Replacement Method and Enhanced Replacement Method Versus the Genetic Algorithm Approach for the Selection of Molecular Descriptors in QSPR/QSAR Theories

Andrew G. Mercader,[†,‡] Pablo R. Duchowicz,[*,†] Francisco M. Fernández,[†] and Eduardo A. Castro[†]

Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (INIFTA, UNLP, CCT La Plata-CONICET), Diag. 113 y 64, Sucursal 4, C.C. 16, 1900 La Plata, Argentina, and PRALIB (UBA-CONICET), Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Junín 956, C1113AAD Buenos Aires, Argentina

Received March 15, 2010

We compare three methods for the selection of optimal subsets of molecular descriptors from a much greater pool of such regression variables. On the one hand is our enhanced replacement method (ERM) and on the other is the simpler replacement method (RM) and the genetic algorithm (GA). These methods avoid the impracticable full search for optimal variables in large sets of molecular descriptors. Present results for 10 different experimental databases suggest that the ERM is clearly preferable to the GA that is slightly better than the RM. However, the latter approach requires the smallest amount of linear regressions and, consequently, the lowest computation time.

INTRODUCTION

A generally accepted solution for overcoming the lack of experimental data in complex chemical phenomena is the analysis based on quantitative structure−property/−activity relationships (QSPR/QSAR).[1] For that reason, there is great interest in the development of such kind of predictive techniques.[2−6] The ultimate role of the QSPR/QSAR theory is to develop mathematical models for the estimation of relevant properties and activities of chemical and biological interest, especially when, for some reason, they cannot be experimentally determined. Such studies rely on the basic assumption that the molecular structure of a compound determines entirely its properties or activities. The structure is commonly translated into the so-called molecular descriptors that are calculated through mathematical formulas obtained from several theories, such as chemical graph theory, information theory, quantum mechanics, etc.[7−11]

At the present time, there are thousands of descriptors available in the literature, and one has to select those that characterize the property or activity under consideration in the most reliable and efficient way. One is thus faced with the mathematical problem of selecting a subset $\mathbf{d}$ of d descriptors from a much larger set $\mathbf{D}$ with D ones, where $d \ll D$. The search for the optimal set of descriptors may be guided by the minimization or maximization of a chosen function; for example, we may be interested in a model that makes the standard deviation (S) as small as possible. In this case we look for the global minimum of $S(\mathbf{d})$, where $\mathbf{d}$ is a point in a space of $D!/(D-d)!d!$. Consequently, a full search (FS) of the optimal variables is impractical because it requires $D!/(D-d)!d!$ linear regressions.

We consider that the linear algorithms are most convenient for analyzing QSPR/QSAR data sets for to two main reasons: (i) They exhibit a higher predictive capability and perform more efficiently on external test sets not considered during the model calibration; and (ii) When few experimental observations are available, it is necessary to employ the lowest number of optimized parameters during the model development, a condition that linear models fulfill. In addition, some recent studies have applied the Gram−Schmidt procedure in the recent proposed spectral-structure activity relationship (S-SAR) method.[12] The advantage of this technique is that it enables us to replace the classic multivariable linear regression analysis by purely algebraic models with some conceptual and computational advantages, having both ecotoxicological, environmental, and anticancer bioactivity applications.[12−15]

Some time ago we proposed the replacement method (RM)[16−18] and later the enhanced replacement method (ERM)[19] that produced linear regression QSPR/QSAR models that were quite close, the FS ones with much less computational work. Both those techniques approach the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a subset of d descriptors. The RM produced models with better statistical parameters than the forward stepwise regression procedure[20] and variants of the more elaborated genetic algorithms.[21] The ERM lead to similar or even better statistical parameters although with slightly more computational work.[22]

The RM is a rapidly convergent iterative algorithm that produces linear regression models with small S in remarkably short computer times.[23−26] However, in some difficult cases, the RM can get trapped in a local minimum of S that is not able to leave without some kind of additional constraint. Although such local minima provided quite acceptable models for QSPR-QSAR, as shown in all earlier applications of the RM, there is still room for improvement, and for that reason, we developed the ERM.

The ERM follows the same philosophy but is less likely caught into local minima as well as less dependent on the

[*] Corresponding author. E-mail: pabloducho@gmail.com or prduchowicz@yahoo.com.ar. Telephone: (+54)(221)425-7430/(+54)(221)425-7291.

[†] Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas.

[‡] PRALIB Universidad de Buenos Aires.

10.1021/ci100103r
Published on Web 08/19/2010

initial solution. It has a close resemblance with the simulated annealing (SA), which is an adaptation of the Metropolis−Hastings algorithm, a Monte Carlo method,[27] and generates sample states of a thermodynamic system. The name and inspiration has come from annealing in metallurgy, a technique involving heating and controlled cooling of a material to increase the size of its crystals and reduce their defects. The heat causes the atoms to become unstuck from their initial positions (a local minimum of the internal energy) and wander randomly through states of higher energy; the slow cooling gives them more chances of finding configurations with lower internal energy than the initial one.[28]

Earlier comparison of ERM, RM, and GA[22,29,30] were based on only one case at the time: a given number of descriptors in the model and a specific data set. The main purpose of this paper is to provide a wider comparison using several data sets and numbers of descriptors in order to draw more satisfactory conclusions.

METHODS

The following subsections briefly describe the theory of RM, ERM, and GA as variable subset selection methods.

Replacement Method. In this case we choose an optimal subset $\mathbf{d}_m = \{X_{m1}, X_{m2}, ..., X_{md}\}$ of d descriptors from a large set $\mathbf{D} = \{X_1, X_2, ..., X_D\}$ of D ones ($d \ll D$) provided by some available commercial program, with a minimum standard deviation (S):

$$S = \frac{1}{(N - d - 1)} \sum_{i=1}^{N} \mathrm{res}_i^2 \qquad (1)$$

where N is the number of molecules in the training set, and res_i is the residual for molecule i (difference between the experimental and predicted property). Notice that $S(\mathbf{d}_n)$ is a distribution on a discrete space of $D!/d!(D - d)!$ disordered points $\mathbf{d}_n$. The full search (FS) that consists of calculating $S(\mathbf{d}_n)$ on all those points always enables us to arrive at the global minimum of S, but it is computationally prohibitive if D is sufficiently large. The RM consists of the following steps:

(i) We choose an initial set of descriptors $\mathbf{d}_k$ at random, replace one of the descriptors, say X_{ki}, with all the remaining $D - d$ descriptors, one by one, and keep the set with the smallest value of S. This is one step of the procedure.

(ii) In the resulting set, we choose the descriptor with the greatest standard deviation in its coefficient and substitute all the remaining $D - d$ descriptors, one by one, for it. We repeat this procedure until the set remains unmodified. In each cycle we do not modify the descriptor optimized in the previous one. Thus, we obtain the candidate $\mathbf{d}_m^{(i)}$ that came from the so-constructed path i. It is worth noting that if the replacement of the descriptor with the largest error by those in the pool does not decrease the value of S, then we do not change that descriptor.

(iii) We carry out the process above for all the possible paths $i = 1, 2, ..., d$ and keep the point $\mathbf{d}_m$ with the smallest standard deviation: $\min_i S(\mathbf{d}_m^{(i)})$.

Enhanced Replacement Method. The ERM is a three-step combination of two algorithms: first the RM already described above, then a modified RM (MRM), and finally the RM again. The MRM follows the RM strategy except that in each step we substitute the descriptor with the largest error even if that substitution is not accompanied by a smaller value of S (we chose the next smallest value of S). Thus the MRM adds some sort of noise that prevents the searching process from staying in a local minimum of S.[19]

Genetic Algorithm. The GA is a search technique based on natural evolution principles where variables play the role of genes (in this case a set of descriptors) in an individual of the species. An initial group of random individuals (population) evolves according to a fitness function (in this case the standard deviation) that determines the survival of the individuals. The algorithm searches for those individuals that lead to better values of the fitness function through selection, mutation, and crossover genetic operations. The selection operators guarantee the propagation of individuals with better fitness in future populations. The GAs explore the solution space combining genes from two individuals (parents) by using the crossover operator to form two new individuals (children) and also by randomly mutating individuals using the mutation operator. The GAs offer a combination of hill-climbing ability (natural selection) and a stochastic method (crossover and mutation) and explore many solutions in parallel, processing information in a very efficient manner. The practical application of GAs requires the tuning of some parameters, such as population size, generation gap, crossover rate, and mutation rate. These parameters typically interact among themselves nonlinearly and cannot be optimized one at a time. There has been considerable discussion about parameter settings and approaches to parameter adaptation in the evolutionary computation literature; however, there does not seem to be conclusive results on which may be the best ones.[31]

MATERIALS

Experimental Data. Ten different experimental data sets that had previously been analyzed were used to test and contrast the performance of ERM, RM, and GA. These data sets were: a fluorophilicity data set (FLUOR), consisting of 116 organic compounds characterized by 1268 theoretical descriptors;[32] a growth inhibition data set (GI), with growth inhibition values to the ciliated protozoan *Tetrahymena pyriformis* by 200 mechanistically diverse phenolic compounds and 1338 structural descriptors;[33] a GABA receptor data set, containing 78 inhibition data for flavone derivatives and 1187 molecular descriptors;[34] 100 pED_{50} antiepileptic activities of enaminones with 1306 descriptors (MES);[35] 166 aqueous solubilities of drug-like compounds with 1497 descriptors (SOL);[24] 470 pIGC_{50} aqueous toxicities of heterogeneous aliphatic compounds with 1505 descriptors (TOX1);[36] 392 pIGC_{50} aqueous toxicities of benzene derivatives with 1497 descriptors (TOX2);[37] 17 acetylcholinesterase inhibitor activities of substituted indanone and benzylpiperidine analogs using 300 descriptors (ACET);[38] 35 glass transition temperatures of structurally diverse polymers and 442 descriptors (GTT);[38] and finally 30 melt transition temperatures of structurally diverse polymers and 368 descriptors (MTT).[38]

Calculation of Molecular Descriptors. In the first seven data sets, the structures of the compounds were first preoptimized with the molecular mechanics force field (MM+) procedure included in Hyperchem version 6.03,[39]

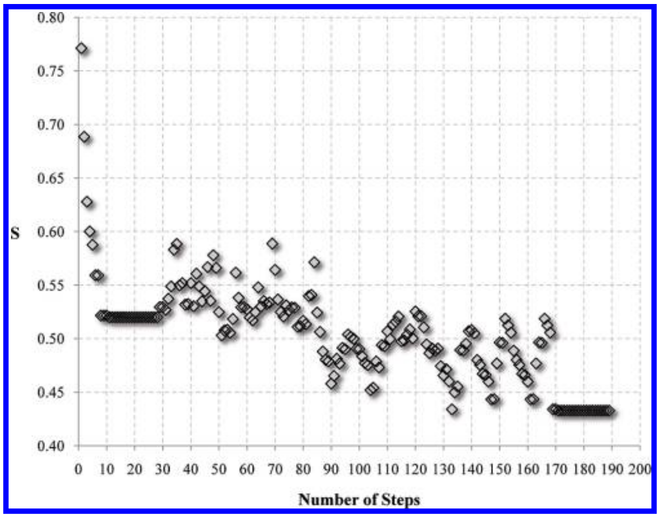

Figure 1. Standard deviation vs number of steps for the ERM.

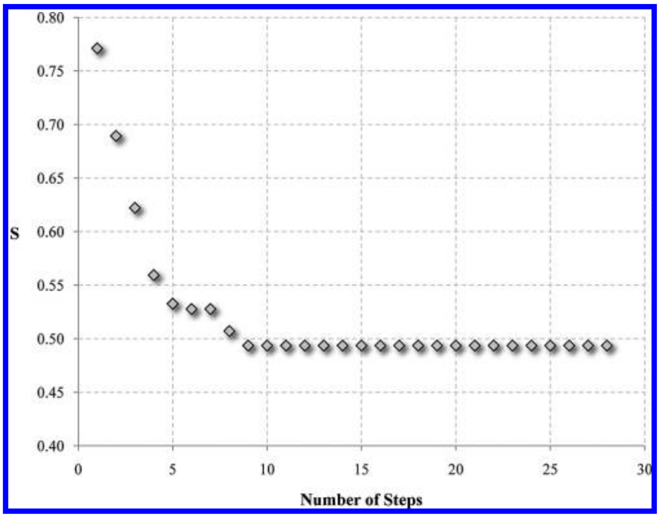

Figure 2. Standard deviation vs number of steps for the RM.

and the resulting geometries were further refined by means of the semiempirical molecular orbitals theory parametric method-3 (PM3) method using the Polak−Ribiere algorithm and a gradient norm limit of 0.01 kcal $Å^{-1}$. For each database, more than a thousand molecular descriptors were calculated using the software Dragon,[40] including parameters of all types, such as constitutional, topological, geometrical, charge, GETAWAY (Geometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk counts, BCUT descriptors, 2D-autocorrelations, aromaticity indices, randic molecular profiles, radial distribution functions, functional groups, atom-centred fragments, empirical, and properties. All the algorithms were programmed in the computer system Matlab 7.6.[41] The calculation of the descriptors in the ACET, GTT, and MTT data sets differs from the rest and is discussed elsewere.[38]

RESULTS AND DISCUSSION

In recent years, research has focused increasing attention on finding the most efficient tool for variable selection in QSAR/QSPR studies.[26] In fact, many techniques were not successful for all the cases under consideration, especially for too large data sets. This is the main reason why we used different databases (10 in total) for comparing the performance of the algorithms tested in this work.

With the purpose of providing a graphical inspection of the behavior of our two methods, i.e., RM and ERM, Figures 1 and 2 show *S* as a function of the number of steps for both algorithms and the optimization of a seven-parameter model on the FLUOR data set.[32] Figure 1 reveals that ERM exhibits three sections: a first one due to RM, a second one that simulates a higher temperature or a higher noise than the RM, although maintaining the overall decreasing tendency of the *S* function, and finally a third section where a second RM forces *S* to decrease. This apparent thermal agitation makes the ERM less likely to get trapped by a local minimum at the cost of a slower convergence and a greater computer time.[19] Figure 2 displays the great convergence rate of RM as the number of steps is increased. Finally, in order to illustrate the behavior of the GA with its parameters, Figures 3−5 show *S* as a function of the population number for three cases, where the number of individuals was 20, 5, and 100, respectively. For all those cases, the generation gap was 0.9, single-point crossover probability was 0.6, and mutation probability was 0.7/*d*.

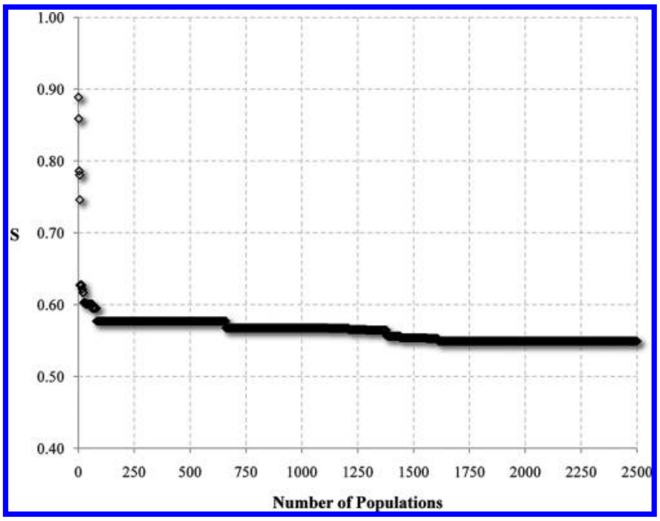

Figure 3. Standard deviation vs population number for GA with number of individuals = 20, generation gap = 0.9, single-point crossover probability = 0.6, and mutation probability = 0.7/*d*.

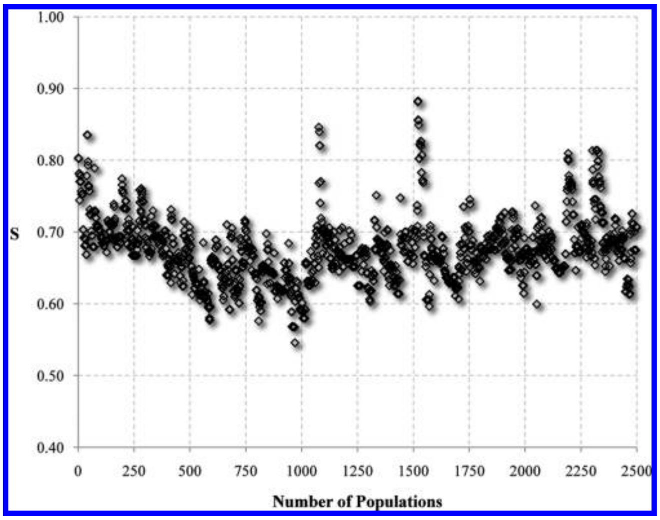

Figure 4. Standard deviation vs population number for GA with number of individuals = 5, generation gap = 0.9, single-point crossover probability = 0.6, and mutation probability = 0.7/*d*.

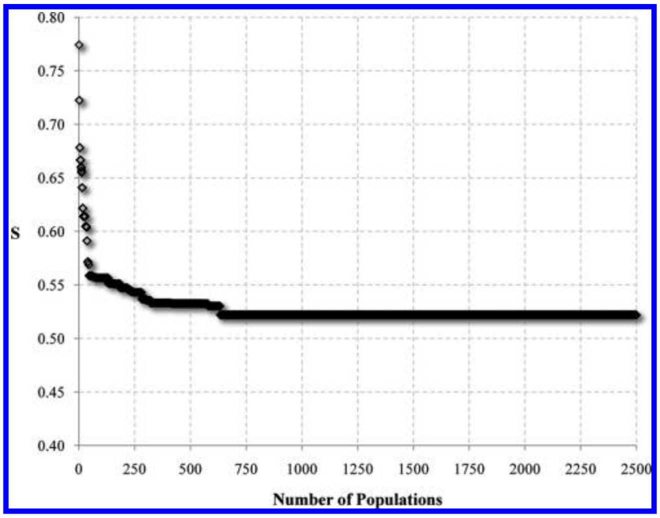

Figure 5. Standard deviation vs population number for GA with number of individuals = 100, generation gap = 0.9, single-point crossover probability = 0.6, and mutation probability =0.7/*d*.

We carried out all the numerical tests by taking values of *d* from 3 to 7, the last one as an example of a computationally demanding search with a reasonable number of descriptors for a potential model in common QSPR/QSAR studies. The number of variables $d = 1$ and 2 were excluded because in such cases one can carry out a full search in relatively short times, as shown in Table 1. As discussed before, such FS is impractical for greater *d* even for a small database (GABA, $D = 1187$).

It is worth mentioning that in the application of the RM and ERM to QSAR/QSPR studies, models with increasing number of descriptors are easily obtained. The optimal *d* is then determined using a criteria to select the model with the best statistical parameters and the additional requirement that the model does not overfit the data.[42] The selection of the optimal sets of *d* descriptors for all the algorithms and databases is not presented in here for space reasons. Another practical aspect of the application of the search methods to particular problems consists of easily avoiding models with strongly correlated descriptors by taking out of the pool those descriptors with a correlation higher than a limiting value.

The GA optimization requires several runs. In almost all the cases, the optimized parameters were: number of individuals = 250; generation gap = 0.9; single-point crossover probability = 0.6; mutation probability = 0.7/*d*. The exceptions were: in the cases of FLUOR and TOX1 with $d = 7$: number of individuals = 20, generation gap = 2, single-point crossover probability = 0.6, and mutation probability =0.7/*d*; and TOX1 database with $d = 6$: number of individuals = 100, generation gap = 0.9, single-point crossover probability = 0.6, and mutation probability = 0.7/*d*. All GA algorithms were stopped when one individual occupied more than 90% of the population or when the number of generations reached 2500.

We used as a starting point three random initial sets for ERM by following the same strategy as our previous work.[19] For the case of RM, 10 initial sets of descriptors were selected as follows: one random set, and nine sets of descriptors searched with the stepwise inclusion procedure in such a way that the maximum correlation coefficient between descriptors for each of the nine linear models were of 0.1−0.9, respectively. These sets of initial solutions for RM presented the best results for the application of this method.

Since the computational demand was different for ERM and RM, and in order to present the results in a clearer way,

Table 1. Number of Necessary Linear Regressions and Estimated Computation Times for a Full Search on a Set of $D = 1187$ Descriptors[a]

d	1	2	3	4	5	6	7
number of regressions	1187	703 891	2.8×10^{8}	8.2×10^{10}	1.9×10^{13}	3.8×10^{15}	6.5×10^{17}
minutes	0.003	1.9	7.7×10^{2}	2.3×10^{5}	5.4×10^{7}	1.1×10^{10}	1.8×10^{12}
hours	5.5×10^{-5}	3.2×10^{-2}	12.8	3791.6	9.0×10^{5}	1.8×10^{8}	3.0×10^{10}
days	2.3×10^{-6}	1.4×10^{-3}	5.3×10^{-1}	158.0	3.7×10^{4}	7.4×10^{6}	1.2×10^{9}
years	6.2×10^{-9}	3.7×10^{-6}	1.5×10^{-3}	$4.3.10^{-1}$	1.0×10^{2}	2.0×10^{4}	3.4×10^{6}

[a] Using an AMD Athlon 64 2800+ processor.

Table 2. Standard Deviation and Number of Linear Regressions for ERM and GA Using Seven Different Data Sets and $d = 3-7$[a]

data set	*d*	3	4	5	6	7
MES	ERM	**0.351**	**0.334**	**0.323**	**0.303**	**0.290**
	GA	**0.351**	**0.334**	**0.323**	0.308	0.297
GI	ERM	**0.512**	**0.495**	**0.478**	**0.455**	**0.437**
	GA	**0.512**	**0.495**	0.479	0.461	0.442
GABA	ERM	**0.607**	**0.553**	**0.490**	**0.444**	**0.396**
	GA	**0.608**	**0.553**	0.513	**0.444**	**0.396**
FLUOR	ERM	**0.673**	**0.602**	**0.538**	**0.499**	**0.433**
	GA	**0.673**	**0.602**	0.563	0.500	0.460
SOL	ERM	**0.921**	**0.861**	**0.828**	**0.806**	**0.779**
	GA	**0.921**	0.905	0.846	0.817	0.803
TOX1	ERM	**0.474**	**0.448**	**0.427**	**0.408**	**0.398**
	GA	**0.474**	**0.448**	**0.427**	0.410	**0.398**
TOX2	ERM	0.368	**0.331**	**0.311**	**0.303**	**0.289**
	GA	**0.362**	**0.331**	0.320	0.308	0.299
number of regressions	ERM	962 031	1 765 060	2 586 950	4 251 480	5 014 618
	GA	1 167 781	1 881 094	2 681 219	4 310 438	5 045 313

[a] The best results appear in boldface numbers.

Table 3. Standard Deviation and Number of Linear Regressions for RM and GA Using Seven Different Data Sets and $d = 3-7$[a]

data set	d	3	4	5	6	7
MES	RM	0.353	0.342	0.328	0.316	0.307
	GA	**0.351**	**0.334**	**0.323**	**0.313**	**0.297**
GI	RM	0.553	0.520	**0.479**	**0.464**	**0.442**
	GA	**0.512**	**0.503**	0.501	**0.462**	0.452
GABA	RM	**0.608**	0.560	**0.513**	0.476	0.431
	GA	0.608	**0.553**	**0.513**	**0.444**	**0.396**
FLUOR	RM	**0.673**	0.636	0.579	0.519	0.494
	GA	**0.673**	**0.620**	**0.563**	**0.502**	**0.460**
SOL	RM	**0.926**	**0.861**	**0.828**	**0.806**	**0.789**
	GA	0.969	0.913	0.866	0.817	0.817
TOX1	RM	**0.474**	**0.448**	**0.427**	0.410	**0.398**
	GA	**0.474**	**0.448**	**0.427**	**0.410**	0.400
TOX2	RM	0.370	0.344	**0.314**	0.317	**0.301**
	GA	**0.362**	**0.331**	0.320	**0.308**	0.301
number of regressions	RM	551 301	955 460	1 485 470	2 134 497	2 883 321
	GA	700 669	940 547	1 532 125	2 394 688	3 027 188

[a] The best results appear in boldface numbers.

Table 4. Comparison between Linear Methods Using Three Different Data Sets Analyzed with GA in Ref 38[a]

data set	method	d	S	descriptors involved
ACET	**RM**	**4**	**0.288**	$<C_4 - 0.027>$, $<2.301 - U_t>$, $<U_t - 2.855>^2$, $<-9.631 - \text{homo}>^2$
	ERM	**4**	**0.288**	$<C_4 - 0.027>$, $<2.301 - U_t>$, $<U_t - 2.855>^2$, $< -9.631 - \text{homo}>^2$
	GA	4	0.289	$<C_4$, $<2.301 - U_t>$, $<U_t - 2.845>^2$, $<-9.631 - \text{homo}>^2$
GTT	**RM**	**5**	**14.045**	$<-1.28 - \bar{E}_D>$, $<\bar{E}_+ + 0.14>$, $<\bar{E}_+ + 2.13>$, $<-0.41 - \bar{E}_->$, $<-0.04 - \bar{E}_->$
	ERM	**5**	**14.045**	$<-1.28 - \bar{E}_D>$, $<\bar{E}_+ + 0.14>$, $<\bar{E}_+ + 2.13>$, $<-0.41 - \bar{E}_->$, $<-0.04 - \bar{E}_->$
	GA	5	16.054	$<\bar{E}_+ + 1.95>$, $<-1.52 - \bar{E}_D>$, $<-0.04 - \bar{E}_->$, $<-0.37 - \bar{E}_->$, S_B
MTT	RM	5	40.093	$<0.8 - S_B>$, $<S_S - 1.7>$, $<\bar{E}_D + 1.66>$, $<-0.17 - \bar{E}_+>$, $<\bar{E}_- + 0.66>$
	ERM	**5**	**39.554**	$<0.8 - S_B>$, $<S_S - 2.91>$, $<\bar{E}_D + 1.65>$, $<0.4 - \bar{E}_+>$, $<\bar{E}_- + 0.66>$
	GA	5	42.643	$\bar{E}_+$, $<\bar{E}_D + 1.66>$, $<\bar{E}_- + 0.8>$, $<S_S - 2.91>$, $<M_S - 15>$

[a] The best results appear in boldface.

we decided to contrast them against GA in separate tables (Tables 2 and 3). The computational demand in GA was set by modifying the number of runs in order to have a similar number of regressions as for ERM and RM. It follows from Table 2 that ERM outperforms or equals GA for all cases except when $d = 3$ for the TOX2 data set. This particular case appears to be fortuitous, since ERM is preferable to GA in all the other cases. It should to be kept in mind that since the GA is a nondeterministic methodology, then its results may change for different runs using exactly the same initial conditions. Table 3 also suggests that the GA is better than the RM in 51.4% of the cases, the latter approach is preferable in 31.4% of the cases and both methods produce similar results in 17.1% of the cases. However, it should be kept in mind that the RM is a much simpler algorithm. It is also clear that the ERM is preferable to the RM, as previously reported.[19] This fact is not surprising if we take into account that the ERM is an improved algorithm that contains the RM, as outlined above.

As a further analysis, we compared present RM and ERM results with the GA ones already published for three different data sets: ACET, GTT, and MTT.[38] Table 4 shows that both the RM and ERM lead to models with values of S that are smaller or similar to those of GA. The meaning of the molecular descriptors appearing in this table were discussed elsewhere.[38]

Table 5. Correlation Coefficient of Leave-One-Out Cross Validation for ERM and GA Using Seven Different Data Sets and $d = 3-7$[a]

data set	d	3	4	5	6	7
MES	ERM	**0.581**	**0.636**	**0.652**	**0.698**	**0.726**
	GA	**0.581**	**0.636**	0.643	0.694	0.722
GI	ERM	**0.778**	**0.795**	**0.805**	**0.825**	**0.837**
	GA	**0.778**	**0.795**	0.805	0.822	0.835
GABA	ERM	**0.787**	**0.831**	**0.867**	**0.895**	**0.912**
	GA	**0.787**	**0.831**	0.850	**0.895**	**0.912**
FLUOR	ERM	–	**0.967**	**0.971**	**0.978**	**0.983**
	GA	–	**0.967**	0.971	0.977	0.981
SOL	ERM	**0.855**	**0.874**	**0.884**	**0.888**	**0.896**
	GA	**0.855**	0.860	0.876	0.886	0.889
TOX1	ERM	0.868	**0.895**	**0.906**	**0.912**	**0.919**
	GA	**0.872**	**0.895**	**0.902**	0.909	**0.911**
TOX2	ERM	0.878	**0.892**	**0.902**	**0.912**	**0.910**
	GA	**0.878**	**0.892**	0.902	0.911	0.910

[a] The best results appear in boldface numbers.

In order to facilitate the discussion of the following analysis of the search methods, we only compare the ERM and GA.

A well-known theoretical validation of the linear models is the leave-one-out cross validation procedure (loo).[43] Table 5 shows results for all the cases except one for which it was not possible to implement the loo. This fact is not a serious drawback because one can resort to the second best model for future practical applications. In earlier papers we tested the reliability of our methods by means of additional validation methods like the leave-more-out cross-validation[43] and the external test set validation.[17,22,23,29,30,32,33,44,45] However, we did not try them on all the models in this paper because it would have been extremely time consuming. According to the specialized literature, R^2_{loo} should be greater than 0.50 for a properly validated model.[46] The analysis in Table 5 suggests ERM outperforms or equals GA for all the cases except $d = 3$ of the TOX2 data set. It should be taken into account that the implementation of the GA technique required the tuning of the previously mentioned parameters, making its execution much more laborious and complicated. If we had taken the computational demand of the GA into consideration, then the advantage brought by the application of the ERM and RM would have been more noticeable.

CONCLUSIONS

In this work we performed an extensive and reliable comparison of enhanced replacement (ERM) and replacement (RM) methods versus genetic algorithm (GA). The results of this paper support earlier ones based on comparisons for only one case at a time[22,29,30] and suggested that ERM was preferable to GA. To the quality of the results, we should add the fact that the ERM is much simpler than the GA. We have also shown that although the GA is slightly better than the RM, simplicity and lower computational cost make the latter more attractive.

Finally, it is worth mentioning the three methods RM, ERM, and GA can be used under different conditions as alternative strategies for the construction of models for chemical properties and activities from quite large pools of descriptors for molecular structure.

ACKNOWLEDGMENT

The authors want to thank the National Research Council of Argentina (CONICET) and Universidad de Buenos Aires (UBA) for financial support and the MINCYT for the electronic library facilities. A.G.M. thanks CONICET for a Postdoctoral Research Fellowship in PRALIB (CONICET, UBA), on leave from INIFTA (CONICET, UNLP). P.R.D. thanks the National University of La Plata. We thank the helpful comments from the reviewers of this manuscript.

REFERENCES AND NOTES

(1) Hansch, C.; Leo, A. *Exploring QSAR. Fundamentals and Applications in Chemistry and Biology*; American Chemical Society: Washington, D.C., 1995.

(2) Diudea, M. V. E. *QSPR/QSAR Studies by Molecular Descriptors*; Nova Science Publishers: New York, 2001.

(3) Noringer, U. In silico modeling of ADMET-a minireview of work from 2000 to 2004. *SAR QSAR Environ. Res.* **2005**, *16*, 1–11.

(4) Benfenati, E. *Quantitative Structure-Activity Relationship (QSAR) for Pesticide Regulatory Purposes*; Elsevier: Amsterdam, The Netherlands, 2007.

(5) Kubinyi, H. *QSAR: Hansch Analysis and Related Approaches*; Wiley-Interscience: New York, 2008.

(6) Puzyn, T.; Leszczynski, J.; Cronin. M. T. *Recent Advances in QSAR Studies: Methods and Applications*, 1st ed.; Springer: New York, 2009.

(7) Karelson, M. *Molecular Descriptors in QSAR/QSPR*; Wiley-Interscience: New York, 2000.

(8) Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; Wiley-VCH: Weinheim, Germany, 2009; Vol. 2.

(9) Trinajstic, N. *Chemical Graph Theory*; CRC Press: Boca Raton, FL, 1992.

(10) *Coral Version 1.4*, http://www.insilico.eu/coral (accessed Apr 3, 2010).

(11) *Recon*, version 5.5; Rensselaer Polytechnic Institute: Troy, New York; http://www.drugmining.com. Accessed April 3, 2010.

(12) Putz, M. V.; Lacrama, A.-M. Introducing spectral structure activity relationship (S-SAR) analysis. Application to ecotoxicology. *Int. J. Mol. Sci.* **2007**, *8*, 363–391.

(13) Lacrama, A.-M.; Putz, M. V.; Ostafe, V. A spectral-SAR model for the anionic-cationic interaction in ionic liquids: application to Vibrio fischeri ecotoxicity. *Int. J. Mol. Sci.* **2007**, *8*, 842–863.

(14) Chicu, S. A.; Putz, M. V. Köln-Timisoara molecular activity combined models toward interspecies toxicity assessment. *Int. J. Mol. Sci.* **2009**, *10*, 4474–4497.

(15) Putz, M. V.; Putz, A.-M.; Lazea, M.; Ienciu, L.; Chiriac, A. Quantum-SAR Extension of the spectral-SAR algorithm. Application to polyphenolic anticancer bioactivity. *Int. J. Mol. Sci.* **2009**, *10*, 1193–1214.

(16) Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. A New Search Algorithm of QSPR/QSAR Theories: Normal Boiling Points of Some Organic Molecules. *Chem. Phys. Lett.* **2005**, *412*, 376–380.

(17) Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. QSAR of Non-Nucleoside Inhibitors of HIV-1 Reverse Transcriptase. *Bioorg. Med. Chem.* **2006**, *14*, 5876–5889.

(18) Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179–192.

(19) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A. Modified and Enhanced Replacement Method for the Selection of Molecular Descriptors in QSAR and QSPR Theories. *Chemom. Intell. Lab. Syst.* **2008**, *92*, 138–144.

(20) Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley&Sons: New York, 1981.

(21) So, S. S.; Karplus, M. Evolutionary Optimization in Quantitative Structure-Activity Relationship: An Application of Genetic Neural Networks. *J. Med. Chem.* **1996**, *39*, 1521–1530.

(22) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Bennardi, D. O.; Autino, J. C.; Romanelli, G. P. QSAR prediction of inhibition of aldose reductase for flavonoids. *Bioorg. Med. Chem.* **2008**, *16*, 7470–7476.

(23) Duchowicz, P. R.; Vitale, M. G.; Castro, E. A.; Fernandez, M.; Caballero, J. QSAR analysis for heterocyclic antifungals. *Bioorg. Med. Chem.* **2007**, *15*, 2680–2689.

(24) Duchowicz, P. R.; Talevi, A.; Bruno-Blanch, L. E.; Castro, E. A. New QSPR Study for the Prediction of Aqueous Solubility of Drug-Like Compounds. *Bioorg. Med. Chem.* **2008**, *16*, 7944–7955.

(25) Duchowicz, P. R.; Goodarzi, M.; Ocsachoque, M. A.; Romanelli, G. P.; Ortiz, E. V.; Autino, J. C.; Bennardi, D. O.; Ruiz, D.; Castro, E. A. QSAR Analysis on Spodoptera litura Antifeedant Activities for Flavone Derivatives. *Sci. Total Environ.* **2009**, *408*, 277–285.

(26) Goodarzi, M.; Duchowicz, P. R.; Wu, C. H.; Fernández, F. M.; Castro, E. A. New Hybrid Genetic Based Support Vector Regression as QSAR Approach for Analyzing Flavonoids-GABA(A) Complexes. *J. Chem. Inf. Model.* **2009**, *49*, 1475–1485.

(27) Metropolis, N.; Rosenbluth, A. W.; Rosenbluth, M. N.; Teller, A. H.; Teller, E. Equation of State Calculations by Fast Computing Machines. *J. Chem. Phys.* **1953**, *21*, 1087–1092.

(28) Kirkpatrick, S.; Gelatt, C. D., Jr.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(29) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Wolcan, E. QSPR study of solvent quenching of the $5D_{0}\rightarrow{}^{7}F_{2}$ emission of Eu(6,6,7,7,8,8,8-heptafluoro-2,2-dimethyl-3,5-octanedionate)$_3$. *Chem. Phys. Lett.* **2008**, *462*, 352–357.

(30) Mercader, A. G.; Duchowicz, P. R.; Fernández, F. M.; Castro, E. A.; Cabrerizo, F. M.; Thomas, A. H. Predictive Modeling of the Total Deactivation Rate Constant of Singlet Oxygen by Heterocyclic Compounds. *J. Mol. Graphics Modell.* **2009**, *28*, 12–19.

(31) Melanie, M. A. *An Introduction to Genetic Algorithms*; The MIT Press: Cambridge, MA, 1998.

(32) Mercader, A. G.; Duchowicz, P. R.; Sanservino, M. A.; Fernandez, F. M.; Castro, E. A. QSPR analysis of fluorophilicity for organic compounds. *J. Fluorine Chem.* **2007**, *128*, 484–492.

(33) Duchowicz, P. R.; Mercader, A. G.; Fernández, F. M.; Castro, E. A. Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR. *Chemom. Intell. Lab. Syst.* **2008**, *90*, 97–107.

(34) Duchowicz, P. R.; Vitale, M. G.; Castro, E. A.; Autino, J. C.; Romanelli, G. P.; Bennardi, D. O. QSAR Modeling of the Interaction of Flavonoids with GABA(A) Receptor. *Eur. J. Med. Chem.* **2007**, *43*, 1593–1602.

(35) Garro-Martínez, J. C.; Duchowicz, P. R.; Estrada, M. R.; Zamarbide, G. N.; Castro, E. A. Anticonvulsant Activity of Ringed Enaminones: A QSAR Study. *QSAR Comb. Sci.* **2009**, *28*, 1376–1385.

(36) Duchowicz, P. R.; Ocsachoque, M. A. Quantitative Structure-Toxicity Models for Heterogeneous Aliphatic Compounds. *QSAR Comb. Sci.* **2009**, *28*, 281–295.

(37) Castillo-Garit, J. A.; Marrero-Ponce, Y.; Escobar, J.; Torrens, F.; Rotondo, R. A novel approach to predict aquatic toxicity from molecular structure. *Chemosphere* **2008**, *73*, 415–427.

(38) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure-Activity Relationships and Quantitative Structure-Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854–866.

(39) *Hyperchem*, version 6.03; Hypercube, Inc.: Gainesville, 2007.

(40) *Dragon*; Milano Chemometrics and QSAR Research Group: Milano, Italy; http://michem.disat.unimib.it/chm. Accessed April 3, 2010.

(41) *Matlab*, version 7.6; The MathWorks, Inc.: Natick, MA, 2008.

(42) Mercader, A. G. Selection of an optimal set of descriptors: use of the Enhanced Replacement Method. In *QSPR-QSAR Studies on Desired Properties for Drug Design*; Castro, E. A., Ed.; Signpost Design: India, 2009.

(43) Hawkins, D. M.; Basak, S. C.; Mills, D. Assessing Model Fit by Cross Validation. *J. Chem. Inf. Model.* **2003**, *43*, 579–586.

(44) Helguer, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. Application of the Replacement Method as Novel Variable Selection Strategy in QSAR. 1. Carcinogenic Potential. *Chemom. Intell. Lab. Syst.* **2006**, *81*, 180–187.

(45) Duchowicz, P. R.; Gonzalez, M. P.; Helguera, A. M.; Cordeiro, M. N. D. S.; Castro, E. A. Application of the Replacement Method as Novel Variable Selection in QSPR. 2. Soil Sorption Coefficients. *Chemom. Intell. Lab. Syst.* **2007**, *88*, 197–203.

(46) Golbraikh, A.; Tropsha, A. Beware of q2! *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

CI100103R