MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **55** (2006) 179-192

ISSN 0340 - 6253

Alternative Algorithm for the Search of an Optimal Set of Descriptors in QSAR-QSPR Studies

Pablo R. Duchowicz[*]*, Eduardo A. Castro and Francisco M. Fernández*

INIFTA, División Química Teórica, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, C.C. 16, (1900) La Plata, Argentina

(Received April 15, 2005)

We propose an algorithm for the search of an optimal set of descriptors from a pool of such regression variables. Our approach requires a smaller number of linear regressions than the full search and produces almost identical results. As an illustrative example, we model the meltingpoints of 30 quinoxalines derivatives by means of several subsets of 14 descriptors each, generated by the Dragon 5 evaluation version.

---

[*] Corresponding author: phone: (+54)(221)425-7430, FAX number: (+54)(221)425-4642, e-mail: pabloducho@yahoo.com.ar or duchow@inifta.unlp.edu.ar

I. Introduction

During the last decades there has been great interest in the development of Quantitative Structure-Property/Activity Relationships QSPR/QSAR for the reliable prediction of physicochemical, biological and pharmacological properties of chemical compounds, solely from the knowledge of their molecular structure. Such relationships are most welcome when the experimental values have not been determined in the laboratory due to economical or time consuming reasons, or technical difficulties[1-4]. In this kind of studies one looks for a relationship of the form $P = f(\mathbf{d})$, where P is the property being studied and $\mathbf{d}$ is a set of mathematical or empirical molecular descriptors quantifying the molecular structure and carrying information about it, represented by simple numerical quantities. The simplest descriptors are, for instance, the numbers and types of chosen atoms or bonds in the structure of the molecule. More elaborated descriptors can be derived from various different theories, such as the Chemical Graph Theory, Quantum Mechanics, Information Theory, etc.[5-7]. The function $f(\mathbf{d})$ is commonly unknown and depends on the property P, the set of descriptors $\mathbf{d}$, and the number and type of compounds under study. Typically, this function is chosen so that it generates the best predictions for the property being modeled.

Nowadays there are thousands of descriptors available in the literature[7-9], and one is faced with the problem of selecting the best set of d descriptors out of a much larger set of D ones, according to some criterion such as the smallest total standard deviation S [10-13]. A full search (FS) of such optimal set requires $D!/[(D-d)!d!]$ linear regressions; a number that increases so rapidly with D that soon becomes impracticable. Moreover, if D is smaller than the number of molecules M, then one may look for the global optimal set of descriptors and the necessary comparison of the best sets of $d = 1, 2, \ldots, D$ descriptors requires a total of $2^D - 1$ linear regressions.

There are several methods available that replace the time consuming full search[14-17]; a famous one that has long been used in QSAR-QSPR applications is the Stepwise Regression Method[18]. We also mention the Genetics and Evolutionary Algorithms[19-21], and the Elimination Method (EM) which, in spite of its remarkable simplicity, yields results in close agreement with the FS ones[22,23].

In this paper we propose an alternative method that focuses mostly on the relative errors of the coefficients of the linear regressions because one expects the total standard deviation S to be a function of those errors. This guiding idea emerges from the fact that the best model given by the FS always exhibits comparatively small relative errors in its regression coefficients. We are not aware of other technique available in the literature for the search of optimal variables based solely on the errors of the coefficients. The reason may be that numerical experiments conducted some time ago suggested that the regression coefficients and their associated errors had a random behaviour, because of the instability of the regression equations[24,25].

In Section II we present the method, in Section III we treat an illustrative example detail, and in Section IV we discuss our results, and suggest possible improvements of the method.

II. The Algorithm

The EM provides a close solution to the optimal set of descriptors with just D linear regressions. First, we do a linear regression with all $D < M$ descriptors, calculate the relative error $\Delta c_j / c_j$ of each coefficient c_j in the model, and remove the descriptor with the greatest error. Second we repeat that calculation for $D-1, D-2, \ldots 1$ descriptors and look for the best model according to the criterion chosen (smallest S, for example). The whole process requires just D linear regressions, a number much smaller than the FS one. Two disadvantages of the EM are that it applies only when $D < M$, and that the resulting models commonly have more descriptors than one would desire. The former reason makes the EM unsuitable for the selection of an optimal set of descriptors out of the thousands that are usually available.

Here we propose an alternative method that leads to a set of d descriptors close to the optimal one with a number of linear regressions that is much smaller than the one required by the FS. Instead of removing variables, this new procedure consists of replacing a chosen variable of the set by another one that minimizes S. For this reason we call it Replacement Method (RM) from now on. To this end, choose d descriptors $\{X_1, X_2, \ldots X_d\}$ at random and do a linear regression. Choose one of the descriptors of this set, say X_i, and replace it by each of the D descriptors of the pool (except itself) keeping the best resulting set. Since one can start replacing any of the d

descriptors in the initial model, then a regression equation with d variables has d possible paths to achieve the final result; for example, the choice above will develop into path i. Next, choose the variable with greatest relative error in its coefficient (except the one replaced in the previous step) and replace it with all the D descriptors (except itself) keeping again the best set. Replace all the remaining variables in the same way bypassing those replaced in previous steps. When finishing, start again with the variable having greatest relative error in the coefficient and repeat the whole process. Repeat this process as many times as necessary until the set of descriptors remains unchanged. At the end, we have the best model for the path i. Proceed in exactly the same way for all possible paths $i = 1, 2, \ldots d$, compare the resulting models, and keep the best one. Our numerical experiments show that in this way one obtains a model almost as good as the best one with much less than $D!/[(D-d)!d!]$ linear regressions when this combinatorial number is large.

III. Illustrative Example

Table I shows a data set of 30 melting points of quinoxalines derivatives reported in references[26,27]. In the case of a temperature interval we chose the mean value. The structures of the compounds were preoptimized by means of the Molecular Mechanics Force Field (MM+) included in Hyperchem version 6.03. Since several molecules contained sulfur atoms, final refined molecular structures were obtained using the semiempirical method PM3 (Parametric Method-3). We chose a gradient norm limit of 0.01 kcal/Å for geometry optimization.

We derived several types of molecular descriptors, such as constitutional, topological, geometrical, charge, Randic Profiles, Atom Centered Groups, etc., by means of the software Dragon version 5 evaluation software available in the web[28]. For simplicity, we decided to restrict our search to sets of no more than 14 descriptors of each family.

Table I. Data set of melting points of quinoxalines derivatives used in the present analysis.

No	Compound Name	MP (ºC)
1	3H-Pyrazolo[3,4-b]quinoxalin-3-one, 1,2-dihydro 1-phenyl-	235.5 [235-236]
2	4H-1,2-Thiazino[5,6-b]quinoxaline-3-acetonitrile, 4-oxo-	242
3	4H-1,2-Thiazino[5,6-b]quinoxaline-3-acetamide, 4-oxo-	255.5 [255-256]
4	4H-1,2-Thiazino[5,6-b]quinoxaline-3-acetic acid, 4-oxo-, ethyl ester	195
5	Benzo[g]pteridine-2,4(1H,3H)-dione, 2-hydrazone	285
6	Benzo[g]pteridin-4(1H)-one, 2-(methylthio)-	168
7	Benzo[g]pteridin-4(1H)-one, 2-(ethylthio)-	156
8	5H-Benzo[g]thiazolo[2,3-b]pteridin-5-one, 3-phenyl-	274
9	5H-Benzo[g]thiazolo[2,3-b]pteridin-5-one, 3-(4-chlorophenyl)-	248
10	5H-Benzo[g]thiazolo[2,3-b]pteridine-2-carbonitrile, 3-amino-5-oxo-	234
11	Benzo[g]-1,2,4-triazolo[3,4-b]pteridin-5(1H)-one, 3-phenyl-	284
12	Benzo[g]-1,2,4-triazolo[3,4-b]pteridin-5(1H)-one, 3-(4-nitrophenyl)-	311 [310-312]
13	Benzo[g]-1,2,4-triazolo[3,4-b]pteridin-5(1H)-one, 2,3-dihydro-3-thioxo-	293
14	Benzo[g]-1,2,4-triazolo[3,4-b]pteridin-5(1H)-one, 3-(ethylthio)-	196.5 [196-197]
15	Benzo[g]-1,2,4-triazolo[3,4-b]pteridin-5(1H)-one, 3-methyl-	296 [295-297]
16	3-Ethoxycarbonyl-quinoxalin-2(1H)thione	187
17	1,2H-(Pyrazolo[4,5-b]quinoxaline)-3-one	220.5 [220-221]
18	1,2,3,4 Tetrahydro-4-oxo-Pyrimido[4,5-b]quinoxalin-2-thione	275
19	3-(3'-Mercapto-1',2',4'-oxadiazol-5'-yl)quinoxalin-2(1H)-one	290
20	2(1H)-Quinoxalinone, 3-[5-(methylthio)-1,3,4-oxadiazol-2-yl]-	175
21	2(1H)-Quinoxalinone, 3-[5-[(2-oxo-2-phenylethyl)thio]-1,3,4-oxadiazol-2yl]-	164
22	Oxazolo[4,5-b]quinoxalin-2(3H)-one	335
23	N,N'[bis(quinoxalin-2(1H)-one-3-yl)] urea	340
24	3-Piperidinocarbonylamino-quinoxalin-2(1H)-one	265
25	2-Chloro-3-piperidinocarbonylaminoquinoxaline	140
26	3-Piperidinocarbonylaminoquinoxalin-2(1H)-thione	225

27	2-Hydrazino-3-piperidinocarbonylaminoquinoxaline	285
28	3-Methoxycarbonylamino-quinoxalin-2(1H)-one	218
29	2(1H)-Quinoxalinone, 3-[[5-(methylthio)-1H-1,2,4-triazol-3-yl]amino]-	210
30	2(1H)-Quinoxalinone,3-[[5-[(2-oxo-2-phenylethyl)thio]-1H-1,2,4-triazol-3-yl]amino]-	228

Intervals of temperatures are given within brackets.

Before discussing our results, we outline the application of our method to a sample case of $d = 6$ topological descriptors out of the 14 ones given in Table II labeled as $X_2 - X_{15}$. We arbitrarily choose the initial model to be $\{X_2 - X_7\}$ that gives $S^{(0)} = 54.20^{\circ} C$. The resulting model is

$$\begin{aligned} P \; &= -287.09(\pm 268.70) + 4.41(\pm 113.61) X_2 - 19.08(\pm 93.90) X_3 \\ &+ 382.97(\pm 141.29) X_4 + 1.01(\pm 1072.63) X_5 + 0.75(\pm 77.79) X_6 \\ &- 1.00(\pm 673.31) X_7 \end{aligned} \qquad (1)$$

were the relative errors of the regression coefficients are given between parentheses.

Table II. Labels of the descriptors used in the illustrative example.

descriptor	label	Description
X_2	ISIZ	information index on molecular size
X_3	IAC	total information index of atomic composition
X_4	AAC	mean information index on atomic composition
X_5	ZM1	first Zagreb index M1
X_6	ZM1V	first Zagreb index by valence vertex degrees
X_7	ZM2	second Zagreb index M2
X_8	ZM2V	second Zagreb index by valence vertex degrees
X_9	Qindex	Quadratic index

X_{10}	SNar	Narumi simple topological index (log)
X_{11}	HNar	Narumi harmonic topological index
X_{12}	GNar	Narumi geometric topological index
X_{13}	Xt	Total structure connectivity index
X_{14}	Dz	Pogliani index
X_{15}	Ram	ramification index

As path 1 we start replacing X_2, and from now on we indicate a substitution by (X_{old}, X_{new}). Of all the 14 variables, the substitution that minimizes S is (X_2, X_{15}) yielding $S^{(1)} = 48.97^0C$, and the following equation:

$$\begin{aligned} P \;=\; & 598.86(\pm 35.93) \;+\; 82.31(\pm 40.41)X_{15} - 5.99(\pm 56.76)X_3 \\ & - 203.03(\pm 57.76)X_4 \;+\; 21.77(\pm 56.58)X_5 \;+\; 0.20(\pm 263.80)X_6 \\ & - 20.62(\pm 46.56)X_7 \end{aligned} \qquad (2)$$

We now replace the variable with the greatest relative error X_6 with all the 14 descriptors and find that the substitution (X_6, X_{11}) yields the smallest standard deviation $S^{(2)} = 46.36^0C$ and the linear combination

$$\begin{aligned} P \;=\; & -287.39(\pm 190.42) + 151.71(\pm 32.39)X_{15} - 4.30(\pm 64.55)X_3 \\ & -209.77(\pm 49.43)X_4 \;+\; 34.15(\pm 27.68)X_5 + 449.20(\pm 59.45)X_{11} \\ & -34.98(\pm 48.79)X_7 \end{aligned} \qquad (3)$$

Of all the variables not yet replaced, X_3 is the one with the largest relative error. After its replacement by all the 14 descriptors, we obtain that the substitution with smallest deviation is (X_3, X_6) that yields $S^{(3)} = 46.17^0C$ and the model:

$$\begin{aligned} P \;=\; & -450.34(\pm 111.66) \;+\; 15806(\pm 29.95)X_{15} + 0.600(\pm 61.98)X_6 \\ & -273.67(\pm 37.43)X_4 + 24.33(\pm 42.42)X_5 + 579.89(\pm 39.63)X_{11} \\ & -30.21(\pm 35.32)X_7 \end{aligned} \qquad (4)$$

The replacement of any of the remaining variables $\{X_4, X_5, X_7\}$ as indicated above does not lead to a new model; therefore, we should start the process from the beginning. The variable with the largest relative error in equation (4) is X_6; however, its replacement does not lead to a new model. Exactly the same situation occurs with the next one X_5, and also with X_4 and X_7. We are thus left with $\{X_{11}, X_{15}\}$. The replacement of the one with the greatest relative error X_{11} leads to the substitution (X_{11}, X_9), the standard deviation $S^{(4)} = 45.76^0C$ and the model:

$$\begin{aligned} P \;=\; & 617.83(\pm 31.43) + 70.00(\pm 42.68)X_{15} + 0.76(\pm 46.83)X_6 \\ & -284.41(\pm 35.68)X_4 + 26.46(\pm 40.36)X_5 + 49.60(\pm 38.05)X_9 \\ & -33.33(\pm 34.02)X_7 \end{aligned} \qquad (5)$$

The last substitution results to be (X_{15}, X_{10}) with $S^{(5)} = 45.01^0C$ and the function:

$$\begin{aligned} P \;=\; & 612.85(\pm 30.41) \;-\; 75.19(\pm 39.41)X_{10} + 0.96(\pm 36.20)X_6 \\ & -287.11(\pm 34.26)X_4 + 36.92(\pm 34.58)X_5 + 57.68(\pm 32.93)X_9 \\ & -30.98(\pm 34.05)X_7 \end{aligned} \qquad (6)$$

Restarting the process once again we get the trivial substitutions (X_6, X_6), (X_4, X_4) and (X_5, X_5), but (X_9, X_{12}) leads to an improved model with $S^{(6)} = 44.70^0C$ and:

$$\begin{aligned} P \;=\; & -1456.26(\pm 48.84) - 180.98(\pm 27.32)X_{10} + 1.02(\pm 33.85)X_6 \\ & -283.41(\pm 34.42)X_4 + 49.87(\pm 32.14)X_5 + 1037.35(\pm 32.14)X_{12} \\ & -23.07(\pm 34.64)X_7 \end{aligned} \qquad (7)$$

If we start the process once again, we do not obtain any new model with smaller S; consequently, the optimal model for path 1 is $\{X_4, X_5, X_6, X_7, X_{10}, X_{11}\}$, with $S(1) = 44.70^0C$. Proceeding exactly in the same way for the other possible paths we find that the best result is that for path 1 which is the one that appears in Table III that we will discuss in what follows.

Table III S and number of linear regressions for the Full Search (FS) and the Replacement Method (RM).

d	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Constitutional Descriptors														
FS	54.6	50.0	49.0	46.8	44.6	42.3	42.5	42.4	43.2	43.8	44.7	45.8	47.2	48.8
	14	*91*	*364*	*1001*	*2002*	*3003*	*3432*	*3003*	*2002*	*1001*	*364*	*91*	*14*	*1*
RM	54.6	51.1	50.2	48.1	44.7	42.3	42.6	42.4	43.2	43.8	44.7	45.8	47.2	48.8
	14	*182*	*216*	*396*	*500*	*648*	*952*	*1064*	*1080*	*1000*	*968*	*864*	*676*	*1*
Topological Descriptors														
FS	55.8	52.2	51.6	49.5	47.2	44.5	42.6	42.2	42.3	41.7	42.6	43.7	45.0	46.4
RM	55.8	52.2	51.7	50.3	48.0	44.7	43.6	42.2	42.3	41.7	42.6	43.7	45.0	46.4
	14	*104*	*216*	*352*	*550*	*864*	*952*	*896*	*1134*	*1000*	*968*	*864*	*676*	*1*
WHIM Descriptors														
FS	51.1	49.0	43.4	41.4	40.9	41.3	41.6	41.9	42.3	42.8	43.9	45.1	46.5	48.0
RM	51.1	49.0	43.4	41.4	40.9	41.3	41.6	41.9	42.3	42.8	43.9	45.1	46.5	48.0
	14	*156*	*216*	*352*	*500*	*702*	*840*	*896*	*972*	*1000*	*968*	*864*	*676*	*1*
Galvez Topological Charge Indexes														
FS	55.7	56.0	55.8	55.6	54.1	53.6	53.9	53.5	53.4	54.6	55.8	57.4	59.0	61.0
RM	55.7	56.2	56.1	55.9	54.1	53.6	53.9	53.5	53.4	54.6	55.8	57.4	59.0	61.0
	14	*52*	*216*	*440*	*650*	*756*	*952*	*896*	*972*	*1400*	*968*	*936*	*676*	*1*
Molecular Walk Counts														
FS	56.6	56.6	53.3	53.2	53.3	53.3	51.7	49.3	48.5	48.5	49.4	50.5	51.8	53.5
RM	56.6	57.0	53.3	54.1	53.3	53.3	51.7	49.3	48.5	48.5	49.4	50.5	51.8	53.5
	14	*52*	*216*	*176*	*500*	*648*	*784*	*896*	*972*	*1000*	*968*	*864*	*676*	*1*
GETAWAY														
FS	51.9	50.8	46.8	42.8	41.1	38.3	36.3	34.8	35.4	36.0	36.7	37.6	38.8	40.0
RM	51.9	50.8	46.8	44.8	42.0	38.3	36.3	34.8	35.4	36.0	36.7	37.6	38.8	40.0
	14	*104*	*216*	*352*	*700*	*756*	*896*	*896*	*972*	*1000*	*968*	*864*	*676*	*1*

3D-MoRSE														
FS	52.9	49.6	49.7	49.2	48.5	48.6	48.7	48.6	48.9	49.1	50.2	51.3	52.7	54.4
RM	52.9 *14*	49.6 *130*	49.7 *252*	49.5 *396*	48.5 *600*	48.6 *810*	49.1 *784*	48.6 *896*	48.9 *972*	49.1 *1000*	50.2 *968*	51.3 *864*	52.7 *676*	54.4 *1*
Atom Centered Fragments														
FS	51.3	46.6	43.1	41.6	41.1	39.7	38.1	37.8	37.6	37.6	38.5	39.6	40.7	42.1
RM	51.3 *14*	46.6 *104*	43.1 *288*	42.5 *352*	41.1 *550*	39.7 *756*	38.1 *840*	38.1 *896*	37.6 *1026*	37.6 *1000*	38.5 *968*	39.6 *864*	40.7 *676*	42.1 *1*
BCUT														
FS	57.1	55.7	53.6	52.1	48.5	48.0	45.8	45.3	45.9	46.2	47.1	48.4	49.8	51.4
RM	57.1 *14*	55.7 *104*	53.6 *216*	52.1 *352*	48.5 *500*	48.0 *648*	45.8 *896*	45.3 *896*	45.9 *1404*	46.2 *1000*	47.1 *968*	48.4 *900*	49.8 *676*	51.4 *1*
Functional Groups														
FS	53.2	51.3	48.3	48.0	46.7	45.5	46.2	45.0	46.0	47.1	48.3	49.7	51.2	52.9
RM	53.2 *14*	51.3 *52*	48.3 *108*	48.0 *352*	48.2 *250*	45.5 *756*	46.2 *840*	45.0 *896*	46.0 *972*	47.1 *1000*	48.3 *968*	49.7 *864*	51.2 *676*	52.9 *1*
2D-Autocorrelations														
FS	57.5	55.5	55.3	53.2	53.9	54.1	53.8	54.4	55.3	56.3	57.8	59.4	61.2	63.2
RM	57.5 *14*	55.5 *130*	55.6 *216*	56.6 *484*	53.9 *1000*	54.1 *1296*	53.8 *1008*	54.4 *1624*	55.3 *972*	56.3 *1000*	57.8 *968*	59.4 *864*	61.2 *676*	63.2 *1*
Geometrical														
FS	54.7	51.8	45.6	44.1	41.8	39.5	35.3	31.4	31.9	29.3	28.7	29.4	30.3	31.2
RM	54.7 *14*	51.8 *104*	45.6 *288*	44.1 *352*	41.8 *550*	39.9 *648*	35.3 *952*	31.4 *1288*	31.9 972	29.3 *1000*	28.7 *968*	29.4 *864*	30.3 *676*	31.2 *1*
Randic Molecular Profiles														
FS	57.6	58.5	58.8	58.0	54.2	53.9	54.3	55.2	56.5	57.9	59.4	61.0	62.8	64.9
RM	57.6 *14*	58.5 *156*	59.4 *288*	59.4 *176*	54.5 *700*	53.9 *648*	56.1 *1176*	56.9 *896*	56.5 *1458*	57.9 *1500*	59.4 *1012*	61.0 *900*	62.8 *754*	64.9 *1*

Charge														
FS	53.7 *13*	53.0 *78*	44.0 *286*	42.0 *715*	42.6 *1287*	42.0 *1716*	42.9 *1716*	43.1 *1287*	43.2 *715*	43.9 *286*	44.2 *78*	45.4 *13*	46.8 *1*	-
RM	53.7 *13*	54.0 *88*	51.2 *198*	50.3 *360*	42.6 *585*	42.5 *672*	42.9 *1029*	43.1 *816*	43.2 *1170*	43.9 *800*	44.2 *759*	45.4 *576*	46.8 *1*	-

Numbers in italics represent the number of required linear regressions.

IV. Results and Discussion

Table III shows the total standard deviation S for each model obtained by FS and RM, corresponding to $d = 1, 2, \ldots, 14$ descriptors, as well as the number of required linear regressions. In the case of the FS this number depends only on d and is displayed only in the first row because it is exactly the same for the other ones. On the other hand, the number of regressions for the RM depends also on the type of descriptors in the model. It follows from Table III that in nearly all cases $S(RM)$ is in close agreement with $S(FS)$, especially for large numbers of descriptors. This result suggests that the more variables available for replacement the more chances to hit the exact FS result. We appreciate that the RM requires less linear regressions except for the limit cases of d close to 1 and d close to D which are of much lesser importance from a practical point of view. In the cases of reasonable numbers of descriptors, and in particular, when S is close to its minimum, the RM is considerably faster than the FS. We are presently investigating the possibility of selecting the paths that most probably will lead to the best model in order to improve the RM even further. However, it does not seem to be a simple task and we do not yet have a systematic criterion for that purpose. However, in its present form, the RM seems to be a most promising alternative to the much more time consuming FS in the case of a large number of possible descriptors (say, several thousands).

Table IV shows that the best models for each family of descriptors obtained by means of the EM, the RM, and the FS are in quite close agreement. In those cases the EM is most preferable because it requires only $D = 14$ linear regressions. However, as said before, the EM does not apply when $D > M$ and is therefore unsuitable for the kind of search that many researchers want to do nowadays.

Table IV. Best models for the different families of descriptors labeled by $S(d)$.

FS	RM	EM
Constitutional Descriptors		
42.34 (6)	42.34 (6)	42.41 (8)
Topological Descriptors		
41.75 (10)	41.75 (10)	42.24 (10)
WHIM Descriptors		
40.97 (5)	40.97 (5)	41.61 (6)
Galvez Topological Charge Indexes		
53.48 (9)	53.48 (9)	53.48(9)
Molecular Walk Counts		
48.58 (9)	48.58 (9)	48.58 (9)
GETAWAY		
34.89 (8)	34.89 (8)	34.89 (8)
3D-MoRSE		
48.59 (5)	48.59 (5)	48.93 (9)
Atom Centered Fragments		
37.65 (9)	37.65 (9)	37.65 (9)
BCUT		
45.35 (8)	45.35 (8)	45.81 (7)
Charge		
42.07 (4)	42.50(6)	43.23 (9)
Functional Groups		
45.07 (8)	45.07 (8)	45.07 (8)
2D-Autocorrelations		

53.27 (4)	53.80 (7)	53.87 (7)
Geometrical		
28.72 (11)	28.72 (11)	28.72 (11)
Randic Molecular Profiles		
53.94 (6)	53.94 (6)	54.28 (6)

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

V. References

(1) Sexton, W. A., Chemical Constitution and Biological Activity, D. Van Nostrand, New York, **1950**.

(2) Hansch, C., Fujita, T., ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure, *J. Am. Chem. Soc.*, **1964**, *86*, 1616-1626.

(3) Hansch, C., Quantitative Approach to Biochemical Structure-Activity Relationships, *Acc. Chem. Res.*, **1969**, *2*, 232-239.

(4) King, R. B., Ed., Chemical Applications of Topology and Graph Theory; Studies in Physical and Theoretical Chemistry, *Elsevier*, Amsterdam, **1983**, *28.*

(5) Kier, L. B., Hall, L. H., Molecular Connectivity in Structure-Activity Analysis, *Wiley*, New York, **1986**.

(6) Trinajstic, N., Chemical Graph Theory, 2nd revised ed., CRC Press, Chapter 10, Boca Raton, FL, **1992**.

(7) Katritzky, A. R., Lobanov, V. S., Karelson, M., QSPR-The Correlation and Quantitative Prediction of Chemical and Physical Properties from Structure, *Chem. Soc. Rev.*, **1995**, *24*, 279-287.

(8) Todeschini, R., Handbook of Molecular Descriptors, *Wiley-VCH*, Berlin, **2000**.

(9) Basak, S. C., Harris, D. K., Magnuson, V. R., POLLY (version 2.3), Copyright of the University of Minnesota, **2001**.

(10) Lucic, B., Nikolic, S., Trinajstic, N., Juretic, D., The Structure-Property Models Can Be Improved Using the Orthogonalized Descriptors, *J. Chem. Inf. Comput. Sci.*, **1995**, *35*, 532-538.

(11) Lucic, B., Trinajstic, N., New Developments in QSPR/QSAR Modeling Based on Topological Indices, *SAR QSAR Environ. Res.*, **1997**, *7*, 45-62

(12) Lucic, B., Trinajstic, N., Multivariate Regression Outperforms Several Robust Architectures of Neural Networks in QSAR Modeling, *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 121-132.

(13) Lucic, B., Trinajstic, N., Sild, S., Karelson, M., Katritzky, A. R., A new Efficient Approach for Variable Selection Based on Multiregression: Prediction of Gas Chromatographic Retention Times and Response Factors, *J. Chem. Inf. Comput. Sci.*, **1999**, *39*, 610-621.

(14) Livingstone, D. J., Rahr, E., CORCHOP-An Interactive Routine for the Dimension Reduction of Large QSAR Data Sets, *Quant. Struct-Act. Relat.*, **1989**, *8*, 103-108.

(15) McFarland, J. W., Gans, D. J., On Identifying Likely Determinants of Biological Activity in High Dimensional QSAR Problems, *Quant. Struc.-Act. Relat.*, **1994**, *13*, 11-17.

(16) Héberger, K., Rajkó, R., Variable Selection Using Pair-Correlation Method, *SAR- QSAR Environm. Res.*, **2002**, *13*, 541-554.

(17) Héberger, K., Rajkó R., Generalization of Pair Correlation Method (PCM) for nonparametric variable selection, *J. Chemom.*, **2002**, *16*, 436-443.

(18) Draper, N. R., Smith, H., Applied Regression Analysis, Second Edition ed., *John Wiley & Sons*, New York, **1981**.

(19) Kubinyi, H., Evolutionary variable selection in regression and PLS analysis, *J. Chemom.*, **1996**, *10*, 119-133.

(20) Kubinyi, H., Variable selection in QSAR studies, I. An Evolutionary Algorithm, *Quant. Struct.- Act. Relat.*, **1994**, *13*, 285-294.

(21) Kubinyi H., Variable selection in QSAR studies, II. A highly efficient combination of systematic search and evolution, *Quant. Struct.- Act. Relat.*, **1994**, *13*, 393-401.

(22) Duchowicz, P. R., Fernández, F. M., Castro, E. A., Alternative Statistical and Theoretical Analysis of Fluorophilicity, *J. Fluor. Chem.*, **2004**, *125*, 43-48.

(23) Nesterov, I. V., Toropov, A. A., Duchowicz, P. R. and Castro E. A., An Improved QSPR Modeling of Hydrocarbon Dipole Moments, TheScientifcWorld Journal (In Press).

(24) Randic, M., Resolution of Ambiguities in Structure-Property Studies by Use of Orthogonal Descriptors, *J. Chem. Inf. Comput. Sci.*, **1991**, *31*, 311-320.

(25) Randic, M., Retro-regression- Another important multivariate regression improvement, *J. Chem. Inf. Comput. Sci.*, **2001**, *41* , 602.

(26) Moustafa, O. S., Abbady, M. S., Triazolopyrimidoquinoxalines and Thiazolopyrimidoquinoxalines, *AFINIDAD*, **2001**, *495*, 335-340.

(27) Moustafa, O. S., Synthesis and some reactions of quinoxalinecarboazides, *J. Chin. Chem. Soc.,* **2000**, *47*, 351-357.

(28) http://www.talete.mi.it/dragon.htm