Chemometrics and Intelligent Laboratory Systems 92 (2008) 138–144

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Modified and enhanced replacement method for the selection of molecular descriptors in QSAR and QSPR theories

Andrew G. Mercader *, Pablo R. Duchowicz, Francisco M. Fernández, Eduardo A. Castro

INIFTA (UNLP, CCT La Plata-CONICET), Diag. 113 y 64 (S/N), Sucursal 4, Casilla de Correo 16, 1900 La Plata, Argentina

ARTICLE INFO

Article history:
Received 9 January 2008
Received in revised form 13 February 2008
Accepted 17 February 2008
Available online 29 February 2008

Keywords:
QSPR-QSAR Studies
Multivariable Linear Regression
Replacement Method
Full Search
Simulated Annealing Methodology

ABSTRACT

We improve a recently developed Replacement Method (RM) for the selection of an optimal set of molecular descriptors from a much greater pool of such regression variables. Our approach yields almost optimal results with a much smaller number of linear regressions than the full search. We test our method on four different experimental full data sets and four sub datasets. The resulting algorithm, which was named Enhanced Replacement Method (ERM), resembles a simulated annealing procedure and improves our RM, yielding models with better statistical parameters than the ones previously published. The number of linear regressions increases only to a small extent so that the new algorithm is still suitable for databases with as many as 63912 descriptors.

1. Introduction

A generally accepted remedy 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 exists a permanently renewed interest focused on the development of such kind of predictive techniques [2–5]. The ultimate role of the QSPR/QSAR theory is to suggest mathematical models capable of estimating relevant properties of interest, especially when those cannot be experimentally determined for some reason. Such studies rely on the basic assumption that the structure of a compound determines entirely its properties, which can therefore be translated into so-called molecular descriptors. These parameters are calculated through mathematical formulae obtained from several theories, such as Chemical Graph Theory, Information Theory, Quantum Mechanics, etc [6,7].

Nowadays, there are thousands of descriptors available in the literature [8], and one has to decide how to select those that characterize the property/activity under consideration in the most efficient way. One is thus faced to the mathematical problem of selecting a subset $\mathbf{d}$ of d descriptors from a much larger set $\mathbf{D}$ of $D \gg d$ ones.

The search for the optimal set of descriptors may be monitored by the minimization or maximization of a chosen statistical parameter; for example, we may be interested in a model that makes the Standard Deviation (S) as small as possible. In other words, we look for the global minimum of $S(\mathbf{d})$, where $\mathbf{d}$ is a point in a space of $D!/[d!(D-d)!]$ ones. Consequently, a full search (FS) of the optimal variables is impractical because it requires $D!/[d!(D-d)!]$ linear regressions.

Some time ago we proposed the Replacement Method (RM) [9–11] that produces linear QSPR–QSAR models that are quite close the FS ones with much less computational work. This technique approaches the minimum of S by judiciously taking into account the relative errors of the coefficients of the least-squares model given by a set of d descriptors $\mathbf{d}=\{X_1, X_2,\ldots, X_d\}$. It has been shown [12] that the RM gives models with better statistical parameters than the Forward Stepwise Regression (FSW) procedure [13] and similar or better ones than the more elaborated Genetics Algorithms (GA) [14]. We believe that the RM is preferable to the GA [11,15] because the former takes into account the error in the regression coefficient and as a result the replacement of the descriptor is not at random as in the GA. In addition to it, the practical application of the GA requires the tuning of some parameters such as mutation probability, crossover probability, generation gap, etc., which is not a simple problem [16].

The RM is a rapidly convergent iterative algorithm that produces linear regression models with small S in a remarkably little computer time [11,12,17]. However, in some cases, the RM can get trapped in a local minimum of S that is not able to leave without some kind of constraint. Although such local minima provide acceptable models, as shown in all earlier applications of the RM [11,12,17], there is still room for improvement.

In this paper we propose a Modified Replacement Method (MRM) that follows the same RM philosophy but exhibits less propensity for remaining in local minima and at the same time is less dependent on the initial solution.

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

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.02.005

We will also discuss the resemblance of the new algorithm with the Simulated Annealing (SA) which is an adaptation of the Metropolis–Hastings algorithm, a Monte Carlo Method [18] to generate sample states of a thermodynamic system. The name and inspiration 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 [19].

In Section 2 we discuss the data sets and develop the method. In Section 3 we apply the alternative algorithms to some QSPR problems and compare the results. Finally, in Section 4 we draw conclusions.

2. Experimental data and theoretical methods

2.1. Data sets

Four different experimental data sets already previously analyzed were used to test and contrast the performance of both RM and MRM: a fluorophilicity data set (FLUOR), consisting of 116 organic compounds characterized by 1268 theoretical descriptors [12]; 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 [17]; a GABA receptor data set (GABA), containing 78 inhibition data for flavone derivatives and 1187 molecular descriptors [20] and a 100 ED50 MES mice ip for enaminones (MES) with 1306 descriptors [21–23]. Additionally a data set of 209 Polychlorinated Biphenyls (PCB) with measured Relative Response Factor containing 63912 molecular descriptors [24] was used to test whether the application of the improved algorithm on an extremely large dataset is possible.

In all cases the structures of the compounds were firstly pre-optimized with the Molecular Mechanics Force Field (MM+) procedure included in Hyperchem version 6.03 [25], and the resulting geometries were further refined by means of the semi empirical method PM3 (Parametric Method-3) using the Polak–Ribiere algorithm and a gradient norm limit of 0.01 kcal/Å. More than a thousand molecular descriptors were calculated using the software Dragon 5.0 [26], including parameters of all types such as constitutional, topological, geometrical, quantum mechanical, etc. Most of the 62,873 descriptors of the last dataset were calculated by the molecular descriptors family methodology [24]. All the algorithms were programmed in the computer system Matlab 5.0 [27].

2.2. The algorithm

Since present algorithm is a slight variant of the RM [9] we begin with the discussion of the latter. We have a large set $\mathbf{D}=\{X_1, X_2,\ldots, X_D\}$ of D descriptors provided by some available commercial program. It is our purpose to choose an optimal subset $\mathbf{d}_m=\{X_{m1}, X_{m2},\ldots, X_{md}\}$ of $d \ll D$ descriptors with 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 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 enable us to arrive at the global minimum, but it is computationally prohibitive if D is sufficiently large. The RM consists of the following steps:

- 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. That is what we define as a 'step'
- From this resulting set we choose the descriptor with the greatest standard deviation in its coefficient (we do not consider the one changed previously) 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 comes from the so-constructed path i.
- 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 we do not change that descriptor.
- We carry out the process above for all the possible paths $i=1, 2,\ldots, d$ and keep the point $\mathbf{d}_m$ with the smallest standard deviation: $\min{}_i S\left(\mathbf{d}_m^{(i)}\right)$.

The MRM follows the same 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 choose the next smallest value of S). The MRM converges to different solutions and commonly bounces from one point to another, occasionally repeating

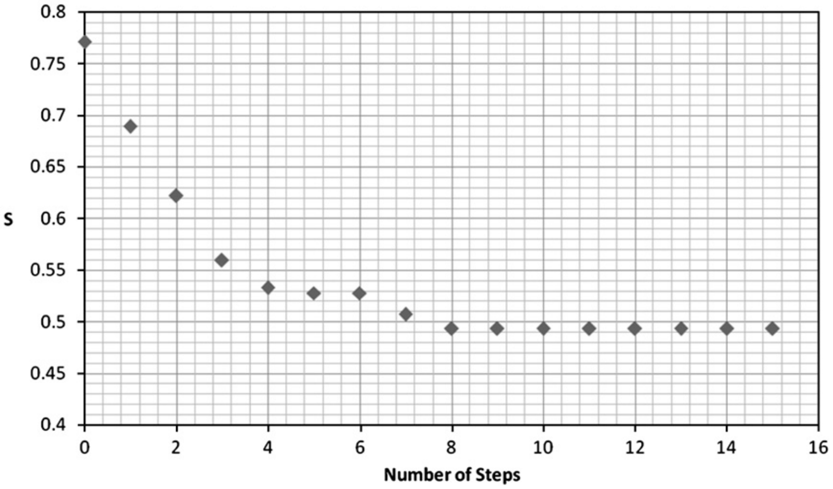

Fig. 1. Standard deviation vs. number of steps of the RM.

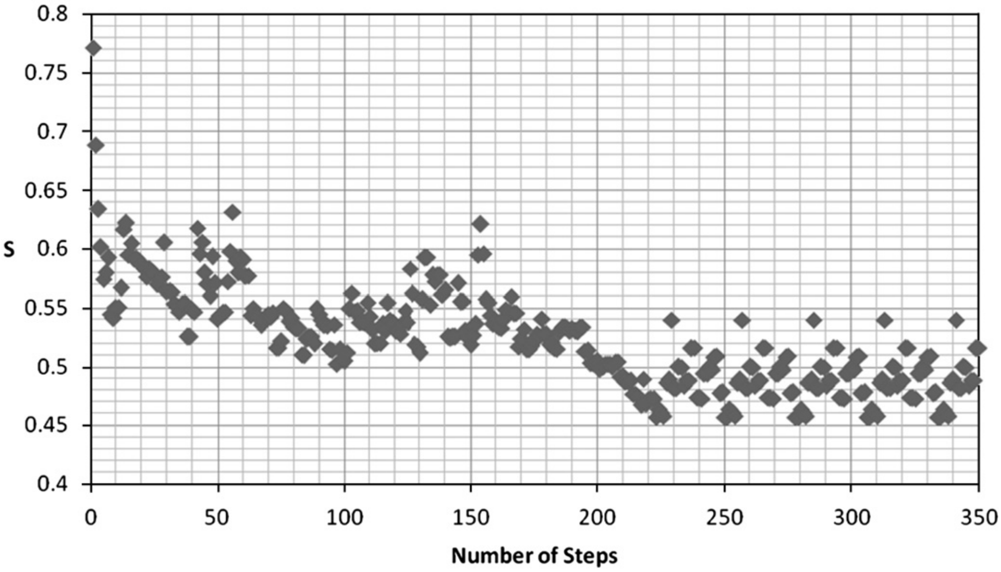

Fig. 2. Standard deviation vs. number of steps of the MRM.

some of them; in such a case we find that a plausible solution is the first one that appears four times.

If convergence is too slow we stop the process after 350 steps, which does not lead to a great loss because the resulting *S* is always sufficiently small.

A descriptive example was included in the Appendix A to illustrate the difference between MRM and RM.

3. Results and discussion

With the purpose of providing a graphical visualization of the behavior of our two algorithms, Figs. 1 and 2 show *S* as a function of the number of steps for both RM and MRM, respectively, and for the optimization of a seven-parameter model using the FLUOR data set [12]. The graphs reveal that the MRM simulates a higher temperature or 'a higher noise' than the RM, although maintaining the overall decreasing tendency of the *S* function. This apparent thermal agitation makes the MRM less likely to get trapped by a local minimum at the cost of slower convergence and more computer time.

The behaviour of the RM and MRM shown in the aforementioned figures suggested us to implement a further optimization routine that integrates the two algorithms, associating the MRM to a thermal agitation of the RM. We tried the following combinations: MRM-RM, RM-MRM and RM-MRM-RM. For instance, when RM is applied after MRM (MRM-RM), the starting solution for RM consists of the best set of variables obtained from the previous application of the MRM algorithm. After several runs we decided to discard the MRM-RM-MRM simply because it increases the number of linear regressions significantly without achieving appreciable improvements in the statistical results. As illustrative examples, Figs. 3–5 display *S* vs. the number of steps for the three optimization cases, resorting again to the FLUOR dataset.

In order to carry out a FS in a reasonable time we selected 75 molecular descriptors from the pool and thus reduced the set **D** to just those $D=75$ variables. We then applied the search algorithms described in the preceding section and obtained the optimal sets of $d=1, 2,\ldots, 7$ descriptors with the same random initial solution for each of them. All the models include the constant term. In this way we can compare our approximate search algorithms with the exact FS. Our results are summarized in Table 1 that compares the minimal values of *S* obtained by all those algorithms. The exact minimal value of *S* appears in boldface to facilitate the comparison.

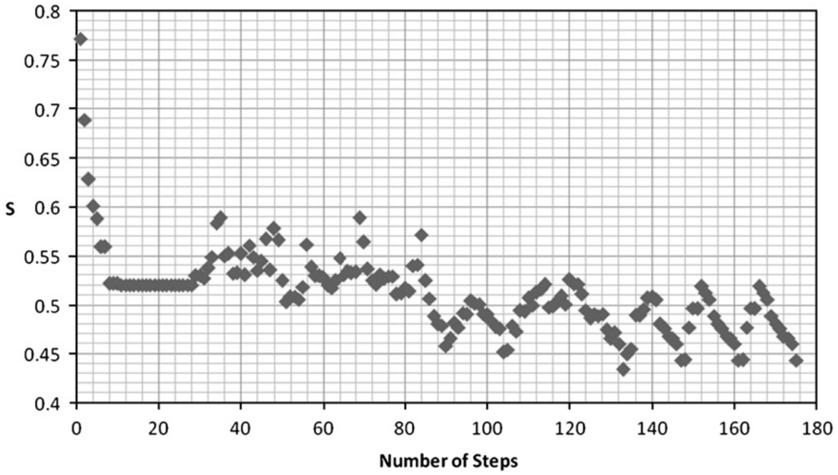

Fig. 3. Standard deviation vs. number of steps of the RM-MRM.

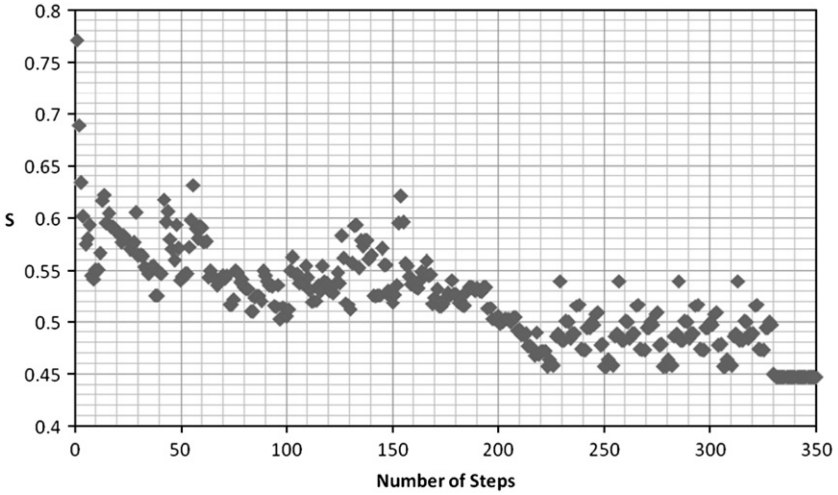

Fig. 4. Standard deviation vs. number of steps of the MRM-RM.

It follows from Table 1 that S_{RM} either agrees or is in close agreement with S_{FS}. It can also be appreciated that for all cases MRM provides better results than RM and can be further improved by the different alternation options; in fact, the RM-MRM-RM composite appears to produce the best results. The number of linear regressions and computation time (shown as an average at the end of Table 1) required for the alternative algorithms are greater than the RM ones but they remain smaller than the FS calculations for $d>2$. The difference between the number of calculations for the approximate algorithms and the FS increases as D increases. Remember that we chose $D=75$ for this manageable benchmark experiment, but in actual applications $D>1000$.

In what follows we apply all the algorithms, the RM and its improvements, to real-life problems; in this case the four full datasets. It is not possible for us to carry out a FS because even for the smallest database (GABA, $D=1187$) the number of linear regressions for $d=7$ amounts to 6.47×10^{17} that would take about 3.4×10^{6} years in a PC with an AMD Athlon 64 2800+ processor. Even in a much more powerful computer the solution would not be reached in a reasonable time.

We carried out all the numerical tests for $d=7$ as an example of a high computational demanding search with a reasonable number of descriptors for a potential model in common QSPR/QSAR studies. It should be mentioned that in the application of the method in QSAR/QSPR studies, models with increasing number of descriptors are easily searched using the algorithm and the optimal d is afterward determined using a criteria that selects the model with better statistical parameters and at the same time avoids models that overfit the data [17]. The selection of d optimal descriptors for all the databases and algorithms is not presented in this work for space reasons. Another thing to consider in practical use of the algorithm is the correlation of the descriptors that can be easily avoided by taking out of the pool those descriptors that have a correlation higher than a set limit.

Since the approximate optimal models normally depend on the initial set of descriptors, we chose the same three random initial sets for all the algorithms and show the results in Table 2. The last column of Table 2 shows the results provided by the well-known FSW regression algorithm [13] as an external comparison point. This procedure consists of a step-by-step addition of descriptors to the model, initially without any independent variable, until there is no variable left outside the equation that minimizes the value of S. No initial set is necessary for this approach.

Table 2 also shows the average percentage improvement over the RM to facilitate visualization of the performance of the alternative algorithms proposed in this paper. At the end of the table we see the ratio of the number of linear regressions for the new algorithms with respect to the RM ones.

As a theoretical validation of the models we used the well-known Leave-One-Out (*loo*) [28], the results can be appreciated between

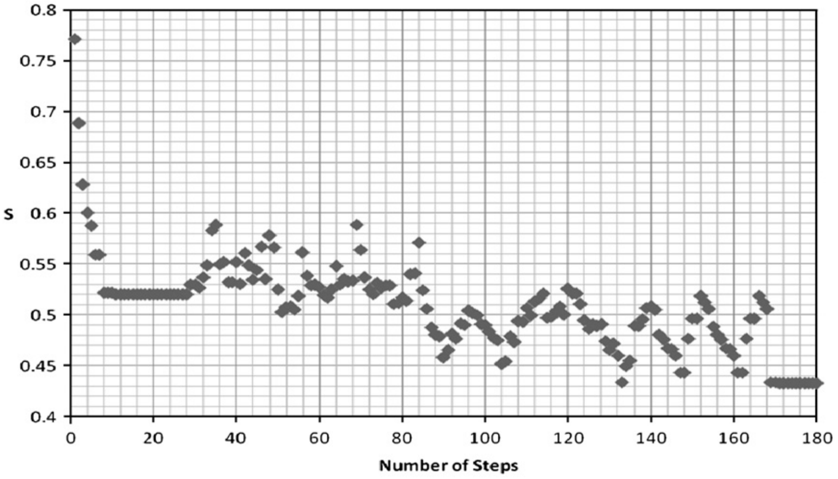

Fig. 5. Standard deviation vs. number of steps of the ERM (RM-MRM-RM).

Table 1
Standard deviation (*S*), number of linear regressions and computation time for the FS, RM, MRM, RM-MRM, MRM-RM and RM-MRM-RM, for four sub data sets of *D*=75 descriptors. The bar "/" separates algorithms that give identical results

Algorithm	*S*						
d	1	2	3	4	5	6	7
				MES			
FS	**0.3 991**	**0. 3666**	**0 .3536**	**0.3443**	**0.3361**	**0 .3254**	**0.3169**
RM	**0.3 991**	**0. 3666**	**0 .3536**	0.3480	**0.3361**	0 .3327	0.3268
MRM/ MRM-RM	**0.3 991**	**0. 3666**	**0 .3536**	**0.3443**	**0.3361**	0 .3290	**0.3169**
RM-MRM/ RM-MRM-RM	**0.3 991**	**0. 3666**	**0 .3536**	**0.3443**	**0.3361**	**0 .3254**	**0.3169**
				GI			
FS	**0.6494**	**0.6000**	**0.5693**	**0.5605**	**0.5487**	**0.5324**	**0.5214**
RM	**0.6494**	**0.6000**	0.5875	**0.5605**	0.5512	0.5415	0.5350
MRM	**0.6494**	**0.6000**	**0.5693**	**0.5605**	0.5492	**0.5324**	**0.5214**
RM-MRM	**0.6494**	**0.6000**	0.5875	**0.5605**	0.5492	0.5350	**0.5214**
MRM-RM	**0.6494**	**0.6000**	**0.5693**	**0.5605**	0.5492	**0.5324**	**0.5214**
RM-MRM-RM	**0.6494**	**0.6000**	0.5875	**0.5605**	0.5492	**0.5324**	**0.5214**
				FLUOR			
FS	**1.1192**	**0. 7587**	**0.7294**	**0.6901**	**0.6440**	**0.6200**	**0.5971**
RM	**1.1192**	0. 7891	0.7329	**0.6901**	0.6549	0.6451	0.6253
MRM/Rest	**1.1192**	**0. 7587**	**0.7329**	**0.6901**	**0.6440**	**0.6200**	**0.5971**
				GABA			
FS	**0.8289**	**0.7335**	**0.6421**	**0.5918**	**0.5719**	**0.5383**	**0.5083**
RM	**0.8289**	**0.7335**	**0.6421**	**0.5918**	**0.5719**	0.5398	0.5120
MRM	**0.8289**	**0.7335**	**0.6421**	**0.5918**	**0.5719**	**0.5383**	0.5088
RM-MRM	**0.8289**	**0.7335**	**0.6421**	**0.5918**	**0.5719**	0.5398	0.5120
MRM-RM/ RM-MRM-RM	**0.8289**	**0.7335**	**0.6421**	**0.5918**	**0.5719**	**0.5383**	**0.5083**
				Average Number of linear regressions			
FS	75	2775	67,525	1.22E+06	1.73E+07	2.01E+08	1.98E+09
RM	75	1260	2850	4756	7638	10,086	14,739
MRM	75	4674	16,095	89,464	81,380	110,958	240,149
RM-MRM/ MRM-RM	75	5934	18,945	94,220	89,018	121,044	254,888
RM-MRM-RM	75	7194	21,795	98,976	96,655	131,130	269,627
				Average computation time in minutes*			
FS	2.07E−04	7.67E−03	1.87E−01	3.36E+00	4.77E+01	5.57E+02	5.49E+03
RM	2.07E−04	3.48E−03	7.88E−03	1.31E−02	2.11E−02	2.79E−02	4.07E−02
MRM	2.07E−04	1.29E−02	4.45E−02	2.47E−01	2.25E−01	3.07E−01	6.64E−01
RM-MRM/ MRM-RM	2.07E−04	1.64E−02	5.24E−02	2.60E−01	2.46E−01	3.35E−01	7.05E−01
RM-MRM-RM	2.07E−04	1.99E−02	6.02E−02	2.74E−01	2.67E−01	3.62E−01	7.45E−01

FS results are given in boldface numbers.
* Using an AMD Athlon 64 2800+ processor.

Table 2
Standard deviation, R from Leave-One-Out validation (between parentheses), number of linear regressions and computational time (between parentheses) for the RM, MRM, RM-MRM, MRM-RM, RM-MRM-RM, and FSR, for the four full data sets with three different initial seven-descriptor sets

Algorithm	**RM**	**MRM**	**RM-MRM**	**MRM-RM**	**RM-MRM-RM**	**FSR**
			***S* (*R*$_{loo}$)**			
MES	0.3089 (0.685)	**0.2896 (0.726)**	0.2919 (0.722)	**0.2896 (0.726)**	0.2919 (0.722)	0.3409 (———)
	0.3077 (0.692)	**0.2973 (0.722)**	**0.2973 (0.722)**	**0.2973 (0.722)**	**0.3209 (0.722)**	
	0.3008 (0.695)	0.2954 (0.710)	**0.2896 (0.726)**	0.2951 (0.710)	**0.2896 (0.726)**	
GI	**0.4421 (0.835)**	**0.4421 (0.835)**	**0.4421 (0.835)**	**0.4421 (0.835)**	**0.4421 (0.835)**	0.4937 (0.789)
	0.4648 (0.821)	0.4421 (0.835)	0.4367 (0.837)	0.4421 (0.835)	**0.4367 (0.837)**	
	0.4445 (0.835)	**0.4445 (0.835)**	**0.4445 (0.835)**	**0.4445 (0.835)**	**0.4445 (0.835)**	
GABA	0.4465 (0.891)	**0.4045 (0.91)**	0.4269 (0.898)	**0.4045 (0.91)**	0.4142 (0.903)	0.46797 (0.876)
	0.4683 (0.878)	**0.3961 (0.912)**	0.4121 (0.903)	**0.3961 (0.912)**	**0.3961 (0.912)**	
	0.4160 (0.905)	**0.3961 (0.912)**	**0.3961 (0.912)**	**0.3961 (0.912)**	**0.3961 (0.912)**	
FLUOR	0.4936 (———)	0.4572 (0.981)	0.4339 (0.983)	0.4470 (0.982)	**0.4328 (0.983)**	0.5718 (0.970)
	0.4328 (0.983)	0.4647 (0.981)	**0.4328 (0.983)**	0.4606 (0.981)	**0.4328 (0.983)**	
	0.4985 (0.976)	0.4426 (0.983)	0.4619 (0.979)	**0.4408 (0.983)**	0.4470 (0.982)	
Average Improvement	**0% (0%)**	**4.76% (1.8%)**	**4.94% (1.8%)**	**5.05% (1.9%)**	**5.73% (2.0%)**	**−10.31% (−2.5%)**
			Number of linear regression s (Computation time in minutes*)			
Average	283878 (0.78)	1629938 (4.51)	1725873 (4.77)	1828165 (5.05)	1926775 (5.33)	8923 (0.02)
Ratio	1	5.74	6.08	6.44	6.79	0.031

The best solutions appear in boldface numbers.
*Using an AMD Athlon 64 2800+ processor.

parenthesis in Tables 2 and 3. There are two cases (one on RM and the other on FSR) that are impossible to calculate since they present problems in the implementation of the *loo* methodology. This is an extra proof of the superiority of the new methods over RM and FSR (the average improvement percentage based on *loo* does not take this into account). Additional validation methods as Leave-More-Out Cross-Validation [28] and external test set validation have shown in the past the prediction ability of models obtained by the presented methodology [10–12,15,17,20] and were not used since they would have been extremely time consuming for the number of models employed in this work.

It follows from Table 2 that the RM gives better results than FSW, confirming previous comparative studies [12]. We appreciate that the MRM outperforms or equals the RM for all cases except for one of the initial solutions of the FLUOR dataset. This particular case appears to be fortuitous since the MRM is clearly better than the RM for the other two initial solutions of that dataset. It has to be kept in mind that the results of the approximate methods depend on the initial solutions, and, therefore, it is always possible that a method may give a smaller value of *S* than a supposedly better algorithm. What is more, there is low but nonzero probability that the poorer method may even hit the global minimum. The improvement percentage in Table 2 suggests that the proposed algorithm combinations are better than the RM alone. In particular, the sequence RM-MRM-RM emerges as the best

Table 3
Standard deviation, *R* from Leave-One-Out validation (between parentheses), number of linear regressions and computational time (between parentheses) for the RM and ERM for the PCB data set with three different initial solutions

Algorithm	**RM**	**ERM**
	***S* (*R*$_{loo}$)**	
PCB	**0.1616 (0.883)**	**0.1616 (0.883)**
	0.1718 (0.866)	**0.1616 (0.883)**
	0.1616 (0.883)	**0.1610 (0.884)**
Average Improvement	**0% (0%)**	**2.1% (0.7%)**
	Number of linear regressions (Computation time in minutes*)	
Average	1.42E+07 (39.25)	7.42E+07 (205.18)
Ratio	1	5.23

The best solutions appear in boldface numbers.
*Using an AMD Athlon 64 2800+ processor.

Table 4
Evolution of the MRM. Number of the descriptors in the model with the corresponding relative errors in the regression coefficients, *S* and *R* for each step of the algorithm

Step No.	Descriptor number/relative errors of the regression coefficients								S	R
0	C	**1**	2	3	4	5	6	7	0.771	0.952
	28.29	90.12	38.95	59.59	20.36	194.94	84.91	50.21		
1	C	**1068**	2	3	4	5	6	7	0.689	0.962
	21.34	18.58	41.89	67.67	15.74	66.44	**796.66**	35.89		
2	C	1068	2	3	4	5	**40**	7	0.634	0.968
	15.62	16.69	31.43	**43.14**	10.58	35.56	22.60	27.24		
3	C	1068	2	**411**	4	5	40	7	0.602	0.971
	16.16	15.48	19.93	23.67	6.45	9.06	20.34	**82.96**		
4	C	1068	2	411	4	5	40	**697**	0.574	0.974
	8.76	17.92	**12.77**	18.99	5.29	7.07	18.49	28.74		
5	C	1068	**1110**	411	4	5	40	697	0.580	0.973
	9.24	23.11	13.12	18.97	6.49	**6.91**	21.53	23.96		
6	C	1068	1110	411	4	**394**	40	697	0.593	0.972
	6.50	26.58	10.69	16.79	**7.41**	7.14	23.55	15.56		
7	C	1068	204	411	**1050**	394	40	697	0.545	0.974
	45.13	**27.90**	25.83	21.09	4.84	5.85	13.76	12.49		
222	C	425	240	40	200	480	1095	256	0.457	0.984
	414.91	16.10	8.35	6.05	3.37	10.05	11.34	12.30		

C stands for regression constant.

algorithm which we will call Enhanced Replacement Method (ERM) from now on. This conclusion is in line with the idea that the ERM is the only algorithm that goes through a complete simulated annealing cycle [19], as shown by Fig. 5. The ERM computational demand is comparable to the MRM one and is almost seven times greater than the RM one. This is the price we have to pay for obtaining QSPR models with better statistical parameters than the ones obtained previously [12].

Table 2 suggests that the ERM results are less sensitive to the initial point than the RM ones. However, the ERM solutions also depend on the starting point and we plan to study this aspect of the algorithm in the future.

As a further test of the ERM on a much more demanding problem, we tried it on the PCB database than contains as many as 63912 descriptors [24]. The ERM converged in a reasonable time and Table 3 shows the results. As expected the ERM gave smaller values of *S* for the same three random initial solutions chosen in the preceding tests.

4. Conclusions

In this paper we propose an improvement on the RM [9–11], which we call MRM, as well as some composite algorithms that resemble a simulated annealing [19]. The most efficient one appears to be ERM=RM-MRM-RM that yields better statistical parameters and is less sensitive to the starting point of the iterative procedure. The greater computational demand of this new algorithm does not appear to counterbalance its advantages and we plan to try it on many interesting problems in the future. Here, we have improved previous results derived earlier from the RM for real problems [12]. In order to show that the greater number of necessary linear regressions is not an obstacle for the application of the ERM to actual problems of chemical and biological interest we tested its performance on a large dataset of 63912 descriptors provided by Jäntschi [24].

Table 5
Information about the descriptors in the best model found in the example shown in the Appendix A

Descriptor			
Number	Name	Type	Meaning
X_{425}	MATS1p	2D Autocorrelations	Moran autocorrelation−lag 1/weighted by atomic polarizabilities
X_{240}	piPC03	Topological	Molecular multiple path count of order 03
X_{40}	IAC	Topological	Total information index of atomic composition
X_{200}	SEigv	Topological	Eigenvalue sum from van der Waals weighted distance matrix
X_{480}	AROM	Aromaticity indices	Aromaticity (trial)
X_{1095}	R3u+	GETAWAY	R maximal autocorrelation of lag 3/unweighted
X_{256}	D/Dr10	Topological	Distance/detour ring index of order 10

Acknowledgment

One of us (AGM) would like to thank the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) for a research fellowship.

Appendix A

In order to illustrate the difference between MRM and RM we apply them to the fluorophilicity data set (FLUOR), that consists of 116 organic compounds characterized by 1268 theoretical descriptors. We will obtain the optimal model with $d=7$ topological descriptors out of the pool of $D=1268$ ones.

We arbitrarily choose the initial set $\mathbf{d}=\{X_1, X_2, X_3, X_4, X_5, X_6, X_7\}$ which yields $S_{(0)}=0.771$ and follow path 1 that leads to the results in Fig. 2.

Table 4 displays a summary of the procedure where one can see the relative error of the regression coefficients for de descriptors and regression constant (*C*), and how *S* decreases and *R* increases in each step of the algorithm.

In path 1 we first change X_1; each change is indicated by the notation $(X_{\mathrm{old}}, X_{\mathrm{new}})$ Of all the 1261 $(D-d)$ variables, the substitution that minimizes *S* is (X_1, X_{1068}) yielding $S(1)=0.689$.

We now replace the variable with the greatest relative error X_6 with all the 1261descriptors (X_{1068} is now out of the descriptor pool and X_1 is in it) and find that the substitution (X_6, X_{40}) that yields the smallest standard deviation $S(2)=0.634$.

Now the variable with greatest relative error is X_3. After its replacement by all the 1261 descriptors, we conclude that the substitution (X_3, X_{411}) yields the minimal value $S(3)=0.602$.

In the following step the variable with greatest relative error is X_7 and after its replacement by all the 1261 descriptors, we have (X_7, X_{697}) and $S(4)=0.574$.

Of all the variables not yet replaced, X_2 is the one with the largest relative error. The replacement by all the 1261 descriptors does not lead to a model with lower *S*.

Up to this point MRM and RM have exactly the same behavior. From now on their difference will become visible.

First we will describe how RM would have continued. Since the replacement of X_2 did not lead to a model with lower S, X_2 remains in its position and is not replaced. Exactly the same situation occurs with the next descriptors X_5 and X_4. Restarting the process once again does not lead to a model with lower S, so the best model found in this case yields $S(4)=0.574$.

Now we will continue with the MRM. Even if the replacement of X_2 does not lead to a model with lower S, the descriptor is replaced anyway by the descriptor that leads to the lowest S from the 1261 remaining descriptors; thus we have (X_2, X_{1110}) with $S(5)=0.580$. Notice that in this step S has increased slightly. As will be seen in the next steps this is far from being a problem since an even lower S will be found later on, showing that the increase in S was necessary to get out of a local S minimum.

In the next step we once again find that the replacement of the descriptor X_5 with higher error in the coefficient that was not previously replaced by all the 1261 descriptors leads to a substitution (X_5, X_{394}) that yields an even higher standard deviation $S(6)=0.593$.

Nevertheless in the next step the replacement of X_4 (the descriptor with higher error in the coefficient that remains untouched) by all the 1261 descriptors leads to a substitution (X_4, X_{1050}) that yields $S(7)=0.545$ which is even lower than the local minimum found in step four: $S(4)=0.574$.

As the procedure continues, S continues the decreasing tendency, as can be seen in Fig. 2, in this case arriving to the lowest value after 222 steps. The best model found yields $S(222)=0.4572$. and $R=0.9835$, having the form:

$$\begin{aligned}\ln P = {} & 0.065(\pm 0.3) - 3.9029(\pm 0.6)X_{425} - 0.054(\pm 0.005)X_{240} \\ & - 0.063(\pm 0.004)X_{40} - 0.3749(\pm 0.01)X_{200} + 1.7051(\pm 0.2)X_{480} \\ & - 23.6913(\pm 2.7)X_{1095} + 0.008(\pm 0.001)X_{256}\end{aligned}$$

The molecular descriptors appearing in the equation combine several two- and three-dimensional aspects of the molecular structure, and can be classified as a 2D Autocorrelations, four Topological descriptors, an Aromaticity Index and a GETAWAY descriptor [26]. The names of this descriptors and their meanings a can be found in Table 5.

References

[1] C. Hansch and A. Leo. American Chemical Society, Washington, D. C., 1995.
[2] W.A. Sexton. 115 D. Van Nostrand, New York, 1950.
[3] C. Hansch, T. Fujita, J. Am. Chem. Soc. 86 (1964) 1616–1626.
[4] C. Hansch, Acc. Chem. Res. 2 (1969) 232–239.
[5] R.B. King, Ed., 28 Elsevier, Amsterdam, 1983.
[6] A.R. Katritzky, V.S. Lobanov, M. Karelson, Chem. Soc. Rev. 24 (1995) 279–287.
[7] N. Trinajstic. CRC Press, Boca Raton, FL, 1992.
[8] R. Todeschini and V. Consonni. Wiley VCH, Weinheim, Germany, 2000.
[9] P.R. Duchowicz, E.A. Castro, F.M. Fernández, MATCH Commun. Math. Comput. Chem. 55 (2006) 179–192.
[10] P.R. Duchowicz, M. Fernández, J. Caballero, E.A. Castro, F.M. Fernández, Bioorg. Med. Chem. 14 (2006) 5876–5889.
[11] A.M. Helguera, P.R. Duchowicz, M.A.C. Pérez, E.A. Castro, M.N.D.S. Cordeiro, M.P. González, Chemom. Intell. Lab. Syst. 81 (2006) 180–187.
[12] A.G. Mercader, P.R. Duchowicz, M.A. Sanservino, F.M. Fernandez, E.A. Castro, J. Fluorine Chem. 128 (2007) 484–492.
[13] N.R. Draper and H. Smith. John Wiley&Sons, New York, 1981.
[14] S.S. So, M. Karplus, J. Med. Chem. 39 (1996) 1521–1530.
[15] P.R. Duchowicz, M.P. González, A.M. Helguera, M.N.D.S. Cordeiro, E.A. Castro, Chemom. Intell. Lab. Syst. 88 (2007) 197–203.
[16] M. Melanie, A Bradford Book, The MIT Press, Cambridge, Massachusetts, 1998.
[17] P.R. Duchowicz, A.G. Mercader, F.M. Fernández, E.A. Castro, Chemom. Intell. Lab. Syst. 90 (2007) 97–107.
[18] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, J. Chem. Phys. 21 (1953) 1087–1092.
[19] S. Kirkpatrick, C.D. Gelatt Jr., M.P. Vecchi, Science 220 (1983) 671–680.
[20] P.R. Duchowicz, M.G. Vitale, E.A. Castro, J.C. Autino, G.P. Romanelli and D.O. Bennardi, Eur. J. Med. Chem., (in press), doi:10.1016/j.ejmech.2007.11.009.
[21] I.O. Edafiogho, C.N. Hinko, H. Chang, J.A. Moore, D. Mulzac, J.M. Nicholson, K.R. Scott, J. Med. Chem. 35 (1992) 2798–2805.
[22] N.D. Eddington, D.S. Cox, M. Khurana, N.N. Salama, J.P. Stables, S.J. Harrison, A. Negussie, R.S. Taylor, U.Q. Tran, J.A. Moore, J.C. Barrow, K.R. Scott, Eur. J. Med. Chem. 38 (2003) 49–64.
[23] I.O. Edafiogho, K.V.V. Ananthalakshmi, S.B. Kombian, Bioorganic Med. Chem. 14 (2006) 5266–5272.
[24] L. Jäntschi, Leonardo Electronic, J. Pract. Technol. 3 (2007) 67–84.
[25] HYPERCHEM 6.03 (Hypercube) http://www.hyper.com/.
[26] DRAGON 5.0 Evaluation Version http://www.disat.unimib.it/chm.
[27] Matlab 5.0 The MathWorks Inc. http://www.mathworks.com/.
[28] D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Model. 43 (2003) 579–586.