Bulletin UASVM Agriculture, 67(1)/2010
Print ISSN 1843-5246; Electronic ISSN 1843-5386

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

Lorentz JÄNTSCHI[1)], Sorana D. BOLBOACĂ[2)], Mircea V. DIUDEA[3)], Radu E. SESTRAŞ[4)]

[1] Technical University of Cluj-Napoca, Department of Chemistry, 103-105 Muncii Bvd., 400641 Cluj-Napoca, Romania; lori@academicdirect.org
[2] "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, Department of Medical Informatics and Biostatistics, 6 Louis Pasteur, 400349 Cluj-Napoca, Romania; sbolboaca@umfcluj.ro
[3] Babeş Bolyai University, Faculty of Chemistry and Chemical Engineering, Arany Janos no. 11, 400028 Cluj-Napoca, Romania; diudea@chem.ubbcluj.ro
[4] University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Mănăştur, 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 on structure-activity relationships. The Molecular Descriptors Family was used for structure characterization of a sample of 206 polychlorinated biphenyls with measured octanol-water partition coefficients as case study. The research aimed to analyze the degree of association between the number of viable genotypes in cultivar in generations when the evolution occurred and the selection and survival strategies. The GA was run by 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; Phenotypes.

INTRODUCTION

Genetic algorithms (GAs) are search and optimization methods inspired from process of natural evolution (Fraser, 1957a; 1957b). Besides their usefulness in finding solutions of hard problems of decision (Juan *et al.*, 2010), classification (Duval and Hao, 2009), optimization (Di Serafino et al., 2010), and simulation (Stamatakis and Zygourakis, 2010), the GAs are also used in optimization quantitative structure-activity relationships – qSAR - (Ghosh and Bagchi, 2009) as well as in drug discovery (Teixidó *et al.*, 2005).

The main researches in using the genetic algorithm for qSAR/qSPR (quantitative Structure-Property Relationships) investigations deal with variable selection (Wu *et al.*, 2010), optimization (Jäntschi *et al.*, 2010a), model identification (Dashtbozorgi and Golmohammadi, 2010), identification of best design of experiment (Yang *et al.*, 2010), in silico screening (Hemmateenejad *et al.*, 2010), etc. Despite of the interest in using the GAs in qSAR/qSPR, there are just few articles which present the assessment of the GA in terms of viable genotypes and phenotypes in cultivar. Our present research was focus on finding the answer to the question "Is the average number of viable phenotypes in cultivar dependent or not by the selection and survival strategies?"

MATERIAL AND METHODS

An analysis of structure-activity relationships was carried out on a sample of 206 PCBs (with octanol-water partition coefficients (Eisler and Belisle, 1996) expressed in logarithmic scale ($\ln(K_{OW})$) as property of interest) using a home-made genetic algorithm (Jäntschi, 2010b; Jäntschi *et al.*, 2010c; Jäntschi *et al.*, 2010d). The molecular descriptors family (MDF) approach (Jäntschi, 2004) was used to link the structure of the PCBs to the octanol-water partition coefficients. The GA experiment was repeated by 46 times and a series of parameters has been counting during the evolution of the heuristics; the number of viable phenotypes (*sum_obs*), average of viable phenotypes (*avg_obs*) according to GA evolution were of interest for the present research.

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 answer the research question. Anderson-Darling test verify if there is a statistical evidence that a sample is from a given probability function. Details about the formulas applied are detailed in (Jäntschi *et al.*, 2010e).

RESULTS AND DISCUSSION

The summary of the number of viable phenotypes expressed as means on thousand of generations (from 0..1000 to 1901..20000) according to each pair of selection and survival strategy is presented in Tab. 1.

Tab. 1

Frequency of viable phenotypes in cultivar (cumulated from 46 experiments) when evolution occurred for each pair of selection and survival strategies

SelSrv	1000G	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
PP	avg_obs	17.9	17.9	17.3	17.8	17.9	18.2	19.2	17.9	17.4	17.9	19.6	18.1	18.0	17.5	17.9	18.0	19.1	19.0	19.4	18.9
	sum_obs	13117	2529	1677	1105	643	799	479	537	610	573	567	289	360	262	143	126	229	400	330	377
PT	avg_obs	18.2	18.5	18.5	19.8	20.7	19.3	19.0	19.3	17.2	18.0	18.8	18.1	18.5	15.9	17.7	18.3	16.5	17.5	16.0	14.6
	sum_obs	12968	3157	2464	1766	1118	851	758	636	551	576	207	508	592	414	177	385	99	140	176	219
PD	avg_obs	18.1	17.7	18.4	17.7	17.4	18.5	17.2	18.4	18.1	18.1	17.7	17.3	18.5	18.0	16.0	17.4	18.9	17.1	17.5	15.8
	sum_obs	13541	2809	1862	1502	1046	1315	413	461	507	381	389	570	445	449	160	261	284	377	245	285
TP	avg_obs	18.0	18.9	18.9	19.2	19.4	19.2	18.0	18.6	19.8	19.0	19.6	19.0	18.6	20.4	18.6	19.6	17.9	18.1	18.5	19.4
	sum_obs	11766	3004	2077	1692	1106	594	666	614	674	514	489	436	317	184	130	157	161	217	185	252
TT	avg_obs	18.5	18.3	17.8	17.9	18.2	18.1	19.1	18.9	19.3	18.6	18.8	17.7	18.1	19.0	18.9	18.2	18.1	19.0	19.9	17.4
	sum_obs	13681	3254	1640	1110	1021	671	726	587	406	427	358	443	254	323	340	456	163	19	139	209
TD	avg_obs	18.1	18.5	18.0	18.3	18.5	18.2	18.0	18.0	18.2	18.6	19.1	20.0	17.9	17.8	20.4	19.0	17.8	18.3	17.7	17.1
	sum_obs	13715	2037	1762	1522	1131	1200	883	703	382	576	553	500	556	284	102	76	249	165	195	274
DP	avg_obs	16.2	15.1	14.6	15.7	15.7	14.6	13.9	14.3	14.7	15.4	16.7	16.2	14.5	12.6	14.6	16.4	14.2	15.5	15.8	17.8
	sum_obs	6817	1418	642	518	676	409	209	613	457	231	251	276	174	215	233	148	213	263	95	89
DT	avg_obs	16.3	15.5	15.3	15.3	15.5	15.6	15.4	14.3	14.9	15.5	13.8	14.0	15.2	15.4	16.4	14.7	15.1	15.6	15.2	12.5
	sum_obs	7044	1704	996	780	745	609	585	643	538	665	193	351	441	200	131	235	121	156	167	25
DD	avg_obs	17.8	18.2	19.1	18.1	18.3	17.3	17.7	18.4	19.9	18.6	17.7	19.2	19.6	17.8	19.0	17.9	17.4	19.5	17.6	16.5
	sum_obs	8284	2189	1547	1195	969	711	424	717	1015	446	195	364	352	409	436	250	330	234	423	33

Sel = selection strategy; Srv = survival strategy; avg_obs = mean of observations; sum_obs = sum of observations; 1000G = thousands of generations; P = proportional; T = tournament; D = deterministic

Anderson-Darling test has been applied in order to accomplish the aim of the research by investigation of a total number of 502 cases. The smallest value of Anderson-Darling statistics of 0.6096 was observed when the following pairs of selection and survival strategies were analyzed: PT (proportional selection accompanied by tournament survival) and DD (deterministic selection accompanied by deterministic survival). The highest value of Anderson-Darling statistics of 15.6013 was observed when the following pairs of selection and survival strategies were analyzed:

PP (proportional selection accompanied by proportional survival) and DT (deterministic selection accompanied by tournament survival). In 7% of investigated cases the hypothesis that the groups are from identical population could not be rejected at a significance level of 5%. In 93% of cases, with a 5% risk to be in error, it could be stated that the investigated groups of selection and survival strategies are from different populations. Moreover, the population interpretation of the average number of viable phenotypes in cultivar in generation when evolution occurred is more complex compared to the population interpretation of the number of genotypes.

The interpretation of the results requires the identification of the largest group of pairs of methods with possible identical populations (hereafter named *list of suspects* – suspected to originate from the population with identical distributions) and elimination of their unique sub-groups of inferior order. Thus, the groups of higher order are group of order 5 (PP, PT, TT, TD, DD) - run 400 - and (PP, PT, PD, TD, DD) - run 448 - that automatically enter in the list of suspects; the inferior order groups are as it results from application of the inclusion algorithm:

÷ Groups of 5^{th} order; list of suspects: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD)}; ■ (PP, PT, TD, DD) - run 384 - is simultaneously in (PP, PT, TT, TD, DD) and (PP, PT, PD, TD, DD) making it impossible to detect its membership, is it added to the list of suspects; ■ (PT, TT, TD, DD) - run 145 - is just in (PP, PT, TT, TD, DD); is deleted; ■ (PP, TT, TD, DD) - run 272 - is just in (PP, PT, TT, TD, DD); is deleted; ■ (PP, PT, TT, DD) - run 392 - is just in (PP, PT, TT, TD, DD); is deleted; ■ (PP, PT, TT, TD) - run 399 - is just in (PP, PT, TT, TD, DD); is deleted; ■ (PP, PT, PD, DD) - run 440 - is just in (PP, PT, PD, TD, DD); is deleted; ■ (PT, PD, TD, DD) - run 193 - is just in (PP, PT, PD, TD, DD); is deleted; ■ (PP, PT, PD, TD) - run 447 - is just in (PP, PT, PD, TD, DD); is deleted;

÷ Groups of at least 4^{th} order; list of suspects: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD)}: ■ (PT, TD, DD) - run 129, (PP, PT, DD) - run 376, (PP, TD, DD) - run 256, (PP, PT, TD) - are simultaneously in (PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD) and (PP, PT, TD, DD); are added; ■ (TT, TD, DD) - run 20, (PT, TT, TD) - run 144, (PT, TT, DD) - run 137, (PP, TT, TD) - run 271, (PP, TT, DD) - run 264 and (PP, PT, TT) - run 391 - are just in (PP, PT, TT, TD, DD); are deleted; ■ (PT, PD, DD) - run 185, (PP, PT, PD) - run 439 and (PP, PD, DD) - run 312 - are just in (PP, PT, PD, TD, DD); are deleted;

÷ Groups of at least 3^{rd} order; list of suspects: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD), (PT, TD, DD), (PP, PT, DD), (PP, TD, DD), (PP, PT, TD)}: ■ (PT, DD) - run 121 - is in runs 376, 129, 384, 448, 400; is added; ■ (TT, TD) - run 19, (PT, TT) - run 136, (TT, DD) - run 12, (PP, TT) - run 263 - in run 400; are deleted; ■ (PP, DD) - run 248 - in runs 256, 376, 384, 448, 400; is added; ■ (TD, DD) - run 5 - in runs 256, 129, 384, 448 and 400; is added; ■ (PT, TD) - run 128 - in runs 383, 129, 384, 448 and 400; is added; ■ (PP, TD) - run 255 - in runs 383, 256, 384, 448 and 400; is added; ■ (DP, DT) - run 3 - is not found in the higher-order groups; is added; ■ (PP, PT) - run 375 - in run 383, 376, 384, 448, 400; is added; ■ (PT, PD) - run 184 and (PD, DD) - in run 448; are deleted; ■ (TP, TT) - run 42 - is not found in the higher-order groups; has the highest susceptibility being closest to the 95 confidence for rejection (c/k = 1.09); is deleted.

÷ Groups of at least 2^{nd} order; list of suspects: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD), (PT, TD, DD), (PP, PT, DD), (PP, TD, DD), (PP, PT, TD), (PT, DD), (PP, DD), (TD, DD), (PT, TD), (PP, TD), (DP, DT), (PP, PT)}; it is imposed to continue the inclusion in the list of distinct populations by following the previous procedure: ■ PP is found in several groups of higher order; is added; ■ PT is found in several groups of higher order; is added; ■ PD is found in just one group of higher order (PP, PT, PD, TD, DD); is deleted; ■ TP is not found in any group of higher order; is added; ■ TT is found in just one group of higher order: (PP, PT, TT, TD, DD); is deleted; ■ TD is found in several groups of higher order; is

added; ■ DP is found in just one group of higher order: (DP, DT); is deleted; ■ DT is found in just one group of higher order: (DP, DT); is deleted; ■ DD is found in several groups of higher order; is added;

÷ All distinct groups; final list: {(PP, PT, TT, TD, DD), (PP, PT, PD, TD, DD), (PP, PT, TD, DD), (PT, TD, DD), (PP, PT, DD), (PP, TD, DD), (PP, PT, TD), (PT, DD), (PP, DD), (TD, DD), (PT, TD), (PP, TD), (DP, DT), (PP, PT), PP, PT, TP, TD, DD};

The list of possible phenotypes populations with the associated level of confidence can now be constructed and is presented in Tab. 2 (the classification is according to the obtained level of confidence).

Tab. 2

Number of phenotypes in cultivar during evolutions: populations parameterized by selection and survival having distinct distribution laws

Population	M(c/k) [n_obs]	Cat	Interpretation and the significance level
(DP, DT)	0.194 [128]	1	With a 5% risk of being in error produce a population different by all other pairs of methods.
TP	0.293 [255]		
(PP, PT, PD, TD, DD)	0.368 [16]	2	The hypothesis of come from the same population could not be rejected (at a 95% confidence)
(PP, PT, TT, TD, DD)	0.409 [16]		
PP	0.390 [255]	3.1	The hypothesis of identical distribution with more populations could not be rejected (at a 95% confidence)
TD	0.405 [255]		
DD	0.417 [255]		
PT	0.417 [255]		
(PP, TD)	0.400 [128]	3.2	The hypothesis of identical distribution with more populations could not be rejected (at a 95% confidence)
(PP, PT)	0.407 [128]		
(PP, DD)	0.417 [128]		
(PT, TD)	0.419 [128]		
(TD, DD)	0.419 [128]		
(PT, DD)	0.444 [128]		
(PP, PT, TD)	0.407 [64]	3.3	The hypothesis of identical distribution with more populations could not be rejected (at a 95% confidence)
(PP, TD, DD)	0.413 [64]		
(PP, PT, DD)	0.424 [64]		
(PT, TD, DD)	0.430 [64]		
(PP, PT, TD, DD)	0.422 [32]	3.4	The hypothesis of identical distribution with more populations could not be rejected (at a 95% confidence)

M(c/k) [n_obs] - Mean of c/k (number of observations in calculation of the mean); Cat = category
Remark: the smallest mean of c/k, the better relevance in the interpretation;
The c/k higher than 1 attract the susceptibility of interpretation rejection;

The analysis of results presented in Tab. 2 give us the possibility to statistically separate the populations. Thus, the 1st category of significance contains populations identified as distinct versus other populations (Fig. 1 – mean values on intervals of thousand generations; Fig. 2 – mean tendencies on intervals of thousand generations).

The analysis of Fig. 1 and 2 showed that the average number of phenotypes is different not only by the mean distribution law on evolutions (as it can be observed from Tab. 1) but also in terms of absolute value of the average number of phenotypes in cultivar.

Tournament selection associated to proportional survival (TP) produce a population of distinct phenotypes in cultivar in terms of their number (its mean is situated above the mean produced by remaining methods - PP, PT, PD, TT, TD, DD - in all moments of evolution – see Fig. 1). Moreover, the mean of phenotypes number obtained by applying the tournament selection and proportional survival systematically increase with almost one phenotype compared with all other methods (mean of 18.94 for TP; mean of 18.18 for PP+PT+PD+TT+TD+DD) and in all cases with more than one phenotype compared to overall mean (17.59), having the increasing tendency19 to

phenotypes during evolution (Fig. 2).

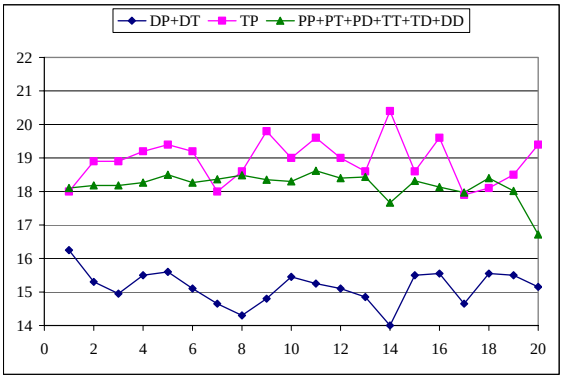

Fig. 1. Segregation of the populations of average number of viable phenotypes in the moments of evolution; populations: (DP and DT), TP and the rest of selection and survival strategies (PP, PT, PD, TT, TD, DD)

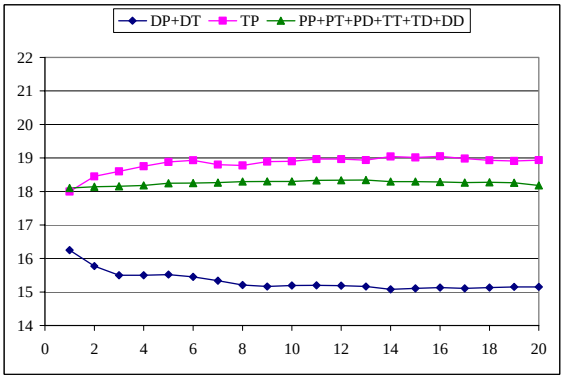

Fig. 2. Monte-Carlo experiment for the populations of viable phenotypes
Legend: Avg(Num. of viable phenotypes) vs. Millennia

Average number of phenotypes produced by deterministic selection accompanied by proportional (DP) or tournament (DT) survival, creates a distinct population of phenotypes in the cultivar in terms of numbers, which in its average lies at all evolution times (Fig. 1) than the average produced by the remaining methods of selection and survival (PP, PT, PD, TT, TD, SD). The average number of population phenotypes produced by deterministic selection accompanied by proportional (DP) or tournament (DT) survival produce a decrease of at least two phenotypes almost everywhere compared to the remaining methods (average of 15.15 for DP + DT, average of 18.18 for PP + PT + PD + TT + TD + DD) and with almost two phenotypes compared to the overall mean (17.59), with a decreasing trend of 15 phenotypes during evolution (Fig. 2).

The second examined case was of the most numerous populations which could not reveal different distributions laws (Fig. 3). On the populations presented in Fig. 3 it is remarkable the presence of four pairs in both populations (PP, PT, TD and DD) the distinct being made in each by the presence of the fifth pairs (PD and TT, respectively).

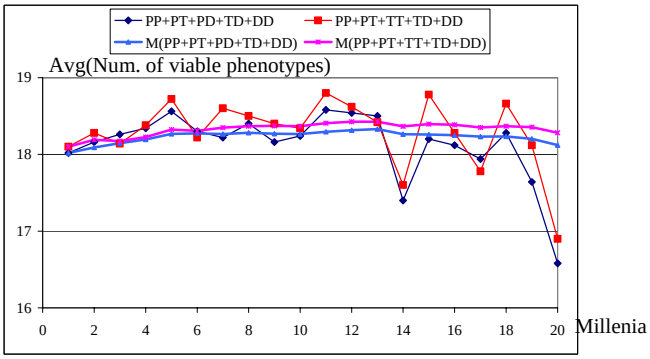

Fig. 3. Largest populations at 5% risk being in error for the number of phenotypes during evolutions

The explanation could be found in Tab. 1, the (PP, PT, TD and DD) population had an average score of nondiscrimination (c/k) of 0.422 which is even decreases with a significant amount at 0.368 when the PD sub-population was included; it was also decreased with an important amount when the TT sub-population was included. Both sub-populations proved to be significantly different (score in the pair of 0.42 - the hypothesis of belonging to the same population being rejected at a significance level of 5%) for their inclusion in the (PP, PT, TD, DD, PD, TT) together may not be

accepted as a single population distribution, the (PP, PT, TD, DD, PD, TT) group having a membership score of identically distributed populations of 0.92 at a significance level of 5% (almost accepted – value close to 1 – but not sufficient). An important remark is that the PD (proportional selection accompanied by deterministic survival) and TT (tournament selection accompanied by tournament survival) were not identified as distinct populations under the scheme of analysis presented in Tab. 1, which made it impossible the decision of their inclusion in the group five pairs of selection-survival strategies that contains the PP, PT, TD and DD.

The analysis of Table 1 brings satisfactory answers in terms of possibilities but not in terms of probabilities. If we fix the objective to obtain the most probable partition of populations obtained by selection and survival strategies, the analysis of pairs of two is the solution (see Tab. 2 - Annex).

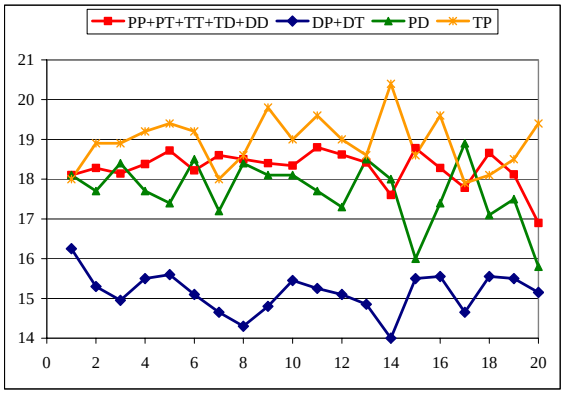

Fig. 4. The most probable partition of the populations of phenotypes number parameterized by selection and survival

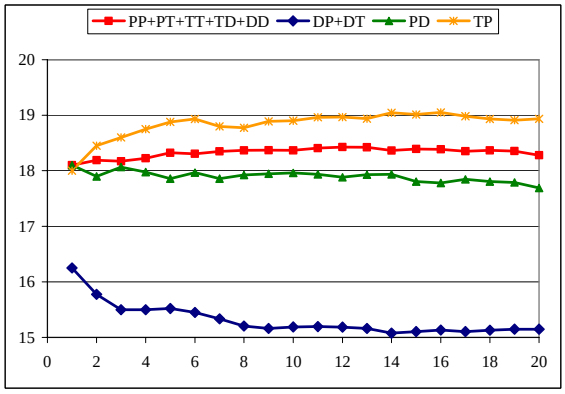

Fig. 5. Distinct populations for the number of phenotypes: The Monte-Carlo experiment
Legend: Avg(Num. of viable phenotypes) vs. Millennia

Thus, the combination obtained by applying the procedure described in Tab. 2 imposed the elimination of the three most week associations of population pairs: (PD, DD), (TP, TT) and (PD, PT. In this case, the solution (most probable solution) is given by the {(DD, PP, PT, TT, TD), (DT, DP), PD, TP} partition. The obtained result was graphically represented in terms of distribution of average number of phenotypes in cultivar (Fig. 4) and the average of means number of phenotype in cultivar (Fig. 5).

CONCLUSIONS

In 7% of investigated cases, the hypothesis that the groups are from the same population could not be rejected at a significance level of 5%. In 93% of cases, with a risk to be in error of 5%, it could be stated that the investigated groups of selection and survival strategies are from different populations.

Tournament selection accompanied by proportional survival (TP) produce a population of distinct phenotypes in cultivar in terms of their number (its mean is situated above the mean produced by remaining methods - PP, PT, PD, TT, TD, DD - in all moments of evolution), with a mean that increase with almost one phenotype compared with all other methods.

REFERENCES

1. Anderson, T. W., 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. Dashtbozorgi, Z., H. Golmohammadi. (2010). Prediction of air to liver partition coefficient for volatile organic compounds using QSAR approaches. European Journal of Medicinal Chemistry 45(6):2182-

90.

3. Di Serafino, D., S. Gomez, L. Milano, F. Riccio, G. Toraldo. (2010). A genetic algorithm for a global optimization problem arising in the detection of gravitational waves. Journal of Global Optimization 48(1):41-55.

4. Duval, B., J.-K. Hao. (2009). Advances in metaheuristics for gene selection and classification of microarray data. Briefings in Bioinformatics 11(1);art. no. bbp035:127-141.

5. Eisler, R., 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

6. Fraser AS. (1957a). Simulation of genetic systems by automatic digital computers. I. Introduction. Australian J. Biol. Sci. 10:484-491.

7. Fraser AS. (1957b). Simulation of genetic systems by automatic digital computers. II. Effects on linkage on rates of advance under selection. Australian J. Biol. Sci. 10:492-499.

8. Ghosh, P., M. C. Bagchi. (2009). QSAR modeling for quinoxaline derivatives using genetic algorithm and simulated annealing based feature selection. Curr. Med. Chem. 16(30):4032-48.

9. Hemmateenejad, B., A. R. Mehdipour, R. Miri, M. Shamsipur. (2010). Comparative qsar studies on toxicity of phenol derivatives using quantum topological molecular similarity indices. Chemical Biology and Drug Design 75(5):521-31.

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

11. Jäntschi, L. (2010b). 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

12. Jäntschi, L., S. D. Bolboacă, M. V. Diudea, R. E. Sestraş. (2010e). Average Trends over Millennia of Evolution Supervised by Genetic Algorithms. 1. Analysis of Genotypes, Submitted to arxiv.org (5 September 2010)

13. Jäntschi, L., S. D. Bolboacă, R. E. Sestraş. (2010a). Recording evolution supervised by a genetic algorithm for quantitative structure-activity relationship optimization. Applied Medical Informatics 26(2):89-100.

14. Jäntschi, L., S. D. Bolboacă, R. E. Sestraş. (2010d). 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

15. Jäntschi, L., S.D. Bolboacă, R.E. Sestraş. (2010c). 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

16. Juan, Y.-K., K. O. Roper, D. Castro-Lacouture, J. H. Kim. (2010). Optimal decision making on urban renewal projects. Management Decision 48(2):207-224.

17. Stamatakis, M., K. Zygourakis. (2010). A mathematical and computational approach for integrating the major sources of cell population heterogeneity. Journal of Theoretical Biology 266(1):41-61.

18. Teixidó, M., I. Belda, E. Zurita, X. Llorà, M. Fabre, S. Villaró, F. Albericio, E. Giralt. (2005). Evolutionary combinatorial chemistry, a novel tool for SAR studies on peptide transport across the blood-brain barrier. Part 2. Design, synthesis and evaluation of a first generation of peptides. Journal of Peptide Science 11(12):789-804.

19. Wu, J., J. Mei, S. Wen, S. Liao, J. Chen, Y. Shen. (2010). A self-adaptive genetic algorithm-artificial neural network algorithm with leave-one-out cross validation for descriptor selection in QSAR study. Journal of Computational Chemistry 31(10):1956-1968.

20. Yang, Y., T. Lin, X. L. Weng, J. A. Darr, X. Z. Wang. (2010). Data flow modeling, data mining and QSAR in high-throughput discovery of functional nanomaterials. Computers and Chemical Engineering. DOI: 10.1016/j.compchemeng.2010.04.018.

ANNEX

Tab. 2

The most probable solution interpreting the results from Tab. 1

The matrix of pairs, a cell of the matrix contains 0 if with a 5% risk of being in error the pair of methods given by the row and column of the cell produce populations with distinct distribution laws and 1 otherwise.

SS	PP	PT	PD	TP	TT	TD	DP	DT	DD
PP	1	1	0	0	1	1	0	0	1
PT	1	1	1	0	1	1	0	0	1
PD	0	1	1	0	0	0	0	0	1
TP	0	0	0	1	1	0	0	0	0
TT	1	1	0	1	1	1	0	0	1
TD	1	1	0	0	1	1	0	0	1
DP	0	0	0	0	0	0	1	1	0
DT	0	0	0	0	0	0	1	1	0
DD	1	1	1	0	1	1	0	0	1

The rows and columns of the matrix of pairs are changed so as to obtain the largest group of a compact of 1; one solution is the permutation on beside column.

SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
TP	0	0	0	0	1	1	0	0	0
DP	0	0	0	0	0	0	0	1	1
DT	0	0	0	0	0	0	0	1	1
PP	1	1	1	0	0	1	1	0	0
PT	1	1	1	1	0	1	1	0	0
TT	1	1	1	0	1	1	1	0	0
TD	1	1	1	0	0	1	1	0	0
DD	1	1	1	1	0	1	1	0	0
PD	1	0	1	1	0	0	0	0	0

The blocks of 1 are marked; the aim is to not have any 1 value in the neighbors of the block, which means that some identified combinations are random at a 5% risk of being in error

SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
DP	0	0	0	0	0	0	0	1	1
DT	0	0	0	0	0	0	0	1	1
TP	0	0	0	0	1	1	0	0	0
PP	1	1	1	0	0	1	1	0	0
PT	1	1	1	1	0	1	1	0	0
TT	1	1	1	0	1	1	1	0	0
TD	1	1	1	0	0	1	1	0	0
DD	1	1	1	1	0	1	1	0	0
PD	1	0	1	1	0	0	0	0	0

The elimination of values of 1 from the table should take into account the chance to be in error; analyzing the values from Table 1 it can be observed that the associations closest to the significance level of 5% are (PD, DD) and (TP, TT), both with a c/k ratio of 1.09. Thus, these combinations are associated to chance and are eliminated.

SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
DP	0	0	0	0	0	0	0	1	1
DT	0	0	0	0	0	0	0	1	1
TP	0	0	0	0	1	0	0	0	0
PP	1	1	1	0	0	1	1	0	0
PT	1	1	1	1	0	1	1	0	0
TT	1	1	1	0	0	1	1	0	0
TD	1	1	1	0	0	1	1	0	0
DD	1	1	1	0	0	1	1	0	0
PD	0	0	1	1	0	0	0	0	0

It is easy to observe that the combination (PD, PT) hindering the partition of populations by the pair (selection method, survival method); Analyzing the value of this combination (Table 1) the assumption is confirmed, after removal of the combinations (PD, DD) and (TP, TT) the (PD, PT) combination remained the combination with the smallest value of the c/k ratio (1.26). The (PD, PT) combination is eliminated.

SS	DD	PP	PT	PD	TP	TT	TD	DT	DP
DP	0	0	0	0	0	0	0	1	1
DT	0	0	0	0	0	0	0	1	1
TP	0	0	0	0	1	0	0	0	0
PP	1	1	1	0	0	1	1	0	0
PT	1	1	1	0	0	1	1	0	0
TT	1	1	1	0	0	1	1	0	0
TD	1	1	1	0	0	1	1	0	0
DD	1	1	1	0	0	1	1	0	0
PD	0	0	0	1	0	0	0	0	0

The rows and columns of the table could be permutated again for convenience.

SS	DD	PP	PT	TT	TD	DT	DP	PD	TP
TP	0	0	0	0	0	0	0	0	1
PD	0	0	0	0	0	0	0	1	0
DP	0	0	0	0	0	1	1	0	0
DT	0	0	0	0	0	1	1	0	0
PP	1	1	1	1	1	0	0	0	0
PT	1	1	1	1	1	0	0	0	0
TT	1	1	1	1	1	0	0	0	0
TD	1	1	1	1	1	0	0	0	0
DD	1	1	1	1	1	0	0	0	0