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

Tendency of Evolution Supervised by Genetic Algorithms

Lorentz JÄNTSCHI[1)], Sorana D. BOLBOACĂ[2)], Mugur C. BĂLAN[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 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 had been developed and implemented in order to identify the optimal determination coefficient of using a multiple linear regression approach for structure-activity relationships. An experiment was conducted using Molecular Descriptors Family as genetic material and a sample of 206 polychlorinated biphenyls with measured octanol-water partition coefficients as environment of adaptation. The GA was repeated for 46 times for every pair of survival and selection strategies from proportional, tournament and deterministic ones. The Fisher-Tippett distribution was found suitable to characterize a moment of evolution. Tendency models of distribution were constructed from the pool of all Fisher-Tippett distributions in every recorded generation from 1 to 20000.

Keywords: Genetic algorithm; Fisher-Tippett distribution; Evolution; Determination coefficient

INTRODUCTION

The aim of the research conducted in (Jäntschi and Sestraş, 2010) were to assess the suitability of genetic algorithms for make inferences about the use of different selection and survival strategies in breeding. The research covered projecting of a genetic algorithm (GA), implementation of an evolutionary program based on it, and then the analysis of the influence of different selection and survival strategies on evolution controlled by the genetic algorithm feed with data for structure-activity relationships (SARs) optimization in a series of biologically active compounds. Three objectives were followed:

- (method) design of the GA (including defining of the hard problem); formulation of the problem in genetic terms; projecting of the GA; implementation and documentation of the evolutionary program embedding the GA;
- (results) simulation of the evolution (defining of the observables; defining of the contingency between selection and survival strategy; projecting of the statistical experiment; run of the experiment;
- (analysis) analysis and interpretation of the runs results about qualitative observables and about evolution objective (was set to r^2 - determination coefficient) - quantitative observable during evolution.

The GA are described in (Jäntschi, 2009) and a series of other papers (Jäntschi *et al.*, 2010mh; Jäntschi *et al.,* 2010ga) analyses a series of the results of interest. The aim of this work is to give inferences about the tendency in evolution using different selection and survival strategies.

MATERIALS AND METHODS

The raw data obtained in 46 independent executions of evolutionary program implementing the GA set to go near to the best multiple linear regression (MLR) model with four Molecular Descriptors Family (MDF) structure descriptors are online available and are given in Table 1.

Tab. 1

Simulation results

Selection*	Survival*	Configuration**	Evolution**
Proportional	Proportional	PCB_4044_cfg.txt	PCB_4044_evo.txt
Proportional	Deterministic	PCB_2441_cfg.txt	PCB_2441_evo.txt
Proportional	Tournament	PCB_9878_cfg.txt	PCB_9878_cfg.txt
Deterministic	Proportional	PCB_5108_cfg.txt	PCB_5108_evo.txt
Deterministic	Deterministic	PCB_6369_cfg.txt	PCB_6369_evo.txt
Deterministic	Tournament	PCB_6690_cfg.txt	PCB_6690_evo.txt
Tournament	Proportional	PCB_5828_cfg.txt	PCB_5828_evo.txt
Tournament	Deterministic	PCB_4872_cfg.txt	PCB_4872_evo.txt
Tournament	Tournament	PCB_1758_cfg.txt	PCB_1758_evo.txt

*There are following pairs of selection and survival strategies (PP, PD, PT, DP, DD, DT, TP, TD, TT)
**Files available at: http://l.academicdirect.org/Horticulture/GAs/MLR_MDF_selection_vs_survival

The raw data from Table 1 were processed with Excel's macro of EasyFit program (EasyFitXL) computing parameters of the Fisher-Tippett (Fisher and Tippett, 1928) theoretical distributions in every pair of selection and survival strategy and every generation (three parameters; nine pairs of strategies; 20000 generations). The proof that Fisher-Tippett (FT) distribution is the distribution of the determination coefficient of multiple linear regressions between MDF may be found in (Jäntschi and Sestraş, 2010).

Location (λ), scale (β) and shape (k) parameters from maximum likelihood of observed distributions agreeing with theoretical FT distributions are given in Figures from 1 to 3, and analysis of tendency (including exponential smoothing of them conducted with Statistica software) are given in Tables from 2 to 4.

To highlight the location (λ) dependence (which is not linear) from Fisher-Tippett distribution of number of generations was performed a regression analysis (Tab. 4) using a sample of 16 frequently observed regression models using SlideWrite application.

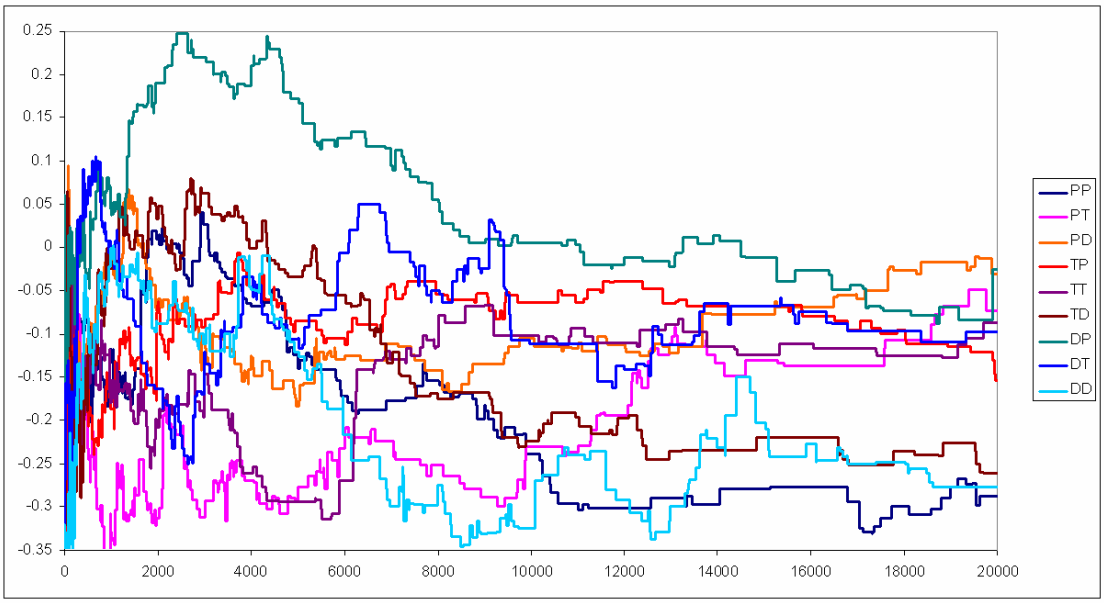


Fig. 1. Shape parameter (k) of FT distribution: MLE estimation from observations

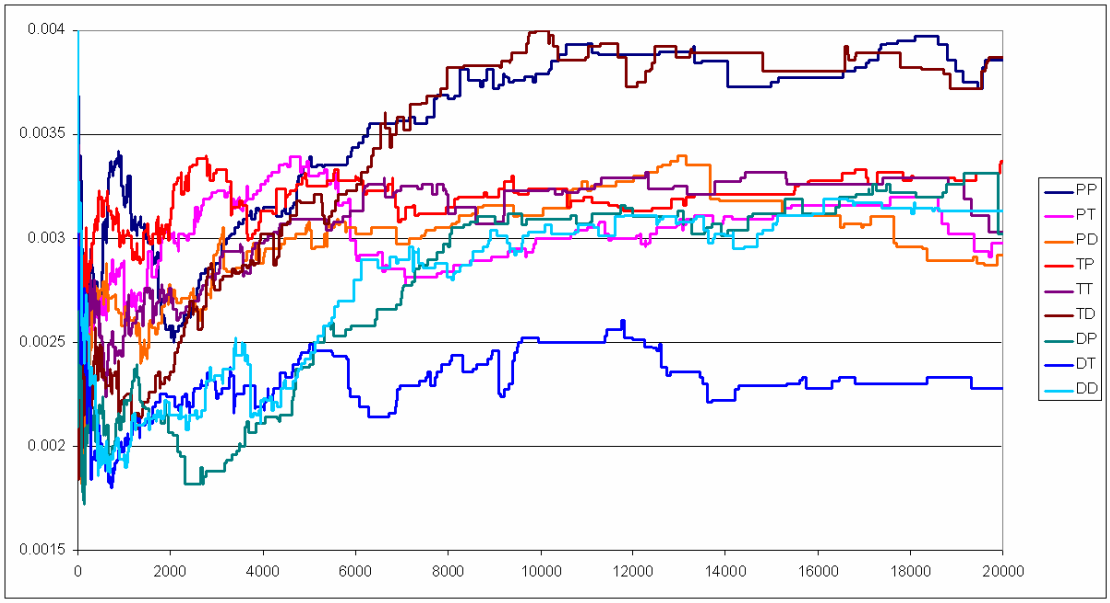


Fig. 2. Scale parameter (β) of FT distribution: MLE estimation from observations

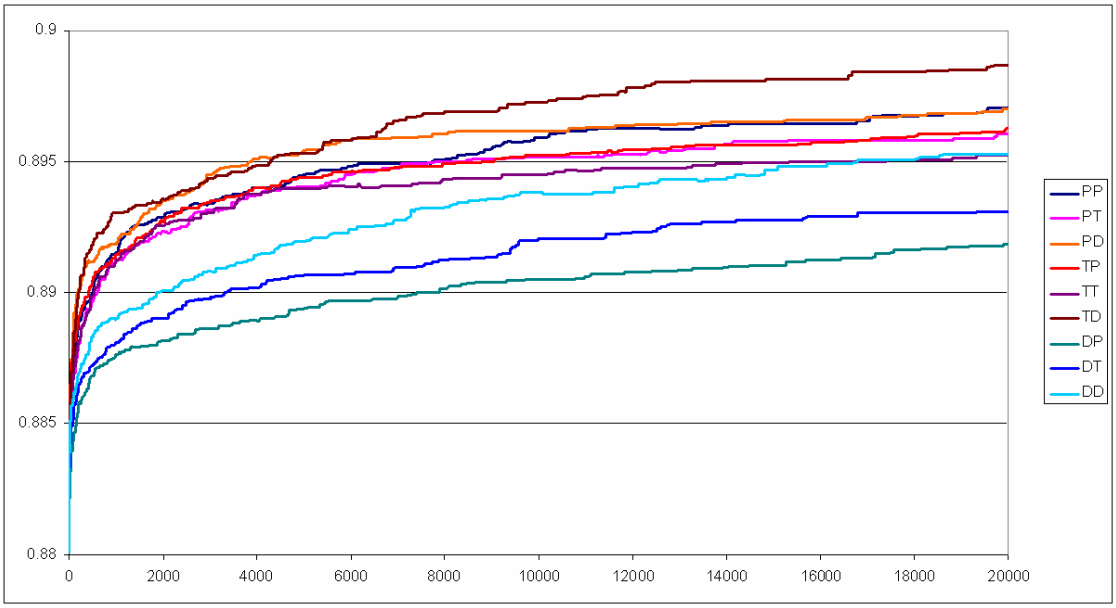


Fig. 3. Location parameter (λ) of Fisher-Tippett distribution: MLE estimation from observations

Tab. 2

Transformation of the shapes of the Fisher-Tippett distributions and tendency equation

k	S0	MAE	SSE	MSE	$S(G) = a_0 + a_1 \cdot G$	r	F	t_0	t_1
PP	-0.206	1.3E-3	0.45	2.3E-5	$a_0 = -0.1912$; $a_1 = -1.47 \cdot 10^{-6}$	0.858	56017	-2661	-237
PT	-0.200	1.6E-3	0.64	3.2E-5	$a_0 = -0.2108$; $a_1 = 1.08 \cdot 10^{-6}$	0.809	38004	-3287	195
PD	-0.093	1.5E-3	1.38	6.9E-5	$a_0 = -0.0961$; $a_1 = 3.12 \cdot 10^{-7}$	0.374	3245	-1519	57
TP	-0.082	1.4E-3	1.54	7.7E-5	$a_0 = -0.0833$; $a_1 = 1.24 \cdot 10^{-7}$	0.173	619	-1444	25
TT	-0.142	1.4E-3	0.97	4.8E-5	$a_0 = -0.1476$; $a_1 = 5.58 \cdot 10^{-7}$	0.510	7044	-1924	84
TD	-0.150	1.6E-3	0.93	4.6E-5	$a_0 = -0.1352$; $a_1 = -1.47 \cdot 10^{-6}$	0.831	44775	-1684	-212
DP	-0.041	1.2E-3	1.19	6.0E-5	$a_0 = -0.0193$; $a_1 = -1.32 \cdot 10^{-6}$	0.800	35649	-238	-189
DT	-0.081	1.6E-3	0.96	4.8E-5	$a_0 = -0.0797$; $a_1 = -1.35 \cdot 10^{-7}$	0.132	354	-961	-19
DD	-0.216	1.8E-3	1.01	5.1E-5	$a_0 = -0.0207$; $a_1 = -9.52 \cdot 10^{-7}$	0.592	10779	-1949	-104

Tab. 3

Transformation of the scales of Fisher-Tippett distributions and the tendency law

B	S0	MAE	SSE	MSE	$S(G) = a_0 + a_1 \cdot G$		t_0	t_1	r	F
PP	0.0036	4.6E-6	8.6E-6	4.3E-10	3.541E-3	5.5E-9	11599	210	0.829	43915
PT	0.0030	4.3E-6	1.1E-5	5.4E-10	2.996E-3	8.2E-10	13116	42	0.283	1739
PD	0.0030	4.2E-6	9.7E-6	4.9E-10	2.983E-3	1.9E-9	11534	84	0.513	7134
TP	0.0032	3.9E-6	5.9E-6	3.0E-10	3.192E-3	8.9E-10	25597	82	0.503	6786
TT	0.0031	4.1E-6	1.3E-5	6.4E-10	3.072E-3	2.9E-9	14164	157	0.743	24578
TD	0.0035	5.1E-6	1.2E-5	6.2E-10	3.419E-3	7.9E-9	7700	205	0.823	41884
DP	0.0028	4.0E-6	9.0E-6	4.5E-10	2.730E-3	7.1E-9	9043	271	0.887	73442
DT	0.0023	3.8E-6	6.5E-6	3.3E-10	2.296E-3	6.1E-10	12484	38	0.263	1482
DD	0.0028	4.5E-6	1.3E-5	6.5E-10	2.745E-3	5.6E-9	10150	241	0.862	58091

Tab. 4

The regression analysis for locations of the Fisher-Tippett distributions

λ	Exponential smoothing				Model			Significance				
SS	S0	MAE	SSE	MSE	a_0	a_1	a_2	t_0	t_1	t_2	r	F
					$\lambda(G) = a_0 + a_1 \cdot \ln(G + a_2)$							
PP	0.8952	1.2E-5	9.4E-4	4.7E-8	0.89357	$1.82 \cdot 10^{-4}$	0.867	640993	1174	9	0.9966	737924
PD	0.8956	1.1E-5	1.0E-3	5.0E-8	0.89422	$1.55 \cdot 10^{-4}$	-0.344	344366	536	-12	0.9833	146308
TP	0.8947	1.1E-5	9.0E-4	4.5E-8	0.89333	$1.54 \cdot 10^{-4}$	-0.213	666193	1027	-8	0.9954	539368
TT	0.8941	1.0E-5	8.0E-4	4.0E-8	0.89286	$1.40 \cdot 10^{-4}$	-0.348	361929	507	-12	0.9814	130838
					$\lambda(G) = a_0 + a_1 \cdot \ln(G)$							
PT	0.8946	1.1E-5	9.4E-4	4.7E-8	0.89309	$1.69 \cdot 10^{-4}$	-	502833	853	-	0.9932	727191
					$\lambda(G) = a_0 + a_1 \cdot G^{a_2}$							
TD	0.8966	1.3E-5	1.1E-3	5.5E-8	0.89465	$6.84 \cdot 10^{-4}$	0.117	38625	39	76	0.9923	321942
DP	0.8901	8.4E-6	4.2E-4	2.1E-8	0.88916	$2.02 \cdot 10^{-4}$	0.171	132950	51	124	0.9947	467811
DT	0.8914	8.5E-6	4.6E-4	2.3E-8	0.89016	$3.19 \cdot 10^{-4}$	0.151	100793	56	125	0.9957	574382
DD	0.8931	9.8E-6	6.2E-4	3.1E-8	0.89173	$2.93 \cdot 10^{-4}$	0.172	134833	75	183	0.9975	1010738

An important observation is the shape parameter value (k) is negative in all cases (Tab. 2), which customizes the Fisher-Tippett distribution to Weibull distribution. Small negative values of shape parameter, ranging between -0.25 and 0 with slight trends (Tab. 2) increasing (for PT, PD, PT and TT) or decreasing (for PP, TD, DP, DT and SD) explains why the hypothesis when testing for Gumbel distribution in most cases could be accepted without that this would entail accepting the hypothesis of Gumbel distribution.

Other important observation is the tendency of scale parameter (β) is increasing in all cases, indicating the increasing trend (during evolution) of variability for all pairs of strategies under observation. It can reveal distinct groups of forms and scales between the strategy pairs using principal components analysis on data from Tables 2 and 3. Table 5 gives the groups that were created by the shape (k) and scale (β) tendencies of the distribution.

Tab. 5

Groups of shape and scale in selection and survival

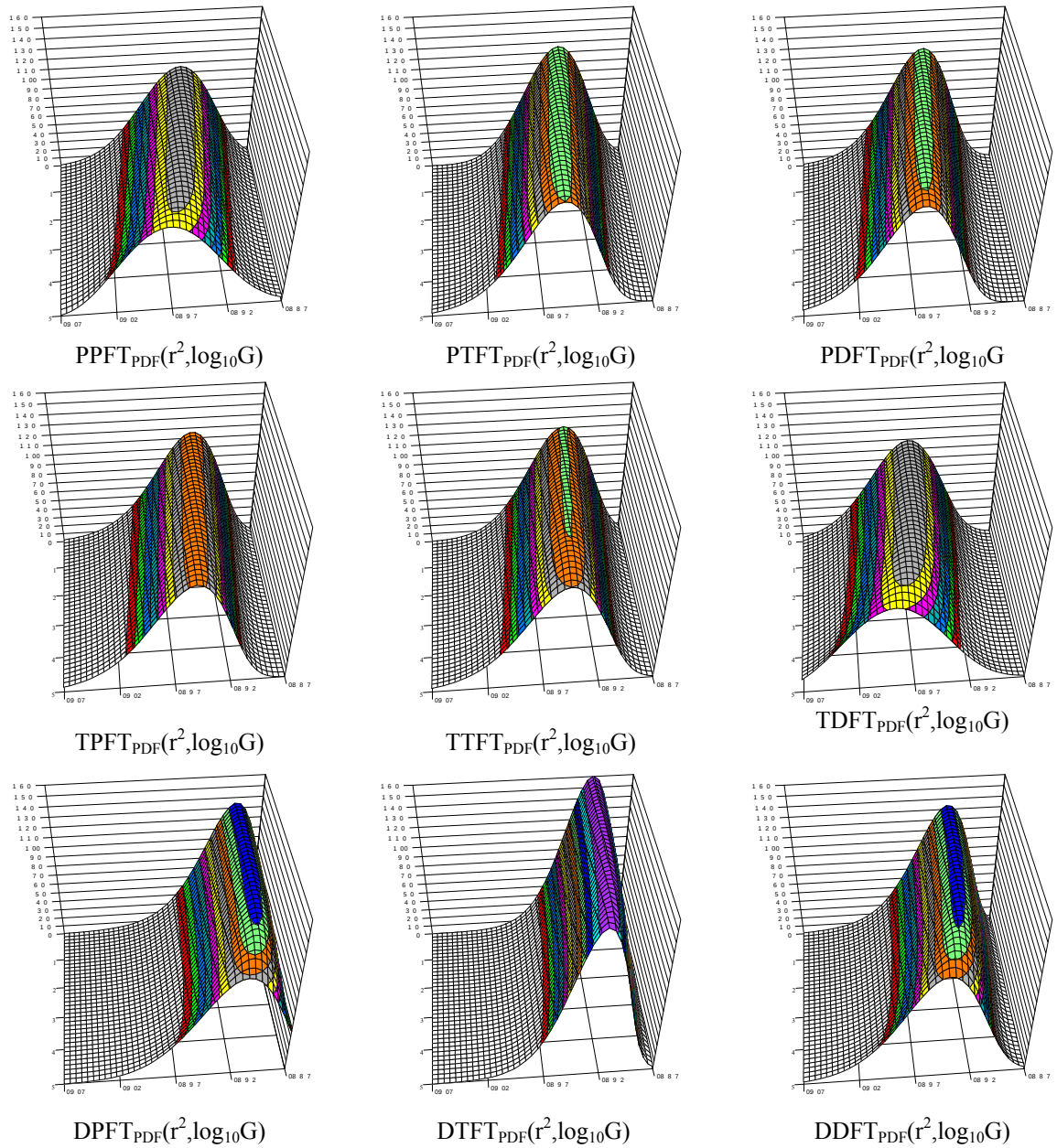
Parameter of the Fisher-Tippett distribution	Groups of selection and survival
Shape (k)	DD, DP; PP, TP
Scale (β)	PT, PD; DD, DP

Equations can be expressed as Fisher-Tippett distribution of evolution's objective in time-dependent form (depending on the generation of the evolution). What we have done were the replacing of the expressions obtained for the form (k) - Table 2, scale (β) - Table 3 and location (λ) - Table 4, in the expressions for the probability density function (PDF) and probability distribution function (CDF):

$$FT_{PDF}(X)=\frac{1}{\beta}\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)(1+k\frac{x-\lambda}{\beta})^{-1-1/k},\ FT_{CDF}(X)=\exp\left(-(1+k\frac{x-\lambda}{\beta})^{-1/k}\right)$$

Expressions' giving the probability distribution function tends to time-dependent Fisher-Tippett distributions ($SSFT_{CDF}$, where SS - selection and survival strategies) which are in form (eg. for CDF of PP strategies pair):

$$PPFT_{CDF}(r^2,G)=\exp\left(-\left(1+(-0.1912-1.47\cdot10^{-6}\cdot G)\frac{r^2-0.89357-1.82\cdot10^{-4}\cdot\ln(G+0.867)}{3.541\cdot10^{-3}+5.5\cdot10^{-9}\cdot G}\right)^{1/(0.1912+1.47\cdot10^{-6}\cdot G)}\right)$$



$PPFT_{PDF}(r^2,\log_{10}G)$ $PTFT_{PDF}(r^2,\log_{10}G)$ $PDFT_{PDF}(r^2,\log_{10}G$

$TPFT_{PDF}(r^2,\log_{10}G)$ $TTFT_{PDF}(r^2,\log_{10}G)$ $TDFT_{PDF}(r^2,\log_{10}G)$

$DPFT_{PDF}(r^2,\log_{10}G)$ $DTFT_{PDF}(r^2,\log_{10}G)$ $DDFT_{PDF}(r^2,\log_{10}G)$

Fig. 4. Tendency of time-dependent Fisher-Tippett PDFs

Legend (see Tables from 2 to 4 for k(t), β(t) and λ(t) function expressions):

$$SSFT_{PDF}(x,t)=\frac{1}{\beta(t)}\exp\left(-(1+k(t)\frac{x-\lambda(t)}{\beta(t)})^{-1/k(t)}\right)(1+k\frac{x-\lambda(t)}{\beta(t)})^{-1-1/k(t)},$$

where: SS ∈ {PP, PT, PD, TP, TT, TD, DP, DT, DD}

The expressions that give the trend probability density function (PDF) can be obtained in similar manner as for the CDF. They are more complicated to be given as mathematical expressions, but are more suggestive their three-dimensional representation (Fig. 4). Figure 4 gives three-dimensional representations of probability density functions in which was used instead of variable generation (G) its base-10 logarithm ($\log_{10}$ scale of G).

CONCLUSIONS

The study revealed that the Fisher-Tippett distribution is suitable to describe the evolution in any moment of it under the constraints defined by the genetic algorithm. From whole pool of observed distributions was possible the extraction of the tendency for the distribution, and thus a time-dependent distribution law to be expressed for every pair of selection and survival strategy.

REFERENCES

1. Fisher, R. A. and L. H. C. Tippett (1928). Limiting Forms of the Frequency Distribution of the Largest or Smallest Member of a Sample. Proceedings of the Cambridge Philosophical Society 24:180-190.

2. Jäntschi, L. (2009). A genetic algorithm for structure-activity relationships: software implementation, Manuscript, ArXiv.org deposit (online from June 26, 2009), http://arxiv.org/abs/0906.4846

3. Jäntschi, L. (2010). 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.

4. Jäntschi, L, S. D. Bolboacă and R. E. Sestraş (2010ga). A Study of Genetic Algorithm Evolution on the Lipophilicity of Polychlorinated Biphenyls, Chemistry and Biodiversity, 7(8):1978-1989, http://dx.doi.org/10.1002/cbdv.200900356.

5. Jäntschi, L, S. D. Bolboacă and R. E. Sestraş (2010mh). Meta-heuristics on quantitative structure-activity relationships: study on polychlorinated biphenyls, Journal of Molecular Modeling, 16(2):377-386, http://dx.doi.org/10.1007/s00894-009-0540-z

6. *** EasyFit (©MathWave Tech.): http://mathwave.com

7. *** Estimating Fisher-Tippett distribution parameters, measure the agreement with observations GA (© Lorentz JÄNTSCHI): http://arxiv.org/abs/0906.4846

8. *** Supervising evolution, recording observations HyperChem v. 8.0 (© HyperCube Inc.): http://hyper.com

9. *** Molecular modeling of PCB 3D structures MDF (© Lorentz JÄNTSCHI): http://l.academicdirect.org/Chemistry/SARs/MDF

10. *** Generating the family of 3D structure descriptors SlideWrite (© Advanced Graphics Software Inc.): http://slidewrite.com

11. *** 3D plotting of time-dependent Fisher-Tippett tendency distributions Statistica (© StatSoft Inc.): http://statsoft.com Principal component analysis of scale and shape parameters