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 Associations

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 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 phenotypes 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. Here the records of association of phenotypes in pairs of four were analyzed and the conclusions were highlighted.

Keywords: Anderson-Darling Test, Genetic Algorithm (GA), Distribution; Associations.

INTRODUCTION

This paper continues the analysis started with the evolution of the number of viable genotypes (Jäntschi et al., 2010b) and of phenotypes (Jäntschi et al., 2010c).

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

MATERIAL AND METHODS

A structure-property relationships analysis has been conducted on a sample of 206 PCBs (Eisler and Belisle, 1996) in order to investigate the link between compound structure and octanol-water partition coefficient expressed in logarithmic scale. The 3D optimized geometry of the PCBs and the molecular descriptors family (MDF, which previously proved its usefulness in SPR analyses (Jäntschi and Bolboacă, 2007)) were the input data in this analysis. A heuristic has been developed and implemented in order to identity the optimal solution, the multiple linear regressions (MLR) with highest determination coefficient and estimation abilities (Jäntschi, 2010a). The GA was run for 46 times (46 experiments) and a series of parameters has been counting during the evolution of the heuristics from which the following were the interest for the present research: number of viable associations (*avg_obs*) for the same number of observed evolutions (*num_obs*) on thousand of generations (from 0..1000 to 19001-20000).

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 statistical evidence that a sample is from a given probability distribution. Details about the formulas applied are provided by (Jäntschi et al., 2010b).

RESULTS AND DISCUSSION

Five hundred and two alive regressions in cultivar in generation were the evolution occurred obtained for associations of 4 phenotypes were obtained and investigated. The smallest value of Anderson-Darling statistics was of 0.4239 (the PD - proportional selection accompanied by deterministic survival - & DD - deterministic selection and survival) while the highest value was of 15.6013 (TD - tournament selection accompanied by deterministic selection - & DT - deterministic selection accompanied by tournament survival). The c/k ratio (where k = Anderson-Darling critical value at a significance level of 5%, c = Anderson-Darling statistics) varied from 0.13 to 5.72. In almost 2% of the cases the hypothesis that the groups are from the same population could not be rejected at a significance level of 5%.

The summary of the viable associations (regressions with parameters significantly differed by zero at a maximum 5% risk of being in error) in terms of mean of their number (*avg_obs*) for the same number of observed generations (*num_obs*) grouped on thousand of generations (from 0..1000 to la 19001-20000) are presented in Table 1.

Tab. 1

Frequency (from 46 independent runs) of the viable associations in the moments of evolution according to the selection and survival Sel, Srv ∈ {P, T, D}

SS	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
PP	3793	3965	3532	4539	4569	4531	7444	4030	3599	3917	4594	3414	4301	3936	3788	2614	6620	4885	5054	4602
PT	4144	4872	5088	7798	8018	5068	5272	5347	2957	3959	4215	4104	5058	2404	5051	4403	3037	3448	1557	1574
PD	3264	3514	3800	3140	3024	4028	3343	4282	3440	3564	2919	3174	4354	3540	1868	2994	4747	2912	3501	2192
TP	3743	5506	5361	6637	7540	5593	4168	4687	6611	5077	6093	6260	4630	7280	6301	5013	4301	3704	4531	5459
TT	4909	5279	4958	4569	6587	4426	6074	5967	4917	4348	5513	3446	3762	6979	6163	4539	5769	5910	6100	5029
TD	3456	4023	3321	3905	4011	3655	3453	4034	3443	3641	4683	5390	3758	3905	6060	4418	2621	4099	3070	2590
DP	1786	1448	1026	2005	1698	999	862	1293	1328	1413	2473	1679	1192	772	1205	1788	1250	1422	1777	2800
DT	2172	1595	1513	1524	1532	1878	1726	1180	1512	1992	1117	719	1585	1487	2058	1428	1016	1757	1492	572
DD	2795	3158	4088	3617	3398	2930	3014	3647	4735	3352	2523	4104	5902	3341	4282	2853	2100	4008	2445	1605

D = deterministic; P = proportional; T = tournament; Sel = selection strategy; Srv = survival strategy; SS = selection-survival

The hypothesis that different associations of populations obtained by using different selection and survival strategies in terms of the average numbers of viable associations was verified by applying the Anderson-Darling test. The results showed that for 10 groups of pairs of selection-survival methods could not be identify a difference statistically significant between populations' distributions laws.

The largest groups of maximum non-discrimination order proved to be of 3rd order (PD - proportional selection accompanied by deterministic survival, TD - tournament selection accompanied by deterministic survival, DD - deterministic selection accompanied by deterministic survival) - run 66 - and (PP - proportional selection accompanied by proportional survival, PT - proportional selection accompanied by tournament survival, TD - tournament selection accompanied by deterministic survival) - run 383; these groups enter automatically in the list of suspects. The groups of smaller order are as resulted by applying the inclusion algorithm:

÷ Groups of 2nd order; list of suspects: {(PD, TD, DD), (PP, PT, TD)}

  o (PD, DD) - run 58, (TD, DD) - run 5, and (PD, TD) - run 65 - are present just in (PD, TD, DD); are deleted;

  - (TP, TT) - run 42 and (DP, DT) - run 3 - where not found in any of the higher order groups; are added;
  - (PP, PT) - run 375, (PP, TD) - run 255, and (PT, TD) - run 128 - where found just in (PP, PT, TD); are deleted.
- Groups of at least 2nd order; list of suspects: {(PD, TD, DD), (PP, PT, TD), (TP, TT), (DP, DT)}; be further included in the list of distinct populations using the same procedure:
  - PP, PT were found in one group (PP, PT, TD); are deleted;
  - PD, DD were found in one group (PD, TD, DD); are deleted;
  - TP, TT were found in one group (TP, TT); are deleted;
  - DP, DT were found in one group (DP, DT); are deleted;
  - TD was found in two groups of superior order; is added;
- All distinct groups; final list: {(PD, TD, DD), (PP, PT, TD), (TP, TT), (DP, DT), TD}.

The list of possible associations of phenotypes accompanied by the associated confidence level can now be created. The results are presented in Table 2 and were ordered by the confidence of the obtained results that contains those groups for which belonging to the identical distributed populations was not rejected at a 95% confidence.

Tab. 2

Number of phenotypes associations during evolution: populations with distinct distribution depending on selection and survival strategies

Population	c/k	*Interpretation at 5% risk of being in error*
(PD, TD, DD)	1.29	The hypothesis of belonging to identical populations could not be rejected.
(PP, PT, TD)	1.12	The hypothesis of belonging to identical populations could not be rejected.
(TP, TT)	4.96	The hypothesis of belonging to identical populations could not be rejected.
(DP, DT)	3.15	The hypothesis of belonging to identical populations could not be rejected.
TD	-	Proved to be identical distributed with other two populations.

Note that the c/k values of (TP, TT) and (DP, DT) associations were far away from rejecting the hypothesis of belonging to identically distributed populations. Not in the same situation were the (PD, TD, DD) and (PP, PT, TD) groups for which the c/k were too close (1.29 and 1.12) to be rejected for belonging to identically distributed populations (1.00). Moreover, the number of partition possibilities of the results could be easy calculated and there were exactly 3 possibilities (for which the sum of c/k ratios were calculated):

- {(PD, TD, DD), (TP, TT), (DP, DT), (PP, PT)}, Σc/k = 1.29 + 4.96 + 3.15 + 2.04 = 11.44;
- {(PD, DD), (TP, TT), (DP, DT), (PP, PT, TD)}, Σc/k = 5.72 + 4.96 + 3.15 + 1.12 = 14.95;
- {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD}, Σc/k = 5.72 + 4.96 + 3.15 + 2.04 = 15.87;

Thus, the most probable apparition of populations occurred after the {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD} schema (Table 3).

Tab. 3

The most probable partition in selection vs. survival on average number of phenotypes associations observed in cultivar; groups ranked by the probability of association

	D	P	T
D	1	3	3
P	1	4	4
T	5	2	2

An indicator that the obtained solution is the most probable is that the (PP, PT, TD) association of 3rd order is the weakest association of this order class (c/ k = 1.12). Simultaneously, the (PT, TD) association is the weakest 2nd order association (c/k = 1.04). The both above-presented associations are rejected by the proposed solution {(PD, DD), (TP, TT), (DP, DT), (PP, PT), TD}. Table 3 showed us how to statistical separate the populations while Figure 1 shows the average interval of thousands of generations for each population trends. Figure 2 presented the average tendencies on intervals of thousands of generations.

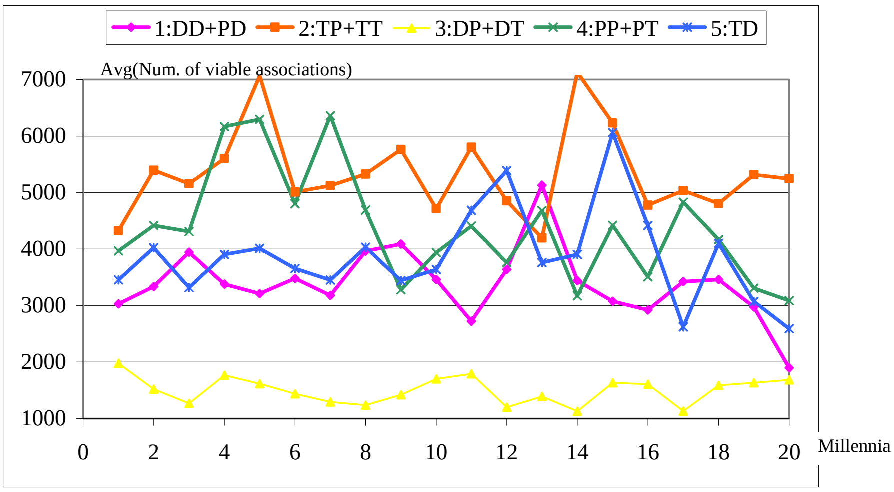

Fig. 1. Number of phenotypes associations in cultivar: significantly distinct populations in relation with selection and survival strategies

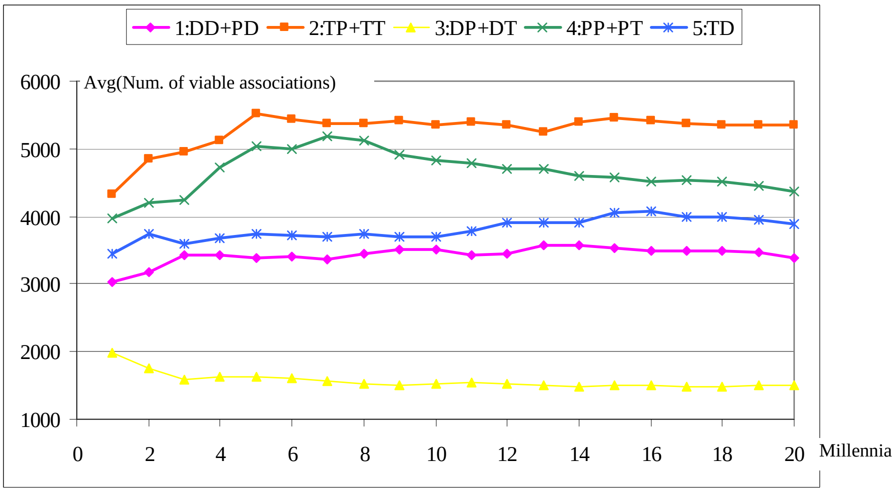

Fig. 2. Mean tendencies of phenotypes associations represented in Figure 1

The analysis of Figure 1 and 2 showed that the average number of associations is different not only in terms of distribution law but also in terms of absolute value of number of alive associations in cultivar.

The tournament selection accompanied by the proportional survival (TP) produce a distinct population of associations in cultivar in terms of their number, which in average is situated in almost all evolution moments (Fig. 1) above the average obtained by all other methods (PP, PT, PD, TT, TD, DD). The average number of associations occurred when tournament selection is accompanied by proportional survival (TP) that increased in almost all cases with almost 1 association compared to all other methods (Fig. 2) (mean of 18.94 for TP; mean of 18.18 for PP + PT + PD + TT + TD +

DD). Moreover, the increase in association is with one over the overall mean (17.59) with an increase tendency to 19 associations during evolution.

The average number of associations produced by deterministic selection accompanied by proportional survival (DP) or tournament survival (DT) create a distinct population of associations in terms of their number, which in its average is situated in all moment of evolution (Figure 1) bellow the average produce by all other methods (PP, PT, PD, TT, TD, DD). Moreover, the deterministic selection accompanied by the proportional survival (DP) or tournament survival (DT) produce the decrease with at least one association in all moments of evolution compared to all other methods (mean of 15.15 for DP+DT; mean of 18.18 for PP+PT+PD+TT+TD+DD) and with almost two associations compared to overall mean (17.59) having a decreasing tendency to 15 associations during the evolution.

CONCLUSIONS

The hypothesis that different associations of selection and survival strategies populations in terms of the average numbers of viable associations was verified by applying the Anderson-Darling test. Ten groups of pairs of selection-survival methods proved not to be statistically significant different in terms of populations' distributions laws.

The average number of associations proved to be different not only in terms of distribution laws but also in terms of absolute value of number of alive associations in cultivar.

The associations produced by deterministic selection accompanied by proportional or tournament survival create a distinct population of associations in terms of their number, with an average situated in all moment of evolution bellow the average produce by all other methods. Moreover, the deterministic selection accompanied by the proportional or tournament survival produce the decrease with at least one association in all moments of evolution (average of 15.15) compared to all other methods (average of 18.18) and with almost two associations compared to overall mean (17.59) having a decreasing tendency to 15 associations during the evolution.

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. 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

3. Jäntschi, L., Bolboacă S. D. (2007). Structure-Activity Relationships on the Molecular Descriptors Family Project at the End. Leonardo Electronic Journal of Practices and Technologies 11:163-80.

4. 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

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

6. Jäntschi, L., S. D. Bolboacă, M. V. Diudea, R. E. Sestraş. (2010c). Average Trends over Millennia of Evolution Supervised by Genetic Algorithms. 2. Analysis of Phenotypes, Submitted to arxiv.org (5 September 2010)