Applied Medical Informatics
Original Research Vol. 26, No. 2/2010, pp: 89-100

Recording Evolution Supervised by a Genetic Algorithm for Quantitative Structure-Activity Relationship Optimization

Lorentz JÄNTSCHI[1*], Sorana D. BOLBOACĂ[2], Radu E. SESTRAŞ[3]

[1] Technical University of Cluj-Napoca, 103-105 Muncii Bdv., 400641 Cluj-Napoca, Cluj, Romania.
[2] "Iuliu Haţieganu" University of Medicine and Pharmacy Cluj-Napoca, 13 Emil Isac, 400349 Cluj-Napoca, Cluj, Romania.
[3] University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca, 3-5 Mănăştur, 400372 Cluj-Napoca, Romania.
E-mail(s): lori@academicdirect.org; sbolboaca@umfcluj.ro; rsestras@usamvcluj.ro.

* Author to whom correspondence should be addressed; Tel.: +4-0264-401775; Fax: +4-0264-401768.

Received: 18 June 2010 / Accepted: 22 June 2010 / Published online: 28 June 2010

Abstract

A genetic algorithm for structure-activity relationships optimization was developed and implemented. The genetic algorithm was designed to be feed with families of molecular descriptors, and was tested on Molecular Descriptors Family. The objective of the genetic algorithm was to optimize the multiple linear regressions with four descriptors for prediction of octanol-water partition coefficient (expressed in logarithmic scale) of a series of 206 polychlorinated biphenyls. Relevant factors for evolution were parameterized in the implementation of the evolutionary program. The configuration file allows running of the genetic algorithm under different settings of parameters. The defined parameters were parameters used to characterize the adaptation to the environment (three parameters), to characterize the breading sample (four), the reproduction (four), the evolution objective (two), the selection (ten), the survival (four), and the program execution (three).

Keywords: Simulating evolution; Genetic algorithms (GAs); Structure-Activity Relationships (SARs); Multiple Linear Regressions (MLRs).

Introduction

Structure-Activity Relationships Optimization

Mathematical approach of SAR (structure-activity relationships) on BAC (biologically active compounds), started in nineteen century and are capitalized through the born of the quantitative structure-activity relationships (QSAR) concept [1], a mathematical tool able to describe the quantitative link between chemical structure and biological activity. SAR records were communicated in scientific literature since 1868, when (first) Crum-Brown & Fraser want to seen if the activity of compounds is or not a function of chemical structure and composition [2]. Only after almost forty years the QSAR paradigm was found useful in agro-chemistry, pharmaceutical chemistry, toxicology, etc [3]. Scientific literature contains many reports on usage of QSARs in the methodology of designing new BACs (the monograph [4] covers a good part of them).

Evolution Supervised by Genetic Algorithms

Hard [5] and soft [6] inheritance, selection and survival [7], traits [8] and genes [9] crossover, a long and contentious debate over the 19th century [10] are all pieces from a puzzle build today the modern genetics [11] and were the sources of inspiration for genetic algorithms (GAs).

First studies on simulating an evolution were published by Nils Aall BARRICELLI [12]. Alex FRASER (1923-2002) also published a series of studies about simulation of artificial selection of organisms having multiple loci controlling a measurable trait. Fraser's simulations [13-25] included all essential elements of modern GAs.

Recording Evolution Supervised by a Genetic Algorithm

The research question "How the evolution can be observed and characterized via different parameters characterizing the sample supervised to evolutes?" are not enough explored in the specialty literature on genetic algorithms subject. Studies of different operators essential for evolution are focused mainly on algorithm efficiency and a representative manuscript for this approach is the collection from [26].

A small number of studies described the influence of the evolution strategy on evolution objective and almost nothing about the influence of different parameters characterizing the evolving sample on evolution objective.

The GAs passed out the border of the informatics field a long time ago, because of their results capitalization potential. PhD theses that had as objectives project, implement and use of genetic algorithms are found practically in all fields of research. Thus, in agriculture GAs were found useful to crop planning [27], in constructions to assess the risk of soil damage [28], in bioengineering to efficient control of pollution at a hydrographic basin level [29], in chemistry at design of sensor-based controlled processes [30], in economics at optimization of problems with multiple options [31], in management at multi-scale processes modeling [32], in mechanics at optimization of composite structures [33], in environment at strategy chousing for water quality control [34], in biology in phylogenetic analysis [35] and evolution studies [36].

Research Aim of the Paper

The research aim to project a GA and to implement it as an evolutionary program, capable to record the influence of different samples, environments and intrinsic parameters of genetic algorithm on evolution supervised by the genetic algorithm feed with data for structure-activity relationships optimization in a series of biologically active compounds.

Material and Method

Material

The chosen set of molecules for the study is

The PCBs data set (with 209 compounds in the series) was the set of compounds choused for this study. The log(Kow) was the available data measured in same experimental conditions for 206 compounds [37-39].

HyperChem (licence v. 8.0/2007) was used for drawing and optimizing the compounds (using AMBER molecular mechanics model, POLAK-RIBIERE optimization algorithm, and AM1 method for semiempirical energy calculations). Molecular Descriptors Family [40-42] was used to create the population of structural descriptors that feed the genetic algorithm. The search was started for multiple linear regressions with four descriptors members of MDF relating the observed log(Kow) of 206 PCBs. Grubbs test [43] was used to identify outliers (PCB209) in the experimental; data.

Method: Search Space and Genetic Algorithm

Every gene codifies an operator used in construction of the chromosome of a molecular descriptor. Every descriptor (of a family of descriptors, such as MDF) is a genotype and all together is the genetic material of the family. Table 1 presents the search space created for MDF.

Table 1. MDF search space

Family	Gene	Genome																											
MDF	D_M	t	g																										
	A_P	C	H	M	E	G	Q																						
	I_D	D	d	O	o	P	p	Q	q	J	j	K	k	L	l	V	E	W	w	F	f	S	s	T	t				
	I_M	r	R	m	M	d	D																						
	F_C	m	M	D	P																								
	S_M	m	M	n	N	S	A	a	B	b	P	G	g	F	f	s	H	h	I	i									
	L_O	I	i	A	a	L	l																						

The working methodology of genetic algorithms suppose a initial prelevation (at random or using a strategy) of a ***sample*** of chromosomes from the genetic material - in this case a array of MDF members - from X_1 to X_p which enters in ***cultivar*** for conducting the ***evolution process***. The ***genetic algorithm*** operates on the sample which is changed (in part) in every ***generation***. Every set of `n` descriptors (where n is the multiplicity order for MLR) is a point in the ***search space*** and a ***possible solution***. The operators which change the genetic code are crossover and mutated. ***Crossover*** of two genotypes suppose chousing of a part from the stream of genes to be crossover (at random or using a strategy) and the values of the parts are switched one in the place of the other, and two descendents are produced. ***Mutation*** of a genotype supposes the changing of the value of a (or more) gene with other allowed value from the list of possible values for a given gene. Both crossover and mutation produces ***descendents***. The ***selection*** of the genotypes is the operation which mutation and crossover calls for, are based on a ***strategy*** and uses a score function (***selection score***). At least a part of the descendents are ***viable*** (descriptors), being able to be part of a ***viable solution*** (MLR) in the next generation(s). Viable descriptors replace a part from the sample through a ***survival*** process. As selection process, survival process uses a score function (***survival score***) and uses a ***strategy***. The evolution objective are recorded during evolution using a score function (***objective score***). The individuals which gives the best objective score in every generation (enters in the best MLR) are ***marked***. The marked individuals are automatically qualified for the next generation (no survival strategy applies on it). Not all individuals of a generation (including parents and descendants) survive and will be part of the next generation. This is done in order to keep constant the sample size (thus the number of replaced individuals is equal to the number of viable descendants).

Selection and survival based on selection and survival scores are applied through selection and survival strategies, using an ***algorithm***. ***PS*** algorithm constructs a ***proportional strategy*** using an array of scores and gives to an individual a chance (to be selected in selection process or to be killed in survival process) proportional with the score; it returns a given number N_Sel of individuals using their chances. ***DS*** algorithm constructs a ***deterministic strategy*** returning the N_Sel individuals with the first N_Sel highest scores (applies a random qualification at equal scores if it is necessary). ***TS*** algorithm constructs a ***tournament strategy*** using the array of scores and qualifies N_Sel individuals through a repeated N_Sel times tournament of two individuals.

The genetic algorithm acts as follows:

- The sample of the given size (N_Gen) is created (it contains predefined or random individuals);
- Repeat steps 1..6 until *objective score is satisfactory* or *a number of generations are exhausted*;
- Step_1: Computes selection, survival and objective scores (and eventually include in the next generation the marked individuals);
- Step_2: Select N_Cro pairs of individuals (using selection strategy);
- Step_3: For every one from 2×N_Cro, using p_Par (low) probability and a discrete uniform distribution pick up a number of N_Mut genes and make a mutation on it (parents); save the result

(whatever mutated or not, 2×N_Cro individuals);

- Step 4: For every one from N_Cro, using a discrete uniform distribution pick up the sequence of genes to be crossover, perform the crossover and save the results (replace the previous one, 2×N_Cro individuals);
- Step_5: For every one from 2×N_Cro, using p_Chi (low) probability and a discrete uniform distribution pick up a number of N_Mut genes and make a mutation on it (child); save the results (whatever mutated or not; replace the previous one, 2×N_Cro individuals);
- Step_6: Replace (using the survival strategy) a part of N_Gen with a part of 2×N_Cro;

Results and Discussion

A series of parameters influence the evolution and should be taken into account when a genetic algorithm is implemented. Table 2 presents a list of parameters that proved to be relevant.

Table 2. Configuration of the evolution: classes and parameters

Class	Parameter	Type (and values for lists)
Adaptation	a_v_ADAPT_Variance	Real
	ajb_ADAPT_JarqueBera	Real
	a_c_ADAPT_Correlation	Real
Sample	sn0_SAMPLE_Size	Integer (natural)
	rn0_REGRESSION_Multiple	Integer (natural)
	e1n_GENERATIONS_max	Integer (natural)
	g_r_GENERATIONS_first_rich	List: {Yes, No}
Reproduction	cn0_CROSSOVER_Pairs	Integer (natural)
	m_m_MUTATION_Genes	Integer (natural)
	mpp_MUTATION_Parent_probability	Real
	mcp_MUTATION_Child_probability	Real
Evolution	b_p_EVOLUTION_parameter	List: {r2, se, Mt, Hr}
	b_o_EVOLUTION_objective	List: {min, max}
Selection	sfs_SELECTION_strategy	List: {proportional, deterministic, tournament}
	sfn_SELECTION_normalized	List: {Yes, No}
	sfr_SELECTION_ranks	List: {Yes, No}
	sfa_SELECTION_accuracy	Integer (natural)
	sff_SELECTION_function	List: {nalive, r2_min, se_min, Mt_min, Hr_min, r2_max, se_max, Mt_max, Hr_max, r2_avg, se_avg, Mt_avg, Hr_avg}
	sfo_SELECTION_objective	{min, max}
	fr2_SELECTION_r2_p	Real
	fse_SELECTION_se_p	Real
	fMt_SELCTION_Mt_p	Real
	fHr_SELECTION_Hr_p	Real
Survival	v_p_SURVIVAL_phenotyping_p	Real
	v_g_SURVIVAL__genotyping_p	Real
	vfs_SURVIVAL_strategy	List: {proportional, deterministic, tournament}
	vfr_SURVIVAL_ranks	List: {Yes, No}
Execution	e0n_RUNS_number	Integer (natural)
	b_k_RUNS_kepp_best_in_sample	List: {Yes, No}
	b_f_RUNS_get_best_from_file	List: {Yes, No} + list of genotypes in a file named c_galg.txt

- *Adaptation parameters* - in a series of three - refers the environment of the evolution in terms of viability of the phenotypes in the environment; these parameters defines limits - Variance and Correlation defines minimal values, JarqueBera defines maximal value - required to phenotypes to be considered alive in the environment.

- *Sample parameters* - in a series of four refer the sample and their cultivar - Sample Size is the 'room' in the environment reserved for breeding; Regression Multiple is the number of phenotypes which participates as independent variables in the MLR; Generations Max is the maximum number of generations in which the GA is run in the process of MLR optimization; Generations First Rich is a choice to repeat the random selection of the initial generation genotypes until all genotypes has at least one alive phenotype in the environment.
- *Reproduction parameters* - in a series of four - defines the way in which reproduction are made in a generation; Crossover Pairs is the number of pairs of genotypes which are crossover; Mutation Genes is the number of genes which are mutated when mutation occurs; Mutation Parent Probability controls when mutation occurs before crossover (at parents) and Mutation Child Probability controls when mutation occurs after crossover (at child).
- *Evolution parameters* - two parameters - define the objective function (Evolution Parameter) and its objective (Evolution Objective).
- *Selection parameters* - in a series of 10 - defines the strategy of selection (Selection Strategy), if the values are normalized between generations (Selection Normalized), if are used their ranks in place of their values (Selection Ranks), an integer defining a factor with which the selection scores is multiplied and rounded to integers (to create a discrete space of probabilities from scores required especially for Proportional Selection), a series of alternatives for selection function (Selection Function), the objective of a given alternative (Selection Objective), a power at which the result of the function to be rising up - rescales the values in terms of probabilities of selection (here four values should be specified, but just one gives the expression used at selection function - this one correspond to the defined alternative).
- *Survival parameters* - in a series of four - defines the strategy for replacing the genotypes (Survival Strategy), if are used ranks in places of values of the survival function (Survival Ranks) and two parameters defining the weight of the genotype similarity (Survival Genotyping) and phenotype similarity (Survival Phenotyping) in the expression of the global similarity score of two individuals.
- *Execution parameters* - in a series of thee - define the behavior in execution of the evolutionary program; Runs Number is the number of independent runs of the genetic algorithm restarting the search from a initial generation; Runs Keep Best In Sample defines if genotypes giving best MLR model in a generation enters in the survival process or are directly qualified for the next generation; Runs Get Best From File are a option useful for repeated optimization procedures in order to improve a previous obtained MLR.

Table 3 shows the parameters which were used in the optimization of a MLR with four MDF descriptors acting as QSAR on log K_{ow} of a series of 206 PCBs.

Only a part of the genotypes from the entire population were adapted to the environment the rules for viability defined by the values from Table 3 were applied. Thus, almost 46% (60337/131328) of the genotypes provided at least one adapted phenotype. About 20% (26355&26324/131328) of the phenotypes created by *Identity* and *Absolute* linearization operators (first and respectively third entries in Genome at L_O gene in Table 1), almost 18% (24085) of *logarithm* (`l` entry in same table), and almost 23% of *Logarithm* of absolute (`L` entry, 29973 adapted phenotypes), *inverse* (`i`, 30178) and *absolute* inversed (`a`, 30174) were adapted.

The average value of the determination coefficients obtained in 46 runs (Table 3) for nine pairs of selection and survival strategies (Table 5) was of 0.8808 (414 observations) with a range between 0.8804 and 0.8811 at a 5% risk of being in error.

The evolution was observed in first generation in 28.5% (118/414) of the cases when the determination was improved to 0.8816 ranging between 0.8810 and 0.8823 at a 5% risk of being in error. The determination coefficient obtained after evolution in first generation was statistically significant different from the determination coefficient obtained in the first generation (the risk of being in error was about 3%) without costs in variability (F ratio comparing variances is 1.1, probability to observe at random such event is 52%).

A list of parameters used to set up the output was defined (Table 4) in order to have detailed observation of the evolution.

Table 3. Configuration of the evolution

Parameter	Value
a_v_ADAPT_Variance	0.1
ajb_ADAPT_JarqueBera	1.0
a_c_ADAPT_Correlation	0.1
sn0_SAMPLE_Size	12
rn0_REGRESSION_Multiple	4
e1n_GENERATIONS_max	20000
g_r_GENERATIONS_first_rich	Yes
cn0_CROSSOVER_Pairs	2
m_m_MUTATION_Genes	2
mpp_MUTATION_Parent_probability	5%
mcp_MUTATION_Child_probability	5%
b_p_SELECTION_parameter	r2
b_o_SELECTION_objective	max
sfn_FITNESS_normalized	No
sfr_FITNESS_ranks	No
sfa_FITNESS_accuracy	10000
sff_FITTEST_function	r2_min
sfo_FITTEST_objective	max
fr2_FITTEST_r2_p	1.0
fse_FITTEST_se_p	1.0
fMt_FITTEST_Mt_p	1.0
fHr_FITTEST_Hr_p	1.0
v_p_SURVIVAL_phenotyping_p	1.0
v_g_SURVIVAL__genotyping_p	1.0
vfr_SURVIVAL_ranks	No
e0n_RUNS_number	46
b_k_RUNS_kepp_best_in_sample	Yes
b_f_RUNS_get_best_from_file	No

Table 4. Configuring the output

Parameter	Value
d_d_SHOW_descriptive_XX (XX = m0, m1, m2, m3, m4, mx, my, v0, g1, g2, jb, r1, r2)	Yes/No
d_f_SHOW_fitness_YY (YY = nalive, r2_min, se_min, Mt_min, Hr_min, r2_max, se_max, Mt_max, Hr_max, r2_avg, se_avg, Mt_avg, Hr_avg)	Yes/No
d_s_SHOW_genotypes	No/Yes
d_p_SHOW_phenotypes	No/Yes
d_m_SHOW_mols	No/Yes
d_g_SHOW_generations	No/Yes
d_e_SHOW_evolutions	Yes/No
d_c_SHOW_configuration	Yes/No

Table 5. Selection and survival strategies in genetic algorithms: run 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

The parameters from Table 4 allow the user to configure its output including or excluding some of them. The meanings of the parameters are as follows:

- The generic parameters *d_d_SHOW_descriptive_XX* and *d_f_SHOW_fitness_YY* are lists (XX = m0, m1, m2, m3, m4, mx, my, v0, g1, g2, jb, r1, r2; YY = nalive, r2_min, se_min, Mt_min, Hr_min, r2_max, se_max, Mt_max, Hr_max, r2_avg, se_avg, Mt_avg, Hr_avg), every item in the list defining an observable (and Supplementary Material (online));
- *d_s_SHOW_genotypes* add a number of columns equal with sample size and containing the genotypes which gives at least an adapted phenotype;
- *d_p_SHOW_phenotypes* add six times of sample size columns in the output file containing the adapted phenotypes;
- *d_m_SHOW_mols* add extra columns in `_cfg.txt`, `_evo.txt`, and `_gen.txt` output files containing the predictions based on the best available MLR;
- *d_g_SHOW_genarations* set to *Yes* produces the `_gen.txt` file containing one row per generation;
- *d_e_SHOW_evolution* set to *Yes* produces the `_cfg.txt` file containing only the results when an evolution occurs;
- *d_c_SHOW_configuration* set to *Yes* produces the `_cfg.txt` file containing only the results after entire cycles of evolution;

The created evolutionary program was used to show the influence of different selection and survival strategies on evolution supervised by the genetic algorithm when was feed with data for structure-activity relationships optimization in the PCB series of biologically active compounds were recorded.

Two parameters (*sfs_FITNESS_strategy* and *vfs_SURVIVAL_strategy* - see Table 2) took different values once at the time for the parameters kept constant (the above table), nine executions of the program were independently started, and the results were recorded in separate files (two files per execution, Table 4).

The following information was available in the *PCB_XXX_evo.txt* files:

- e1i: generation (integer) in which a new evolution occurred (evolution refers the value of the objective function r2 → max.; evolution meant the obtaining of a QSAR with a determination coefficient higher than the one previously obtained);
- sn2: number of viable genotypes (not always all genotypes from crossovers and mutations are viable - has at least one adapted phenotype; maximum number of viable genotypes is the sample size sn0 - see Table 2);
- pn2: number of adapted phenotypes; maximum number of adapted phenotypes is six times of sn2);
- nrali: number of phenotypes associations (regressions) obtained with individuals from cultivar of the sample; phenotypes associations are obtained combining rn0 - see Table 2 - distinct phenotypes and the same adaptation requirements are for the association as are the requirements for a phenotype;
- Gen11..Gen0: the genotype in position from 12 down to 1 in the places of the sample in a given moment of evolution;
- rr2: value of the objective function of evolution (see Table 3); the determination coefficient of the best available SAR expressed as MLR with 4 MDF descriptors existing in the sample;
- rse: the value of the (alternative) objective function "estimation errors sum";
- rMt: the value of the (alternative) objective function "mean of Student t values of the regression parameters";
- rHr: the value of the (alternative) objective function "the entropy of the determination";
- rdr: number of coefficients in the MLR giving the best available MLR; rdr=rn0+1 if the free term are present (is not null at 5% risk being in error) and rdr=rn0 otherwise;
- rme: mean squared error (MSE) of the regression that provided the best available MLR;
- rdt: number of degrees of freedom in the best available MLR (nmo - rdr);
- rm0: the absolute value of the mean of estimate, $M(|\hat{Y}|)$;
- rm1: mean of the estimate, $M(\hat{Y})$;
- rm2: the mean $M((Y-\hat{Y})^2)$;
- rm3: the mean $M((Y-\hat{Y})^3)$;
- rm4: the mean $M((Y-\hat{Y})^4)$;

- rmx: the mean $M(\hat{Y}^2)$;
- rmy: the mean $M(Y\hat{Y})$;
- rv0: $rm2/rm0^2$;
- rg1: skewness (g_1) of the estimated ($\hat{Y}$);
- rg2: kurtosis excess (g_2) of the estimated ($\hat{Y}$);
- rjb: the Jarque-Bera (JB) statistic of the estimated ($\hat{Y}$);
- rr1: $r(Y,\hat{Y})$;
- rr2: $r^2(Y,\hat{Y})$;
- rnalive: 0 (when $\hat{Y}$ are not adapted) or 1 (when $\hat{Y}$ are adapted);
- rr2_min: value "r2_min" of the selection score FS; are given by $\hat{Y}$ to the phenotypes from which are composed (min, max and avg are identical here, referring the min, average and max from a single value); rse_min: value of the alternative "se_min" of the selection score FS; rMt_min: value of the alternative "Mt_min" of the selection score FS; rHr_min: value of the alternative "Hr_min" of the selection score FS;
- rr2_max: idem rr2_min; rse_max: idem rse_min; rMt_max: idem rMt_min; rHr_max: idem rHr_min; rr2_avg: idem rr2_min; rse_avg: idem rse_min; rMt_avg: idem rMt_min; rHr_avg: idem rHr_min;
- gnalive: in the generations which produces evolution (and only these are listed in this file and in the defined configuration of the execution) are identical with nrali; nrali keeps the number of obtained regressions with individuals from last (including the current) generation producing evolution and gnalive refers the current generation (without regarding to evolution);
- gr2_min: value "r2_min" of selection score FS given by the descriptors existing in sample in the generation; gse_min: value of "se_min" alternative; gMt_min: value of "Mt_min" alternative;
- gHr_min: value of "Hr_min" alternative; gr2_max: value of "r2_max" alternative; gse_max: value of "se_max" alternative; gMt_max: value of "Mt_max" alternative; gHr_max: value of "Hr_max" alternative; gr2_avg: value of "r2_avg" alternative; gse_avg: value of "se_avg" alternative; gMt_avg: value of "Mt_avg" alternative; gHr_avg: value of "Hr_avg" alternative;
- regression_equation: best MLR obtained with descriptors with the genotype in the sample;

The following information was available in the *PCB_XXX_cfg.txt* files:

- Configuration part:
  - My_Data=172.27.211.5/MDFSARs[PCB_lkow_data;PCB_lkow_tmpx] - IP of the server, database, tables with input data;
  - Genomes=Genes:[mp;fc;oi;id;ap;dm]/Addre:[fc;ap;id;oi;dm;mp] - definition of the genome and the calculation of the address in the table (a permutation of the definition, used to access fast the data in the `_tmpx` table);
  - G_Codes=mMnNSPsAaBbGgFfHhIi/mMDP/RrMmDd/DdOoPpQqJjKkLlVEWwFfSsTt/CHMEGQ/gt - list of the codes for every gene of the genome;
  - N_Sizes=Sample:12/CrossO:2/RegreM:4/MutGen:2 - sample size; number of crossovers; number of descriptors in MLR; number of genes to be mutated when mutation occurs;
  - Adapted=AbsVariance: 1.6353599751539131E-0003(0.100)/Jarque-Bera: 7.5766145463847375E+0000(1.000)/Determinate:0.100(%)/StudentTVal:[1.9721423227715891E+0000(201); 1.9720824746250930E+0000(202)] - v0 (correspondent of rv0 din PCB_XXXX_evo.txt for the observed (Y); jb (correspondent of rjb din PCB_XXXX_evo.txt for the observed (Y); lowest determination coefficient to be adapted; Student t from Student distribution corresponding to a model of regression with free term (df=nmol-RegM-1) and without free term (df=nmol-RegM) used as limits in the two cases to accept or reject the model;
  - Mutated=ParentsProb:0.050/ChildreProb:0.050 - probability (in percents) to occur mutation before crossover (parents are then mutated); probability (in percents) to occur mutation after crossover (children are then mutated);
  - Objectiv=Parameter:r2/To:max - the name of the objective function chosed to be used in the evolution process; the objective of the objective function;

  - Selected=Parameter:r2_min/To:max/Using:tournament/Normalized?:No/SortedRank?:No/First_Rich?:Yes - which is the observable in selection; which is the selection objective; which is the strategy of selection; if the strategy is applied on normalized (between generations) values; if ranks replaces the values when the strategy are applied; first generation is rich in adapted genotypes (all adapted) or not;
  - Survival=Phenotyping:1.000/Genotyping:1.000/Using:tournament/SortedRank?:No - the value of the weight of the phenotypic similarity in the survival score; the value of the weight of the genotypic similarity in the survival score; the strategy of the survival; if ranks replaces the values when the strategy are applied;
  - Evolution=nalive;r2_min;se_min;Mt_min;Hr_min;r2_max;se_max;Mt_max;Hr_max;r2_avg;se_avg;Mt_avg;Hr_avg; - parameters recorded during evolution;
  - Alive=Genotypes:60337/Phenotypes(I):26355;/Phenotypes(A):26324;/Phenotypes(l):24085;/Phenotypes(L):29973;/Phenotypes(i):30178;/Phenotypes(a):30174; - total number of adapted in the entire population: genotypes, and phenotypes by linearization operator (see Table 1);
  - Repeated=Sampling:45 - number of repetitions is one unit more (here are the lower limit for construction of a statistical significance);
- Repetition: number of repetition in the independent run of the GA starting from a random initial sample;
- Found: number of the generation in current repetition when a evolution occurred;
- All other parameters corresponds to the same ones from PCB_XXXX_evo.txt file: r2=rr2, se=rse, Mt=rMt, Hr=rHr, dr=rdr, me=rme, dt=rdt, m0=rm0, m1=rm1, m2=rm2, m3=rm3, m4=rm4, mx=rmx, my=rmy, v0=rv0, g1=rg1, g2=rg2, jb=rjb, r1=rr1, r2=rr2, nalive=rnalive, r2_min=rr2_min, se_min=rse_min, Mt_min=rMt_min, Hr_min=rHr_min, r2_max=rr2_max, se_max=rse_max, Mt_max=rMt_max, Hr_max=rHr_max, r2_avg=rr2_avg, se_avg=rse_avg, Mt_avg=rMt_avg, Hr_avg=rHr_avg, regression_equation= regression_equation).

Conclusions

A genetic algorithm for identification the most close to the optimum multiple linear regression models was developed and implemented to be use in identification of quantitative structure-activity relationship models. The genetic algorithm was designed to be feed with families of molecular descriptors, and was tested on the Molecular Descriptors Family.

The objective of the genetic algorithm was to optimize the multiple linear regressions with four descriptors for prediction of octanol-water partition coefficient (expressed in logarithmic scale) of a series of 206 polychlorinated biphenyls and was successfully accomplished.

The average of the determination coefficients obtained in 46 runs for nine pairs of selection and survival strategies was of 0.8808 (414 observations) with a range between 0.8804 and 0.8811 (at a risk to be in error of 5%).

The evolution was observed in first generation in 28.5% (118/414) of the cases when the determination was improved to 0.8816 (with a range from 0.8810 to 0.8823, at a 5% risk to be in error). The determination coefficient obtained after evolution in first generation proved to be statistically significant different from the determination coefficient obtained in the initial generation (for a level of significance of 3%) without costs in variability (F ratio comparing variances was 1.1, probability to observe at random such event being 52%).

Conflict of Interest

The author declares that they have no conflict of interest.

Acknowledgements

Financial support is gratefully acknowledged to CNCSIS-UEFISCSU Romania (project PNII-IDEI1051/202/2007).

References

1. Hammett LP. Some Relations between Reaction Rates and Equilibrium Constants. Chemical Reviews 1935;17:125-136.
2. Crum-Brown A, Fraser TR. On the Connection between Chemical Constitution and Physiological Action. Part I. On the Physiological Action of the Salts of the Ammonium Bases, derived from Strychnia, Brucia,Thebaia, Codeia, Morphia, and Nicotia. Philosophical Transactions of the Royal Society of London 1868;25:151-203.
3. Hansch C, Leo A. Substituent Constants for Correlation Analysis in Chemistry and Biology. New York: John Wiley & Sons, 1979.
4. Diudea M, Gutman I, Jäntschi L. Molecular Topology. New York: Nova Science, 2001.
5. Weismann FLA. The Germ-Plasm. A theory of heredity, New York: Charles Scribner's Sons, 1893.
6. Lamarck JBPAM. Philosophie zoologique, ou Exposition des considérations relatives à l'histoire naturelle des animaux; à la diversité de leur organisation et des facultés qu'ils on obtiennent; aux causes physiques qui maintiennent en eux la vie et donnent lieu aux mouvemens qu'ils exécutent; enfin, à celles qui produisent les unes le sentiment et les autres l'intelligence de ceux qui en sont douse, Paris: Dentu; 2 volumes, 1809.
7. Darwin CR. The origin of species by means of natural selection or the preservation of favoured races in the struggle for life, London: John Murray, 1859.
8. Mendel JG. Versuche über Plflanzenhybriden Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr, Abhandlungen:3-47(english translation: Druery CT & Bateson W, 1901. Experiments in plant hybridization, Journal of the Royal Horticultural Society 1866;26:1-32.
9. Morgan TH, Sturtevant AH, Muller HJ, Bridges CB. The mechanism of Mendelian heredity, New York: Henry Holt & Co, 1915.
10. Fisher RA. Retrospect of the Criticisms of the Theory of Natural Selection, In: Evolution as a Process (Huxley JS, Hardy AC, Ford EB), London: Allen&Unwin, 1954p. 84-98.
11. Ayala FJ, Escalante A, O'Hugin C, Klein J. Molecular genetics of speciation and human origins. Proc Natl Acad Sci USA 1994;91(15):6787-6794.
12. Barricelli N. Esempi numerici di processi di evoluzione. Methodos 1954;1954:45-68.
13. Fraser AS. Simulation of genetic systems by automatic digital computers. I. Introduction. Australian J Biol Sci 1957;10:484-491.
14. Fraser AS. Simulation of genetic systems by automatic digital computers. II. Effects on linkage on rates of advance under selection. Australian J Biol Sci 1957;10:492-499.
15. Barker JSF. Simulation of genetic systems by automatic digital computers. III. Selection between alleles at an autosomal locus. Australian J Biol Sci 1958;11:603-612.
16. Barker JSF. Simulation of genetic systems by automatic digital computers. IV. Selection between alleles at a sex-linked locus Australian J Biological Sciences 1958;11:613-625.
17. Fraser AS. Simulation of genetic systems by automatic digital computers. V. Linkage, dominance, and epistasis, In: Biometrical Genetics (Kempthorne O). New York: Pergamon Press, 1960, pp. 70-83.
18. Fraser AS. Simulation of genetic systems by automatic digital computers. VI. Epistasis. Australian J Biol Sci 1960;13(2):150-162.

19. Fraser AS. Simulation of genetic systems by automatic digital computers. VII. Effects of reproductive rate and intensity of selection on genetic structure. Australian J Biol Sci 1960;13(3):344-350.
20. Fraser AS. Simulation of genetic systems VIII. J Theor Biol 1962;2(3):329-346.
21. Fraser AS, Hansche PE. Simulation of genetic systems. IX. Major and minor loci. In: Genetics Today, Proc. XI International Congress of Genetics (Geerts SJ), Oxford: Pergamon Press, 1965;3:507-516.
22. Fraser AS, Burnell D, Miller D. Simulation of genetic systems. X. Inversion polymorphism. J Theor Biol 1966;13:1-14.
23. Fraser AS, Burnell D. Simulation of genetic systems. XI. Inversion polymorphism. Amer J Human Genetics 1967;19:270-287.
24. Fraser A, Burnell D. Simulation of genetic systems. XII. Models of inversion polymorphism. Genetics 1967;57:267-282.
25. Fraser A, Burnell D. Computer Models in Genetics. New York: McGraw-Hill, 1970.
26. Martin WN, Spears WM. Foundations of genetic algorithms 6. San Francisco: Morgan Kaufmann, 2001.
27. Matthews KB. Applying Genetic Algorithms to Multi-objective Land-Use Planning. PhD Thesis (Agriculture) - Supervisor Prof. Kraw S. Robert Gordon University, Craigiebuckler, Aberdeen, NSW, Australia, 2001.
28. Osman NY. The Development of a Predictive Damage Condition Model of Light Structures on Expansive Soils using Hybrid Artificial Intelligence Techniques. PhD Thesis (Engineering and Industrial Sciences) - Supervisor Prof. McManus KAM. Swinburne University of Technology, Melbourne, VIC, Australia, 2007.
29. Veith TL. Agricultural BMP Placement for Cost-Effective Pollution Control at the Watershed Level. PhD Thesis (Biological Systems Engineering) - Supervisor Prof. Wolfe ML. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2002.
30. Dai B. Simulations-guided design of process analytical sensor using molecular factor computing. PhD Thesis (Chemistry) - Supervisor Prof. Lodder RA. University of Kentucky, Lexington, KY, USA, 2007.
31. Aickelin U. Genetic Algorithms for Multiple-Choice Optimisation Problems. PhD Thesis (European Business Management) - Supervisor Prof. Dowsland K. University of Wales Swansea, Swansea, UK, 1999.
32. Sastry KM. Genetic Algorithms and Genetic Programming for Multiscale Modeling: Applications in Materials Science and Chemistry and Advances in Scalability. PhD Thesis (Systems and Entrepreneurial Engineering) - Supervisor Prof. Goldberg DE & Johnson DD. University of Illinois at Urbana-Champaign, Urbana, IL, USA, 2007.
33. Gantovnik VB. An Improved Genetic Algorithm for the Optimization of Composite Structures. PhD Thesis (Engineering Mechanics) - Supervisor Prof. Gürdal Z. Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 2005.
34. Tufail M. Optimal water quality management strategies for urban watersheds using macro-level simulation models linked with evolutionary algorithms. PhD Thesis (Civil Engineering) - Supervisor Prof. Ormsbee LE. University of Kentucky, Lexington, KY, USA, 2006.
35. Zwickl DJ. Genetic Algorithm Approaches for the Phylogenetic Analysis of Large Biological Sequence Datasets Under the Maximum Likelihood Criterion. PhD Thesis (Biology) - Supervisor Prof. Hills DM. University of Texas at Austin, Austin, TX, USA, 2006.
36. Suzuki H. Evolution of a Novel Function Fascilitated by Genetic Recombination in Genetic Algorithms. PhD Thesis (Biology) - Supervisor Prof. Iwasa Y. Kyushu University, Fukuoka, Fukuoka, Japan, 1998.
37. Eisler R, Belisle AA. Planar PCB Hazards to Fish, Wildlife, and Invertebrates: A Synoptic Review. Contaminant Hazard Reviews. Biological Report 31. [online] 1996. [Accessed March 2010]. Available at: URL: http://www.pwrc.usgs.gov/infobase/eisler/chr_31_planar_pcbs.pdf
38. Mullins MD, Pochini CM, McCrindle S, Romkes M, Safe SH, Safe LM. High resolution PCB analysis: synthesis and chromatographic properties of all 209 PCB congeners. Environ Sci Technol 1984;18:468-476.

39. Jäntschi L, Bolboacă SD, Diudea MV. Chromatographic Retention Times of Polychlorinated Biphenyls: from Structural Information to Property Characterization. Int J Mol Sci 2007;8(11): 1125-1157.
40. Jäntschi L. MDF - A New QSAR/QSPR Molecular Descriptors Family. Leonardo Journal of Sciences 2004;3(4):68-85.
41. Jäntschi L. Molecular Descriptors Family on Structure Activity Relationships 1. Review of the Methodology. Leonardo Electronic Journal of Practices and Technologies 2005;6:76-98.
42. Jäntschi L, Bolboacă SD. Results from the Use of Molecular Descriptors Family on Structure Property/Activity Relationships. International Journal of Molecular Sciences 2007;8(3):189-203.
43. Grubbs F. Procedures for Detecting Outlying Observations in Samples. Technometrics 1969;11(1):1-21.