the degree of association between number of viable genotypes in cultivar in generations in which the evolution occurred and the selection and survival strategies. The GA was repeated for 46 times and the Anderson-Darling test was used to compared the distribution laws of populations occurred by using different pairs of survival and selection strategies. The records of genotypes number were analyzed and the conclusions were highlighted.

Keywords: Anderson-Darling Test, genetic algorithm (GA), distribution; genotypes

INTRODUCTION

The quantitative structure-property/activity relationships (qSAR/qSPR) approaches are widely used in characterization of different biological active compound due to their advantages (Boyer, 2009; Ghose *et al.*, 2006). Polychlorinated biphenyls (PCBs), organic compounds with 1 to 10 chlorine atoms attached to biphenyl, were widely used for different applications (e.g. dielectric fluids). Due to their toxicity (El-Shahawi *et al.*, 2010) and their persistency in environment (Werner *et al.*, 2010) the production of PVBs was banned in 1979 in US and in 2001 in Europe.

A previous research on PCBs using the Molecular Descriptors Family (MDF) (Jäntschi and Bolboacă, 2006) identified a multiple linear regression with 4 variables able to explained 91% of the measured activity:

$$\hat{Y} = 3.04 - 0.42 \cdot IIDDKGg + 0.04 \cdot IHDRKEg + 0.07 \cdot aHMmjQt - 37.5 \cdot aSMMjQg$$

$$r^2(Y,\hat{Y}) \approx 0.91, F=554$$

The above equation did not respect all criteria imposed by the phenotypic variability (e.g. variability, deviation from normality and reasonable determination). The phenotypic variability criteria were implemented using a genetic algorithm approach (note that the *HMmjQt* and *SMMjQg* descriptors from the above equation did not accomplished the viability tests) through implementation of a heuristic able to find association of pairs of descriptors (Jäntschi, 2010a). The aim was to identify an optimal qSPR able to characterize the biological activity using a multiple

linear regression (MLR) with four MDF members as variables considering as golden standard the equation previously presented and was successfully accomplished (Jäntschi, 2010a). The implemented heuristic was set to be able to search pairs of descriptors; this strategy was implemented due to the complexity of the calculus (the complexity is proportional to twice of power 2 of sample size compared to power of the sample size in searching the association of each descriptor to each descriptor).

Our present research was focus on finding the answer to the following question "Is the average number of viable genotypes in cultivar dependent or not by the selection and survival strategies?"

MATERIALS AND METHODS

An analysis of structure-property relationships was carried out using a genetic algorithm approach (Jäntschi, 2010a). A sample of 206 PCBs with known octanol-water partition coefficient (Eisler and Belisle, 1996) expressed in logarithmic scale ($\ln(K_{OW})$) was included in the analysis. The geometry of the PCBs was optimized using the HyperChem 8.0 software. The AMBER molecular mechanics method was applied; the Polak-Ribiere was used for geometry optimization (Polak and Ribiere, 1969) while AM1 semi empirical method was applied for energy calculation (Dewar *et al.*, 1985). The optimization results and the calculated energies were saved as *.hin files, defining the chemical structure of each PCB included into analysis. These files represented the input data for SPR analysis. The molecular descriptors family (MDF) approach (Jäntschi, 2004) was used to link the structure of the PCBs and octanol-water partition coefficient. Detail of the genetic algorithm could be found in (Jäntschi *et al.*, 2010b; 2010c). A series of parameters has been counting during the evolution of the heuristics; the following two were of interest for the present research: number of generation in which the evolution occurred and number of viable genotypes occurred in these generation.

A program to analyze the distribution of the pair sampling using the Anderson-Darling test (Anderson and Darling, 1952) was develop and implemented in order to found answer to the research question. Anderson-Darling test verify if there is a statistical evidence that a sample is from a given probability distribution function. Thus, the test verify if the hypothesis that the ascending ordered observations ($(X_i)_{1\le i\le n}$, $X_i<X_{i+1}$) are from a cumulative distribution probability function. The hypothesis is rejected by calculating the value given by the relation:

$$A^2 = -n - \sum_{k=1}^{n} \frac{2k-1}{n}\left(\ln(F(Y_k)) + \ln(1-F(Y_{n+1-k}))\right)$$

However, an application of interest is the Anderson-Darling test for several samples of which the provenance from the same population can be verified without specification of the population distribution law (Scholz and Stephens, 1987; ***, 2002). The formulas used for comparing samples and its interpretation are presented in (Jäntschi, 2009).

When a discrete known distribution is compared to an observed distribution, the variance of A^2 is computed using the formula (Scholz and Stephens, 1987):

$$\mathrm{Var}(A^2) = \frac{2(\pi^2-9)}{3} + \frac{10-\pi^2}{n}$$

where n = number of observations in the sample, π=3.14159265358979323846264338...

RESULTS AND DISCUSSION

The total number of 502 (2^9-C_9^0-C_9^1) possible inferences where investigated. The distribution of the average number of viable genotypes in cultivar according to selection and survival strategies is shown in Table 1.

Tab. 1

Distribution of the average number of viable genotypes in cultivar according to selection and survival strategies

Obs\Group	PP	PT	PD	TP	TT	TD	DP	DT	DD
1	11.43	11.51	11.80	11.41	11.45	11.85	11.23	11.33	11.83
2	11.21	11.46	11.69	10.95	11.17	11.66	10.50	11.12	11.68
3	11.06	11.32	11.55	10.93	11.14	11.68	10.23	11.06	11.72
4	11.29	11.38	11.65	11.27	11.29	11.60	10.58	10.94	11.70
5	11.58	11.28	11.70	11.21	11.16	11.54	10.86	11.10	11.83
6	11.36	11.48	11.68	10.68	11.46	11.65	10.32	11.28	11.85
7	11.40	11.38	11.38	11.08	11.24	11.61	10.00	10.82	11.79
8	11.30	11.61	11.44	11.00	11.36	11.62	10.49	10.60	11.77
9	11.14	11.34	11.64	10.94	11.29	11.62	11.10	10.69	11.69
10	11.19	11.53	11.57	11.11	10.78	11.74	11.13	10.88	11.79
11	11.38	11.55	11.68	11.12	11.47	11.48	11.07	10.93	11.82
12	11.63	11.18	11.64	11.39	10.76	11.64	10.77	10.92	11.84
13	11.05	11.31	11.58	11.24	11.36	11.74	9.67	10.76	11.78
14	11.07	10.73	11.68	11.11	11.47	11.69	9.59	10.69	11.78
15	11.50	11.10	11.60	11.29	11.06	11.80	9.81	11.50	11.74
16	12.00	11.38	11.40	11.50	11.12	12.00	11.00	10.56	11.86
17	11.17	11.50	11.73	11.00	11.33	11.79	10.00	10.88	11.68
18	11.67	11.25	11.64	11.17	12.00	11.67	11.18	11.10	11.92
19	11.82	11.64	11.57	11.50	11.71	11.73	11.33	10.82	11.88
20	11.75	10.73	11.50	11.31	11.33	11.75	11.40	10.00	12.00

Obs = average over 1000 generations;
Group: pairs of selection and survival strategies
P = proportional; T = tournament; D = deterministic

The following were obtained by analyzing the results:

- Analysis of pairs of selection-survival strategies shown that the hypothesis of belonging to the identical population could not be rejected for: DT and DP with a ratio between critical value and obtained value of c/k = 1.2; PP and PT (c/k = 3.0); PP and TT (c/k = 3.2); TT and PT (c/k = 2.1) and TT & TP (c/k = 1.3).
- The analysis of groups of three survival-selection pairs shown statistically that:
  - The hypothesis that PP, PT and TT are from identical populations could not be rejected (c/k = 2.2).
  - With a 5% risk to be in error, all groups containing the TP method are from different populations.
- The analysis of groups of four survival-selection pairs shown statistically that:
  - With a 5% risk to be in error, in the PP-PT-TP-TT group of the pairs of method at least one is from different population (c/k = 0.9).

The main conclusions could be summarizing as:

- The analysis of selection-survival pairs of strategies:
  - With a 5% risk of error the following are from different populations: DD, PD and TD.
  - The hypothesis that DT and DP are from identical populations could not be rejected at 5% significance level.
  - The PP, PT, TT and TP pairs need the investigation of higher order groups.

- The analysis of groups of three survival-selection pairs of strategies:
  - With a 5% risk to be in error, the following are from different populations: DD, PD and TD.
  - The hypothesis that DT and DP are from the same population could not be rejected.
  - The hypothesis that TT and TP are from the same population could not be rejected.
  - The hypothesis that PP, PT and TT are from identical populations could not be rejected.
- The analysis of groups of four survival-selection pairs of strategies:
  - With a 5% risk to be in error, the DD, PD and TD are from different populations.
  - The hypothesis that DT and DP are from the same population could not be rejected.
  - The hypothesis that TT and TP are from the identical populations could not be rejected.
  - The hypothesis that PP, PT and TT are from the identical populations could not be rejected.
  - With a 5% risk to be in error, the hypothesis that PP-PT-TP-TT is from the identical populations is rejected.

The analysis of results allows partitioning of the selection and survival strategies according to the population of the average number of viable genotypes in cultivar into generation when the evolution occurred. The summary of this analysis is presented in Table 2. The Figure 1 presented the average number of viable genotypes in cultivar in generation when evolution occurred for each selection and survival studied strategies.

Tab. 2

Populations for the average number of the viable genotypes in cultivar in the generations producing evolution during evolving depending on the selection and survival strategies

Sel\Srv	Proportional	Tournament	Deterministic
Proportional	PP	PT	PD
Tournament	TP	TT	TD
Deterministic	DP	DT	DD

Sel: Selection (rows) vs. Srv: Survival (columns)

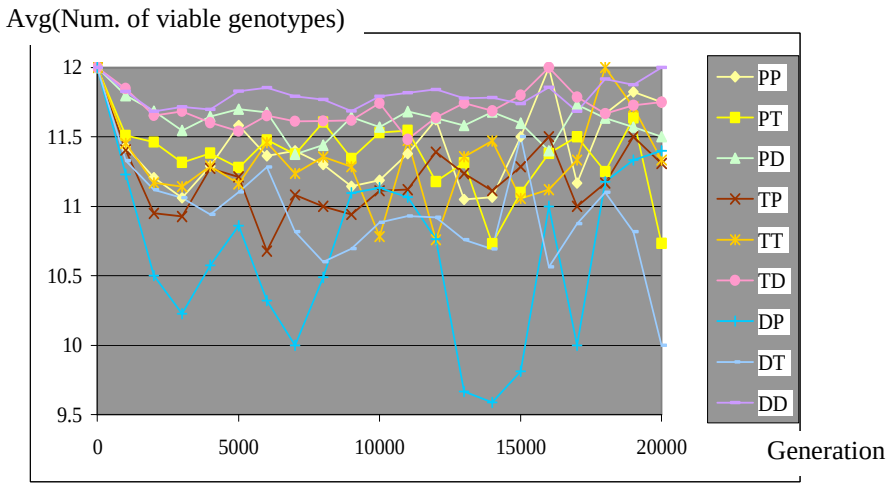

Fig. 1. Average number of viable genotypes according to selection (first letter) and survival (second letter)

The analysis of Figure 1 revealed that when deterministic selection is accompanied by the deterministic survival (DD) ,the average number of viable genotypes in cultivar is high. Note that its population distribution law proved to be statistically different compared to all other investigated selection and survival strategies.

Deterministic selection strategy accompanied by the tournament survival strategy (DT) or proportional survival strategy (DP) produced a population with a distribution statistically different at a significance level of 5% compared to all other investigated selection and survival strategies.

Deterministic survival strategy accompanied by the tournament selection (TD) produce a population distribution statistically different at a significance level of 5% compared to the populations produced by all other investigated survival-selection strategies (Tab. 2 and Fig. 1). Moreover, the mean and the variability of TD are close to the values obtained by DD and DP (Fig. 1).

Deterministic survival strategy accompanied by the proportional selection (PD) produced a population with a distribution statistically different at a significance level of 5% compared to the populations produced by all other investigated survival-selection possibilities (Tab. 2 and Fig. 1). The characteristic of PD in terms of mean and variability are close to the values obtained in deterministic selection accompanied by the tournament survival (DT, Fig. 1).

Selection and survival that are neither deterministic form a relatively compact group (Fig. 1) and produced relatively close populations (Tab. 2) in terms of average number of viable genotypes in cultivar during evolution.

The Anderson-Darling statistics applied on the obtained experimental data was conducted in order to identify statistically significant differences of distribution laws of obtained. Based on the Starting from the obtained results it is possible to separate in statistical terms the populations (see Figure 2, mean values on intervals of thousand of generations).

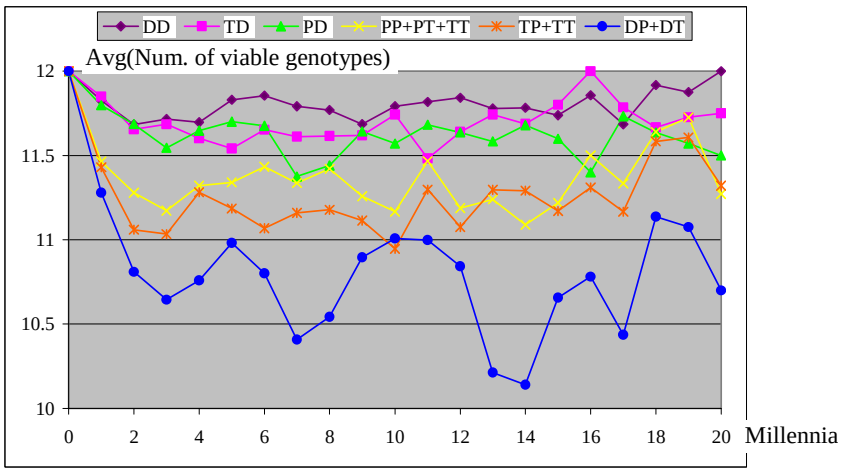

Fig. 2. Average of the number of viable genotypes: significantly different populations partitioned by selection and survival

The average trends of evolution presented in Figure 2 can be obtained by calculating for each frequency class (from 0 to 20000) the mean value of observed variable (the average number of viable genotypes – the mean of means) from generation 0 to the time evolution of the frequency class. The average trends presented in Figure 2 correspond to the values obtained through a Monte-Carlo simulation (Fig. 3). The Monte-Carlo is a class of algorithms related to the genetic algorithms discovered in 1947 (Metropolis and Ulam, 1949) and mainly used in computational statistics.

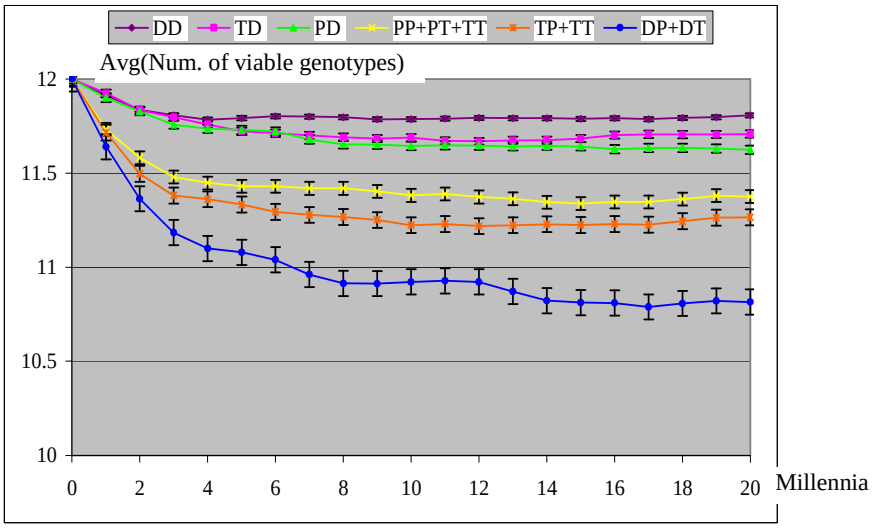

Fig. 3. Average tendencies for the average of the number of viable genotypes: statistically significant different populations partitioned by selection and survival

The analysis of Figure 3 (where were presented also the standard errors) revealed that the differences in average tendencies for the populations produced by DD, TD and PD are very small. Moreover, the intervals of standard errors are frequently overlap on each other for the pair (TD, PD) since the populations produced by DD, TD and PD are distinct on each other as the Anderson-Darling statistics proved.

The proposed solution presented in Table 2 can be further refined to obtain the solution for the most probable combination that is consistent to the obtained possible solutions. If the 95% confidence interval is not considered sufficient for the obtained associations and proceed to successive elimination until the most unlikely combinations are removed (until a partition into the association of the methods) the removal will be stop after elimination of the DP and DT (with c / k = 1.2 and c / k = 1.3) when the partition of the method is obtained {DD, TD, PD, DT, DP, (PP, PT, TT)} (this have a single group where three pairs of selection survival strategies lies - PP, PT and TT). Figure 4 showed the evolution of the average number of viable genotypes in cultivar while the Figure 5 presented the mean of means of average number of genotypes in cultivar.

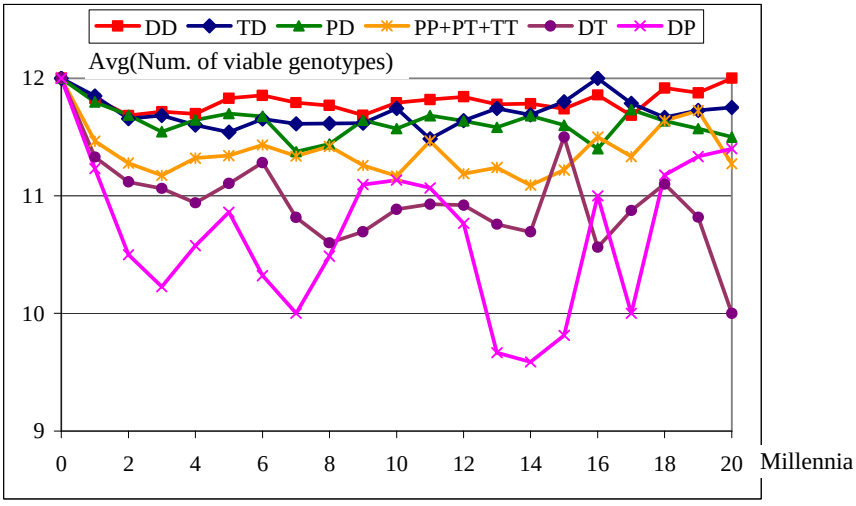

Fig. 4. Partition of the number of genotypes represented in cultivar: the most probable solution

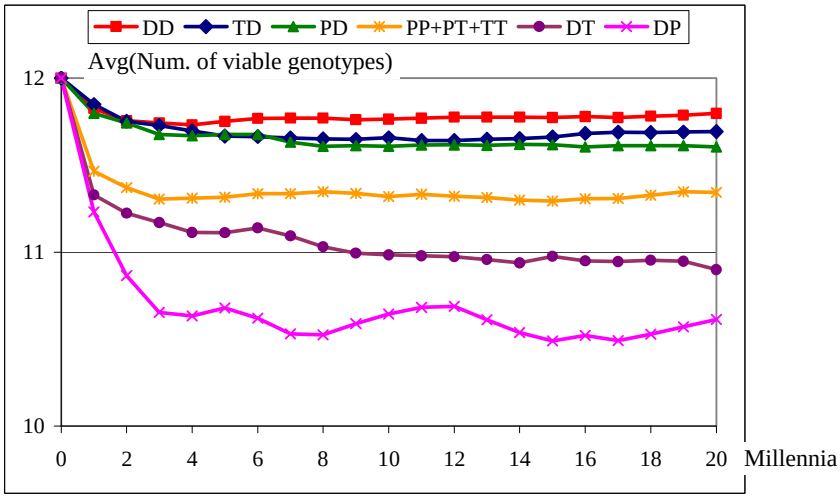

Fig. 5. Partition of the number of genotypes represented in cultivar: the Monte-Carlo experiment

The aim of our research was successfully accomplished. The average numbers of viable genotypes in cultivar where investigated in order to identify their dependence or independence by the selection and survival strategies and this research allowed us to draw important conclusions.

CONCLUSIONS

The deterministic selection accompanied by deterministic survival, proportional selection accompanied by deterministic survival and tournament selection accompanied by deterministic survival proved not belonging to the same population for a significance level of 5%.

The hypothesis that the following pairs of selection strategy and survival strategy were from the same population could not be rejected at a significance level of 5%: DT (deterministic selection accompanied by tournament survival) and DP (deterministic selection accompanied by proportional survival); TT (tournament selection accompanied by tournament survival) and TP (tournament selection accompanied by proportional survival); PP (proportional selection accompanied by proportional survival), PT (proportional selection accompanied by tournament survival) and TT (tournament selection accompanied by tournament survival).

Deterministic selection strategy accompanied by deterministic survival strategy has as effect maintaining a high number of viable genotypes in cultivar; its population distribution law proved to be statistically different compared to all other investigated selection and survival strategies.

REFERENCES

1. Anderson, T. W. and D. A. Darling (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. Annals of Mathematical Statistics 23(2):193-212.

2. Boyer S. (2009). The use of computer models in pharmaceutical safety evaluation. Altern. Lab. Anim. 37(5):467-75.

3. Dewar, M. J. S., E. G. Zoebisch, E. F. Healy and J. J. P. Stewart (1985). AM1: A New General Purpose Quantum Mechanical Molecular Model. J. Am. Chem. Soc. 107:3902-3909.

4. Eisler, R. and A. A. Belisle (1996). Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A

Synoptic Review. Contaminant Hazard Reviews. Biological Report 31. [online] [Accessed march 2009] Available from: URL: http://www.pwrc.usgs.gov/infobase/eisler/chr_31_planar_pcbs.pdf

5. El-Shahawi, M. S., A. Hamza, A. S. Bashammakh and W. T. Al-Saggaf (2010). An overview on the accumulation, distribution, transformations, toxicity and analytical methods for the monitoring of persistent organic pollutants. Talanta 80(5):1587-97.

6. Ghose, A. K., T. Herbertz, J. M. Salvino and J. P. Mallamo (2006). Knowledge-based chemoinformatic approaches to drug discovery. Drug Discov. Today 11(23-24):1107-14.

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

8. Jäntschi, L. and S. D. Bolboacă (2006). Molecular Descriptors Family on Structure Activity Relationships 6. Octanol-Water Partition Coefficient of Polychlorinated Biphenyls, Leonardo Electronic Journal of Practices and Technologies 5(8):71-86.

9. Jäntschi, L. (2009). http://l.academicdirect.org/Statistics/tests/kAD/, k-sample Anderson-Darling.

10. Jäntschi, L. (2010a). Genetic algorithms and their applications. PhD Thesis (Horticulture) - Supervisor Prof. Sestraş R. E., University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, Cluj, RO. http://l.academicdirect.org/Horticulture/GAs/Refs/Jäntschi&Sestras_2010_Thesis.pdf

11. Jäntschi, L., S. D. Bolboacă and R. E. Sestraş (2010b). A study of genetic algorithm evolution on the lipophilicity of polychlorinated biphenyls. Chem. Biodivers. 7(8):1978-89. http://dx.doi.org/10.1002/cbdv.200900356

12. Jäntschi, L., S. D. Bolboacă and R. E. Sestraş (2010c). Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls. J. Mol. Model. 16(2):377-86. http://dx.doi.org/10.1007/s00894-009-0540-z

13. Metropolis, N. and S. Ulam (1949). The Monte Carlo Method. J. Am. Stat. Assoc. 44(247):335-341.

14. Polak, B. and G. Ribiere (1969). Note surla convergence des méthodes de directions conjuguées. Rev. Fr. Imform. Rech. Oper. 16:35-43.

15. Scholz, F. W. and M. A. Stephens (1987). K-sample Anderson-Darling Tests. J. Am. Stat. Assoc. 82(399):918-924.

16. Werner, D., S. E. Hale, U. Ghosh and R. G. Luthy (2010). Polychlorinated biphenyl sorption and availability in field-contaminated sediments. Environ. Sci. Technol. 44(8):2809-15.

17. *** Department of Defense Handbook. (2002). Composite Materials Handbook. Volume 1. Polymer Matrix Composites Guidelines for Characterization of Structural Materials. Chapter 8. Statistical Methods. 8.3.2.2 The k-sample Anderson-Darling test MIL-HDBK-17-1F:8-17.