Chemometrics and Intelligent Laboratory Systems 149 (2015) 117–122

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

Enhanced replacement method integration with genetic algorithms populations in QSAR and QSPR theories

Andrew G. Mercader *, Pablo R. Duchowicz

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

ARTICLE INFO

Article history:
Received 20 March 2015
Received in revised form 4 October 2015
Accepted 19 October 2015
Available online 27 October 2015

Keywords:
Enhanced replacement method
Genetic algorithm
QSAR
QSPR

ABSTRACT

The selection of an optimal set of molecular descriptors from a much larger collection of such regression variables is a vital step in the elaboration of most QSAR and QSPR models. The aim of this work is to continue advancing this important selection process by combining the enhanced replacement method (ERM) and the well-known genetic algorithms (GA). These approaches had previously proven to yield near-optimal results with a much smaller number of linear regressions than a full search. The newly proposed algorithms were tested on four different experimental datasets, formed by collections of 116, 200, 78, and 100 experimental records from different compounds and 1268, 1338, 1187, and 1306 molecular descriptors, respectively. The comparisons showed that the new alternative ERMp (combination of ERM with a GA population) further improves ERM, it has previously been shown that the latter is superior to GA for the selection of an optimal set of molecular descriptors from a much greater pool.

1. Introduction

One generally accepted solution to overcome the lack of experimental data in complex chemical phenomena is the analysis based on quantitative structure-property/activity relationships (QSPR/QSAR) [1]. Therefore, there exists a permanently renewed interest on the development of these techniques [2–6]. The fundamental role of QSPR/QSAR is to suggest mathematical models capable of predicting important properties or activities of interest, particularly when those cannot be experimentally determined for some reason. These techniques rely on the basic assumption that the structure of a compound determines its properties, and that the structures can be represented by molecular descriptors [7], which are calculated through mathematical formulae parameters obtained from several theories, such as chemical graph theory, information theory, and quantum mechanics [8,9].

Currently, there are thousands of molecular descriptors available in the literature [7], and in any QSAR/QSPR study, those that characterize the property/activity under consideration in the most efficient way must be selected. Consequently, a mathematical problem of selecting a subset of d descriptors ($\mathbf{d}$) from a much larger set of D descriptors, arises.

The search for this optimal set of molecular descriptors is generally oriented to find the model that minimizes the standard deviation (S). In other words, the global minimum of $S(\mathbf{d})$ is seek, where $\mathbf{d}$ is a point in a space of $D!/[d!(D-d)!]$ ones. Since D is very large, a full search (FS) of the optimal variables is impractical because it requires $D!/[d!(D-d)!]$ linear regressions.

Some time ago, our research group proposed the replacement method (RM) [10,11]; afterward, the enhanced replacement method (ERM) [12] and later on a modification of the first step that further improved both algorithms (RMfs and ERMfs) [13]. All these methods produce linear regression QSPR/QSAR models that present no relevant difference with FS using much less computational work [12]. These alternative techniques approach the minimum of S by taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d} = \{X_1, X_2, \ldots, X_d\}$. All the methods give models with greatly better estimative and predictive ability than the forward stepwise regression procedure [14]; in addition, the ERM has proven to give better results than the more elaborate genetic algorithms [15,16] (GA).

The first step in ERM does not use the same scheme as the rest of the algorithm; nevertheless, in a recent article, it was proven that by using d different initial sets of descriptors and taking as a first step the replacement of the descriptor with higher relative standard deviation, ERM results were improved [13].

The main target of this work is to combine the latest version of ERM (ERMfs) with GA, to find an algorithm that further improves the previous ones.

2. Methods

2.1. Algorithms

The following subsections briefly describe the theory of the previous state of ERM, GA and the new alternative algorithm (ERMp).

* Corresponding author. Tel.: +54 221 425 7430; fax: +54 221 425 4642.
E-mail address: amercader@inifta.unlp.edu.ar (A.G. Mercader).

http://dx.doi.org/10.1016/j.chemolab.2015.10.007
0169-7439/

All the algorithms were programmed in the computer system Matlab 5.0 [17]. Tests were done using d from 5 to 9 in order to have a high computational demanding search with a reasonable number of descriptors that might be used in a QSPR/QSAR study model.

The comparisons of ERM and ERMp were done through runs with an increasing number of randomly chosen initial sets of descriptors, from 1 to 250. Aiming to determine if increasing the initial number of sets (which in the case of ERM is equal to d) could further minimize the obtained standard deviation.

3. Enhanced replacement method

An optimal subset $\mathbf{d}_m = \{X_{m1}, X_{m2}, \ldots, X_{md}\}$ of $d << D$ was chosen from a large set $\mathbf{D} = \{X_1, X_2, \ldots, X_D\}$ of D descriptors provided by some available commercial program, with minimum standard deviation S:

$$S = \sqrt{\frac{1}{(N-d-1)} \sum_{i=1}^{N} res_i^2} \tag{1}$$

where N is the number of molecules in the training set, and res_i the residual for molecule i (difference between the experimental and predicted property). The fact that $S(\mathbf{d}_n)$ is a distribution on a discrete space of $D!/d!(D-d)!$ disordered points $\mathbf{d}_n$ should be noticed. The full search (FS) that consists of calculating $S(\mathbf{d}_n)$ on all those points always allows to arrive at the global minimum, but as mentioned, it is computationally prohibitive if D is sufficiently large; as an example, using $d = 7$ and $D = 1280$, the number of necessary regressions would be 1.0×10^{18} regressions, which translates to as much as 1.5×10^5 years to complete only one calculation (using an i7 4770K processor). The ERM briefly consists of the following steps:

- An initial set of descriptors $\mathbf{d}_k$ is selected from D at random, one of the descriptors is replaced, denoted as X_{ki}, with all the remaining $D - d$ descriptors, one by one, and the set with the smallest value of S is kept. What was done unto this point is defined as a 'step'.
- From this resulting set, the descriptor with the greatest standard deviation in its coefficient is chosen (the one changed previously is not considered) and substituted with all the remaining $D - d$ descriptors, one by one. This procedure is repeated until the set remains unmodified. In each of these cycles, the descriptors replaced in previous steps are not taken into account. Thus, the candidate $\mathbf{d}_m^{(i)}$ that comes from the so-constructed path i is obtained. The 'paths' are consequently defined as all possible steps to start the algorithm from the initial set of descriptors.
- It should be noticed that if the replacement of the descriptor with the largest error by those in the pool does not decrease the value of S, then that descriptor is not changed.
- The above process is carried out for all the possible paths $i = 1, 2, \ldots, d$ and the point $\mathbf{d}_m$ with the smallest standard deviation: $\min_i S(\mathbf{d}_m^{(i)})$ is kept.

So far, the replacement method (RM) has been described; the ERM is a three-step combination of two algorithms: first, the RM, then a modified RM (MRM) and finally a RM is used again. The MRM follows the same strategy as RM except that, in each step, the descriptor with the largest error is substituted even if that substitution is not accompanied by a smaller value of S (the next smallest value of S is chosen). The main difference in MRM is that it adds some sort of noise that prevents the selected model from staying in a local minimum of S [12].

In the original ERM algorithm, the first step was chosen without taking into account the relative standard deviation (*rsd*) of the coefficient of the descriptor in the model; instead, all possible d paths were followed one at a time. Mainly because in the practical use of the algorithms, it was noticed that the best results did not always depend on the initial *rsd* of the path [10–12,18–26]. However, after further studies, it was determined that if only the initial descriptor substitution with higher *rsd* was used and additional $d - 1$ starting sets of descriptors were used, better results for the same computational cost were found [13]; hence, finding a way to advance the algorithm.

3.1. Genetic algorithms

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) 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 do not seem to be conclusive results on which might be the best ones [27].

In this study, it was necessary to modify an existing GA adapting it to QSAR studies; for this purpose the "GA Toolbox for MATLAB" developed at the Department of Automatic Control and Systems Engineering of The University of Sheffield, UK [28] was used.

3.2. ERM with initial population from GA (ERMp)

Several preliminary trials with the following combinations of ERM and GA were tested:

- GA with mutation operator position determined by *rsd* instead of a random selection

Two options were tested: one where the mutation was oriented by *rsd* but the descriptors were randomly selected, and a second one where the mutation was done by replacing all the rest of the D descriptors and choosing the one that minimizes the S of the model.

- GA with the crossover operator position determined by *rsd* instead a random selection
- GA with both modified mutation and modified crossover operator
- ERM with an initial population (ERMp) similar to GA (this option was made available after the latest algorithm (ERMfs) described in the previous section.)

The results revealed that the first three alternatives did not deserve further studies since they gave much worst results than the original GA in all cases.

Only the final option (ERMp) was further studied since it showed better results than ERM, at the expense of a higher computational cost.

This new algorithm starts with a population of random initial sets of descriptors, and then applies the ERMfs. Increasing the number of individuals in the starting population makes the number of required operations grow accordingly; the computational cost is proportional to the number of initial descriptor sets in the population.

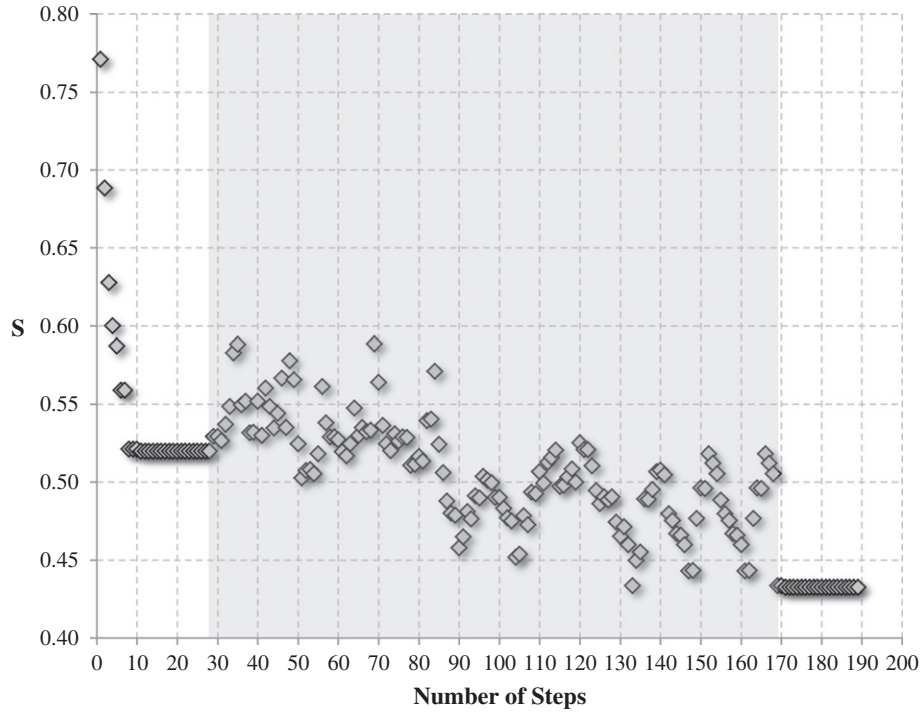

Fig. 1. Standard deviation vs. number of steps for the ERM. The first part of the graph corresponds to a RM procedure, the second to a MRM (marked in gray) and the final to another RM.

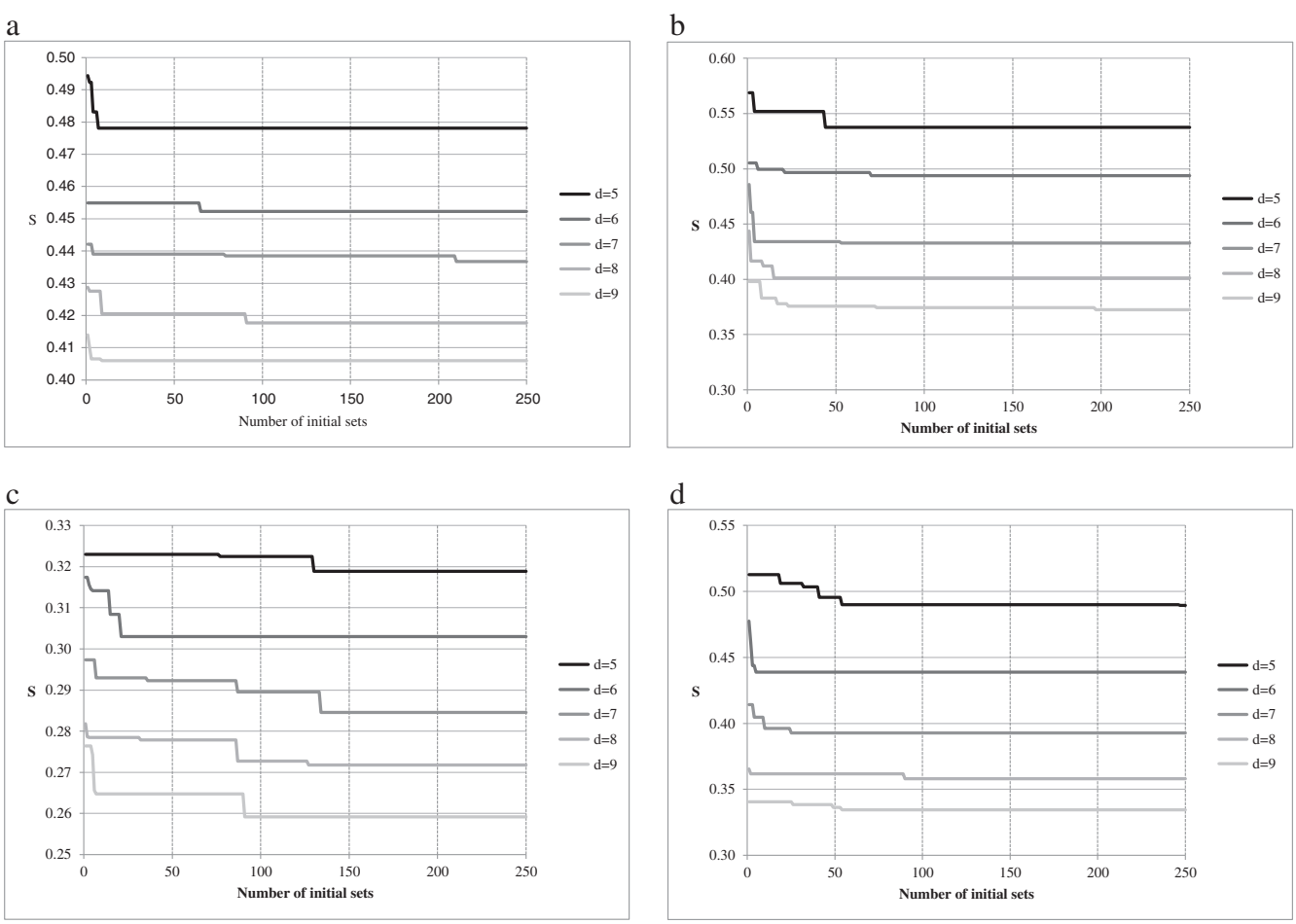

Fig. 2. Standard deviation of the best model vs. number of initial sets for: (a) the GI dataset; (b) the FLUOR dataset; (c) the MES dataset; and (d) the GABA dataset.

4. Materials

4.1. Datasets

Four different experimental datasets previously analyzed were used to test and contrast the performance of RM, ERM and the new alternatives.

A fluorophilicity dataset (FLUOR), consisting of 116 organic compounds characterized by 1268 theoretical descriptors. The fluorophilicity of each compound was quantified through the associated partition coefficient (P) between f1uorous ($CF_3C_6F_{11}$) and organic ($CH_3C_6H_5$) layers.

$$\ln P = \ln \left[\frac{c(CF_3C_6F_{11})}{c(CH_3C_6H_5)}\right] \qquad T = 298\ K \tag{1.2}$$

The tendency of an organic substance to dissolve in fluorous media has continuously gained importance after the disclosure of the fluorous biphase catalysis, as biphasic reactions take advantage of the fact that organic and fluorous phases are typically immiscible at room temperature, but may homogenize at elevated temperatures [26].

A growth inhibition dataset (GI), with growth inhibition values to the ciliated protozoan *Tetrahymena pyriformis* by 200 mechanistically diverse phenolic compounds and 1338 structural descriptors. The aqueous toxicities are expressed as $pIGC_{50} = \log(IGC_{50}^{1})$, with IGC_{50} expressing the concentration [mmol l^{-1}] producing a 50% growth inhibition on *Tetrahymena pyriformis* under a static regime [20].

A GABA receptor dataset (GABA), containing 78 inhibition data for flavone derivatives and 1187 molecular descriptors The dataset consists of the logarithm of the experimental binding affinity constants ($\log_{10} K_i$ [μM]) of flavonoid ligands for the benzodiazepine site of the GABA(A) receptor complex in washed crude synaptosomal membranes from rat cerebral cortex [21].

Additionally, a dataset that consists of 100 $\log_{10} ED_{50}$ mice antiepileptic experimental activity values for enaminones with 1306 descriptors. The activity ED_{50} represents the dose of the chemical compound for which 50% of the individuals reached the desired effect obtained by the 'maximal electroshock seizure' (MES) experimental method [29].

The datasets were divided into training sets and test sets, containing 67% and 33% of the molecules, respectively. For this purpose, a k-means cluster analysis was used to obtain representative molecules from the complete dataset in both the training and test sets [30]. Leading to the following subsets: FLUOR, 78 training and 38 test set molecules; GI, 134 training and 66 test set molecules; GABA, 52 training and 26 test set molecules; and MES, 67 training and 33 test set molecules.

In all cases, the structures of the compounds were pre-optimized with the molecular mechanics force field (MM+)[31] procedure included in Hyperchem version 6.03,[32], and the resulting geometries were further refined by means of the semi-empirical method PM3 (Parametric Method-3)[33] using the Polak–Ribiere algorithm and a gradient norm limit of 0.01 kcal/Å. The molecular descriptors were calculated using the software Dragon 3.0,[34] including parameters of all types such as constitutional, topological, geometrical, quantum mechanical, etc.

5. Results and discussion

To provide a graphical visualization of the behavior of the ERM algorithms, Fig. 1 shows S as a function of the number of steps for

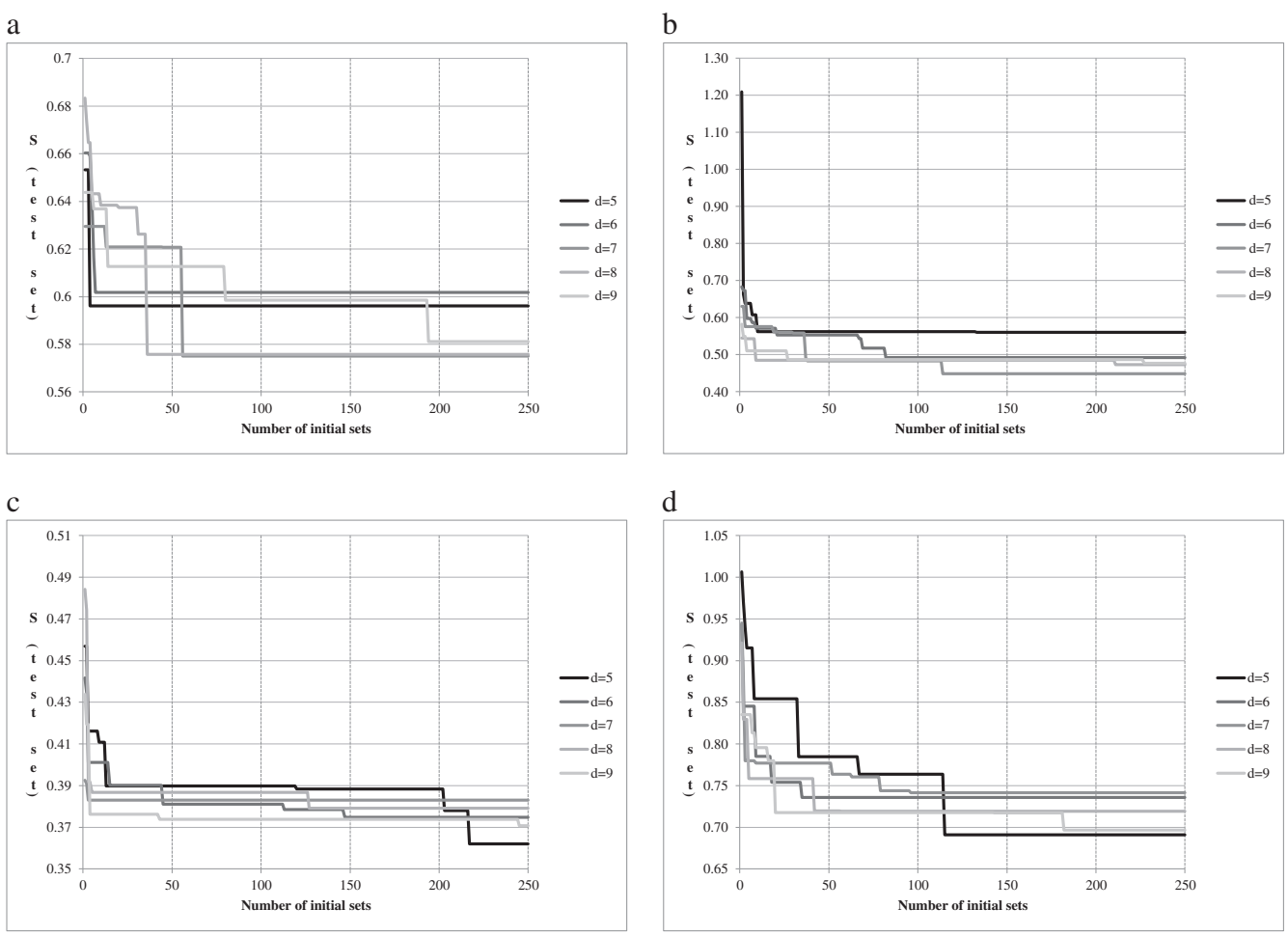

Fig. 3. Test set standard deviation of the best model vs. number of initial sets for: (a) the GI dataset; (b) the FLUOR dataset; (c) the MES dataset; and (d) the GABA dataset.

ERM for the optimization of a seven-parameter model using the MES dataset [29]. Fig. 1 reveal that ERM has three sections, a first section where RM is performed, a second section (MRM) 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 is used to further decrease *S*. This apparent thermal agitation makes the ERM less likely to get trapped in a local minimum [12]. The behavior of the new alternative is similar to the one shown in Fig. 1.

The improvement in the *S* of the best found model as the number of initial sets is increased can be seen in Fig. 2a, b, c and d, for the four different datasets, using models containing 5 to 9 descriptors.

In order to determine possible overfitting of the models by the new algorithm due to the additional decrease of the training set standard deviation, the behavior of the test set standard deviation was studied. In Fig. 3a, b, c and d, it can be seen that from the test set *S* perspective, the best models improve with increasing number of initial individuals; indicating that the new methodology does not appear to be overfitting the models to the training set data. External test set validations have also previously shown the prediction ability of models obtained by former versions of the methodology [19,23–26,35].

The potential overfitting of the models developed by the new methodology will depend mainly on the number of descriptors employed. An excessive number of descriptors with respect to the number of available experimental data will provoke an overfitting of the model to the training set molecules [36]. This was exemplified in Fig. 4, where the output of the ERM_p algorithm with increasing number of descriptors from 1 to 9 using the GI dataset is shown. It can be seen that at the beginning, the number of descriptors lowers the test set standard deviation until an optimal number (seven in this case) and then starts to increase as the models begin to overfit the training set data.

The results were summarized in Table 1, where it can be seen that increasing the number of initial sets significantly further improves the results of ERMfs. The percentage of improvement measured as $(S_{ERM} - S_{ERMp})/S_{ERMp}*100$ was, on average, 2.0% and the highest value was 4.8%, which is a considerable number since, in some cases, it is comparable to the improvement obtained by a unit increase in the number of descriptors in a model (*d*). These results reflect the fact that although ERM has been shown to give better results than GA and RM mainly due to the fact that it has a lower tendency to remain stock in a local *S* minimum, hence being more independent on the initial sets of randomly chosen descriptors, there is still some dependence on the initial set used to start the ERM procedure.

To further compare the new methodology (ERM_p) with the preceding one (ERM_{fs}), and to corroborate the predictive ability of the models obtained by ERM_p, the well-known leave-one-out cross-validation procedure (loo) [37] was used. According to the specialized literature,

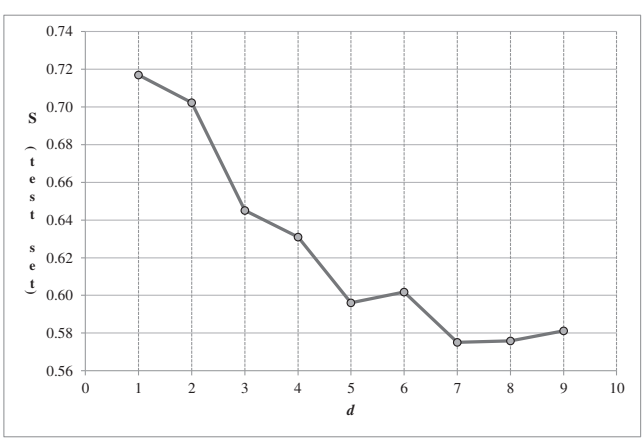

Fig. 4. Test set standard deviation of the best model vs. number of descriptors (*d*) included in the model for the GI dataset.

Table 1
Standard deviation of the previous version of ERM (S_{ERMfs}), the best found model with the proposed alternative (S_{ERMp}) and the relative improvement for the four datasets using *d* from 5 to 9.

	d	5	6	7	8	9
GI	S_{ERMfs}	0.4831	0.4550	0.4391	0.4276	0.4066
	S_{ERMp}	**0.4781**	**0.4523**	**0.4367**	**0.4177**	**0.4060**
	Improvement (%)	1.0%	0.6%	0.5%	2.4%	0.1%
FLUOR	S_{ERMfs}	0.5519	0.5052	0.4342	0.4167	0.3829
	S_{ERMp}	**0.5376**	**0.4939**	**0.4328**	**0.4012**	**0.3725**
	Improvement (%)	2.7%	2.3%	0.3%	3.8%	2.8%
MES	S_{ERMfs}	0.3230	0.3142	0.2930	0.2785	0.2648
	S_{ERMp}	**0.3189**	**0.3030**	**0.2845**	**0.2718**	**0.2592**
	Improvement (%)	1.3%	3.7%	3.0%	2.5%	2.2%
GABA	S_{ERMfs}	0.5129	0.4388	0.4046	0.3617	0.3405
	S_{ERMp}	**0.4896**	**0.4388**	**0.3929**	**0.3581**	**0.3346**
	Improvement (%)	4.8%	0.0%	3.0%	1.0%	1.8%

Boldface numbers indicate better results.

R_{loo} should be greater than 0.707 for a properly validated model [38]. The leave-one-out results were summarized in Table 2, where it can be appreciated that ERM_p outperforms or equals ERM_{fs} in terms of R_{loo} for all the cases except $d = 6$, 7 and 9 from the GI dataset.

The computational cost was also considerably increased, as can be seen in Table 3, the number of necessary cases to obtain the model with best *S* was on average 92, a much larger number than *d* (value used in ERM_{fs}). However, since the selection of initial sets is fortuitous, in some cases, the number to find the optimal model was much lower. Since this number cannot be determine beforehand, the best way to use the new algorithm might be as the follows:

- When sufficient time and computational power is present, the number of initial sets should be as high as possible
- If a reduction in the computational cost or calculation time is required, then all tests and calibrations should be done by using a low number of initial sets (*N* sets $= d \times 2$, is a good starting point).
- In all cases, once the optimal number of descriptors (*d*) for the dataset under investigation is chosen from the preliminary test, then an additional run only for the given *d* with a high number of initial sets (using *N* sets = 250 as reference) is recommended in order to further refine the results obtained in the preliminary runs.

6. Conclusions

In this paper, we studied the possible combinations between ERM and GA algorithms for the search of an optimal set of descriptors from a much bigger group. A single possible improvement was found by increasing the number of randomly initial sets of descriptors that start

Table 2
Correlation coefficient of leave-one-out cross validation of the previous version of ERM (S_{ERMfs}), the best found model with the proposed alternative (S_{ERMp}) and the relative improvement for the four dataset using *d* from 5 to 9.

	d	5	6	7	8	9
GI	S_{ERM}	0.8025	**0.8252**	**0.8385**	0.8452	**0.8602**
	S_{New}	**0.8054**	0.8231	0.8370	**0.8500**	0.8597
	Improvement (%)	0.4%	−0.3%	−0.2%	0.6%	−0.1%
FLUOR	S_{ERM}	0.9714	0.9771	0.9825	0.9840	0.9860
	S_{New}	**0.9745**	**0.9779**	**0.9829**	**0.9849**	**0.9872**
	Improvement (%)	0.3%	0.1%	0.04%	0.1%	0.1%
MES	S_{ERM}	0.6521	0.6606	0.7219	0.7470	0.7699
	S_{New}	**0.6582**	**0.6976**	**0.7378**	**0.7639**	**0.7907**
	Improvement (%)	0.9%	5.3%	2.2%	2.2%	2.6%
GABA	S_{ERM}	0.8499	0.8946	0.9081	0.9248	0.9346
	S_{New}	**0.8665**	0.8946	**0.9145**	**0.9281**	**0.9362**
	Improvement (%)	1.9%	0.0%	0.7%	0.3%	0.2%

Boldface numbers indicate better results.

Table 3
Number of initial sets for ERMfs and ERMp and the additional computational cost expressed as the number of times ERMp is greater than ERMfs.

	N Sets ERMfs	5	6	7	8	9
GI	*N* Sets ERMp	**7**	**65**	**211**	**91**	**26**
	Additional cost (num. times)	1.4	10.8	30.1	11.4	2.9
FLUOR	*N* Sets ERMp	**106**	**70**	**53**	**15**	**197**
	Additional cost (num. times)	75.7	6.5	1.8	1.3	68.2
MES	*N* Sets ERMp	**130**	**21**	**134**	**127**	**91**
	Additional cost (num. times)	1.7	3.3	76.2	96.3	1.3
GABA	*N* Sets ERMp	**247**	**6**	**94**	**92**	**54**
	Additional cost (num. times)	143.9	1.8	1.2	1.0	40.5

ERM, which resembles the initial population in GA. The new algorithm, named ERM_p, showed improved results over the previously existing ones in exchange of adding computational cost to the calculations. For that reason, it is recommended that during preliminary tests on QSAR application studies, the number of sets is kept to a minimum to later refine the models by increasing this number. The new alternative further develops ERM, which has been shown to give better results than GA, for the selection of an optimal set of descriptors in QSAR.

Acknowledgments

The authors want to thank the National Research Council of Argentina (CONICET, project PIP11220100100151) and the National University of La Plata.

Appendix A. Supplementary data

The compound names and their corresponding experimental values for the four data sets are available as supporting information; including marks showing the distribution between test set and training set. Supplementary data to this article can be found online at doi.org/10.1016/j.chemolab.2015.10.007

References

[1] C. Hansch, A. Leo, Exploring QSAR: Fundamentals and Applications in Chemistry and Biology, Am. Chem. Soc, Washington, D.C., 1995
[2] L. Shao, L. Wu, X. Fan, Y. Cheng, Consensus ranking approach to understanding the underlying mechanism with QSAR, J. Chem. Inf. Model. 50 (2010) 1941–1948.
[3] A.M. Wassermann, B. Nisius, M. Vogt, J.R. Bajorath, Identification of descriptors capturing compound class-specific features by mutual information analysis, J. Chem. Inf. Model. 50 (2010) (1935–1940).
[4] H. Yu, R. Kühne, R.-U. Ebert, G. Schüürmann, Comparative analysis of QSAR models for predicting pKa of organic oxygen acids and nitrogen bases from molecular structure, J. Chem. Inf. Model. 50 (2010) 1949–1960.
[5] E.A. Helgee, L. Carlsson, S. Boyer, U. Norinder, Evaluation of quantitative structure-activity relationship modeling strategies: local and global models, J. Chem. Inf. Model. 50 (2010) 677–689.
[6] S. Agarwal, D. Dugar, S. Sengupta, Ranking chemical structures for drug discovery: a new machine learning approach, J. Chem. Inf. Model. 50 (2010) 716–731.
[7] R. Todeschini, V. Consonni, Handbook of Molecular Descriptors, Wiley VCH, Weinheim, Germany, 2000.
[8] A.R. Katritzky, V.S. Lobanov, M. Karelson, QSPR: the correlation and quantitative prediction of chemical and physical properties from structure, Chem. Rev. Soc. 24 (1995) 279–287.
[9] N. Trinajstic, Chemical Graph Theory, CRC Press, Boca Raton, FL, 1992.
[10] P.R. Duchowicz, E.A. Castro, F.M. Fernández, M.P. González, A new search algorithm of QSPR/QSAR theories: normal boiling points of some organic molecules, Chem. Phys. Lett. 412 (2005) 376–380.
[11] P.R. Duchowicz, E.A. Castro, F.M. Fernández, Alternative algorithm for the search of an optimal set of descriptors in QSAR-QSPR studies, MATCH Commun. Math. Comput. Chem. 55 (2006) 179–192.
[12] A.G. Mercader, P.R. Duchowicz, F.M. Fernandez, E.A. Castro, Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories, Chemom. Intell. Lab. Syst. 92 (2008) 138–144.
[13] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Advances in the replacement and enhanced replacement method in QSAR and QSPR theories, J. Chem. Inf. Model. 51 (2011) 1575–1581.
[14] N.R. Draper, H. Smith, Applied Regression Analysis, John Wiley & Sons, New York, 1981.
[15] S.-S. So, M. Karplus, Evolutionary optimization in quantitative structure-activity relationship: an application of genetic neural networks, J. Med. Chem. 39 (1996) 1521–1530.
[16] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, Replacement method and enhanced replacement method versus the genetic algorithm approach for the selection of molecular descriptors in QSPR/QSAR theories, J. Chem. Inf. Model. 50 (2010) 1542–1548.
[17] Matlab, in, 5.0 The MathWorks Inc. http://www.mathworks.com/.
[18] P.R. Duchowicz, M. Fernández, J. Caballero, E.A. Castro, F.M. Fernández, QSAR of non-nucleoside inhibitors of HIV-1 reverse transcriptase, Bioorg. Med. Chem. 14 (2006) 5876–5889.
[19] P.R. Duchowicz, M.P. González, A.M. Helguera, M.N.D.S. Cordeiro, E.A. Castro, Application of the replacement method as novel variable selection in QSPR. 2. Soil sorption coefficients, Chemom. Intell. Lab. Syst. 88 (2007) 197–203.
[20] P.R. Duchowicz, A.G. Mercader, F.M. Fernández, E.A. Castro, Prediction of aqueous toxicity for heterogeneous phenol derivatives by QSAR, Chemom. Intell. Lab. Syst. 90 (2008) 97–107.
[21] P.R. Duchowicz, M.G. Vitale, E.A. Castro, J.C. Autino, G.P. Romanelli, D.O. Bennardi, QSAR modeling of the interaction of flavonoids with GABA(A) receptor, Eur. J. Med. Chem. 43 (2007) 1593–1602.
[22] P.R.F. Duchowicz, M., J. Caballero, E.A. Castro, F.M. Fernández, QSAR of non-nucleoside inhibitors of HIV-1 reverse transcriptase, Bioorg. Med. Chem. 14 (2006) 5876–5889.
[23] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, D.O. Bennardi, J.C. Autino, G.P. Romanelli, QSAR prediction of inhibition of aldose reductase for flavonoids, Bioorg. Med. Chem. 16 (2008) 7470–7476.
[24] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, F.M. Cabrerizo, A.H. Thomas, Predictive modeling of the total deactivation rate constant of singlet oxygen by heterocyclic compounds, J. Mol. Graph. Model. 28 (2009) 12–19.
[25] A.G. Mercader, P.R. Duchowicz, F.M. Fernández, E.A. Castro, E. Wolcan, QSPR Study of solvent quenching of the $^5D_0 \rightarrow {}^7F_2$ emission of Eu(6,6,7,7,8,8,8-heptafluoro-2,2-dimethyl-3,5-octanedionate)$_3$, Chem. Phys. Lett. 462 (2008) 352–357.
[26] A.G. Mercader, P.R. Duchowicz, M.A. Sanservino, F.M. Fernandez, E.A. Castro, QSPR analysis of fluorophilicity for organic compounds, J. Fluor. Chem. 128 (2007) 484–492.
[27] M. Melanie, An Introduction to Genetic Algorithms, A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England, 1998 (3-9,130 -131).
[28] A. Chipperfield, P. Fleming, H. Pohlheim, C. Fonseca, Genetic Algorithm TOOLBOX For Use with MATLAB User's Guide v1.2, in, Sheffield, http://www.shef.ac.uk/acse/research/ecrg/gat.html1994.
[29] Juan C. Garro Martinez, P.R. Duchowicz, M.R. Estrada, G.N. Zamarbide, E.A. Castro, Anticonvulsant activity of ringed enaminones: a QSAR study, QSAR Comb. Sci. 28 (2009) 1376–1385.
[30] L. Kaufman, P.J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis, Wiley-Interscience, New York, 2005.
[31] N.L. Allinger, Y.H. Yuh, J.H. Lii, Molecular mechanics. The MM3 force field for hydrocarbons, J. Am. Chem. Soc. 111 (1989) 8551–8566.
[32] HYPERCHEM, in, 6.03 (Hypercube) http://www.hyper.com.
[33] J.J.P. Stewart, Optimization of parameters for semiempirical methods I, J. Comput. Chem. 10 (1989) 209–220.
[34] DRAGON, in, release 5.0 Evaluation Version http://michem.disat.unimib.it/chm/.
[35] A.M. Helguera, P.R. Duchowicz, M.A.C. Pérez, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Application of the replacement method as novel variable selection strategy in QSAR. 1. Carcinogenic potential, Chemom. Intell. Lab. 81 (2006) 180–187.
[36] D.M. Hawkins, The problem of overfitting, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.
[37] D.M. Hawkins, S.C. Basak, D. Mills, Assessing model fit by cross-validation, J. Chem. Inf. Model. 43 (2003) 579–586.
[38] A. Golbraikh, A. Tropsha, Beware of q2! J. Mol. Graph. Model. 20 (2002) 269–276.