Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Fluorine Chemistry 125 (2004) 43–48
JOURNAL OF FLUORINE CHEMISTRY
www.elsevier.com/locate/jfluchem

Alternative statistical and theoretical analysis of fluorophilicity

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

INIFTA, Departamento de Química, Facultad de Ciencias Exactas, Universidad Nacional de La Plata, Diag. 113 y 64, Suc. 4, Casilla de Correo 16, La Plata 1900, Argentina

Received 23 August 2003; received in revised form 22 September 2003; accepted 26 September 2003

Abstract

We present quantitative structure–property relationships (QSPR) for the partition of 99 organic compounds between organic and fluorinated solvents. The approach consists of straightforward multivariate regression using the simplest topological molecular descriptors. We discuss how to obtain the best model for each set of molecular descriptors and show that present statistical parameters are comparable to those given by more elaborate descriptors.

Keywords: Fluorine; Fluorous partition coefficient; Structure–activity relations; Topological indices; Statistical models

1. Introduction

The graph-theoretical approach to quantitative structure–property relationships (QSPR) is based on a well-defined mathematical representation of the molecular structure according to the basic mathematical formula $P = f(\{D\})$, where P is a property, $\{D\}$is a set of molecular descriptors and f is a properly chosen function.

Some molecular descriptors commonly named ''*topological indices*'' [1,2] are derived from a well-defined mathematical representation of the molecular structure [3,4] and contain relevant information about it. Owing to the complexity of molecular structure, one does not expect that a single set of descriptors will carry all the relevant structural information. Therefore, the search for novel molecular structure descriptors is an active research field within the realm of QSPR theory. However, this search should not be at random but follow some regular procedures based on the desired attributes that a molecular structure descriptor should exhibit [5].

For more than a century most chemists have used constitutional formulae without realising that such representations of ''connectedness'' of atoms are graphs or multigraphs [6]. As a matter of fact, the structural (or constitutional) formula of a chemical compound may be regarded as molecular graph where the vertices represent atoms while the edges represent valence bonds [7]. Evidently, the simplest and most obvious sort of graph-theoretical indices are atoms and chemical bonds. Although they have been considered as suitable molecular descriptors, they have not been widely employed. Several applications by two of us have demonstrated their usefulness to predict physical-chemistry properties and biological activities [8–14]. These parameters may be calculated solely from consideration of the molecular structure and their chemical interpretation is quite direct. They can be computed very readily and have the advantage that they may be applied to quite diverse sets of structures.

Two statistical studies about the fluorophilicity of two different sets of organic molecules have been recently published [15,16]. The first of them reports the estimation of the fluorophilicity of 59 fluorinated organic molecules using a neural network (NN) combination of eight descriptors chosen from a pool of almost 100 possible molecular descriptors. The second article reports linear free energy relation (LFER) models of the partition of 99 organic compounds between organic and fluorous solvents. The authors develop accurate predictive models with a standard deviation of less than three times the estimated experimental error. Both papers resort to standard molecular and topological descriptors and draw conclusions about the fluorophilicity character of the corresponding molecular sets. However, it is well known such kind of statistical studies do not allow one to derive unambiguous physical chemistry interpretations on the basis of the molecular descriptors employed to derive the regression equations. That is to

* Corresponding author. Tel.: +54-221-4214037; fax: +54-221-4259485.
E-mail addresses: castro@quimica.unlp.edu.ar, jubert@arnet.com.ar (E.A. Castro).

0022-1139/$ – see front matter © 2003 Elsevier B.V. All rights reserved.
doi:10.1016/j.jfluchem.2003.09.003

say, there are no universal rules to draw conclusions on the basis of a particular set of molecular or/and topological descriptors because a relatively good predictive power of regression equations does not lead to cause–effect relationships between property and independent variables. For this reason we believe that there is room for the application of another set of molecular descriptors insofar they are readily interpreted and give good results. In this regard, it is particularly tempting to look for a basically simple set of topological descriptors, such as atoms composing each molecular species and chemical bonds connecting them in a classical way.

The purpose of this article is to report results on fluorophilicity for the same 99 organic compounds reported by Huque et al. [16] via a multilinear regression analysis based on first- and second-order fitting polynomials. The paper is organised as follows: in Section 2 we provide some necessary basic definitions on the problem and outline the multivariate linear regression method. In Section 3 we discuss the main results of the paper and compare them with those reported by Huque et al. [16], and in Section 4 we draw some additional conclusions.

2. The method

Molecular tendency to dissolve in fluorous media is most commonly measured by the molecule's partition coefficient P between fluorous and organic phases [17]. This value is currently transformed onto a free energy scale by taking its natural logarithm, and the resulting quantity $\ln P$ is referred to as the "*fluorophilicity*". Throughout this paper we will use the same standard system employed by Huque et al. [16] and originally proposed by Rocaboy et al. [18], that is to say, the partition of molecules between perfluoro (methylcyclohexane), $CF_3C_6F_{11}$, and toluene, given by

$$\ln P = \ln\left[\frac{c(CF_3C_6F_{11})}{c(CH_3C_6H_5)}\right], \qquad T = 298\,\mathrm{K} \tag{1}$$

In what follows we investigate several linear regression models of the form

$$\ln P = \sum_{j=0}^{n} c_j D_j \tag{2}$$

where each D_j, $j = 0, 1, \ldots, n$ is a N-dimensional vector with the values of a given topological descriptor for the N molecules, and D_j is a vector with its N elements equal to unity that accounts for the constant term. We choose the simplest topological descriptors: number of atoms and chemical bonds of each type, and eventually some of their powers. Table 1 shows the labels for the descriptors used in this paper, and for brevity in most discussions we will simply give such labels instead of the actual names of the descriptors (Table 2).

Table 1
Labels for the descriptors used in present calculations

Atoms and bonds	Linear variable	Quadratic variable, only atoms	Quadratic variable, atoms and bonds
C	1	12	24
H	2	13	25
F	3	14	26
O	4	15	27
N	5	16	28
P	6	17	29
I	7	18	30
Br	8	19	31
Si	9	20	32
S	10	21	33
Cl	11	22	34
C–C	12		35
C–H	13		36
C=C	14		37
C=O	15		38
C–O	16		39
C–C aromatic	17		40
P–C	18		41
N–H	19		42
C–N	20		43
C–N aromatic	21		44
O–H	22		45
C=S	23		46

In order to avoid round-off errors in our linear regression calculations we resort to computer algebra systems like Maple and Derive [19,20] that enable one to solve the least-square equations in exact rational mode if necessary. Although this kind of calculation is commonly slow, it is sufficiently fast for our present purposes.

We first tried to obtain the best model for a given set of descriptors according to the criterion of smallest standard deviation S

$$S^2 = \frac{1}{N-n-1}\sum_{j=1}^{N} r_j^2 \tag{3}$$

where r_j are the residuals. Following other authors we tried all the combinations of k descriptors out of n for $k = 1, 2, \ldots, n$ [21]. This procedure is time consuming (especially when using exact rational arithmetic) because the total number of calculations is $2^n - 1$. For that reason we resorted to a different strategy: first do a linear regression with all the descriptors and remove the one with the greatest relative error $\Delta c_j/c_j$. Second, repeat the calculation with the remaining $n-1$ descriptors, and again remove the descriptor with the largest relative error. Proceed exactly in the same way for $n-2$, $n-3, \ldots$ until one descriptor and the constant remain. Then choose the set with the smallest value of S. This procedure requires only n calculations and enables us to single out an optimum model that is very close to or in complete agreement with the one obtained by means of the thorough search mentioned before.

Table 2
Experimental and theoretical values of ln P calculated with the models

Molecule	ln P_{exp}	DS_4 (99)	DS_3 (99)	DS_3 (91)	Huque et al. (91)
1. Decane	−2.86	−3.58	−2.86	−2.79	−3.07
2. Undecane	−3.13	−3.62	−3.11	−3.03	−3.13
3. Dodecane	−3.35	−3.64	−3.36	−3.27	−3.19
4. Tridecane	−3.71	−3.63	−3.61	−3.52	−3.24
5. Tetradecane	−3.94	−3.59	−3.86	−3.77	−3.30
6. Hexadecane	−4.50	−3.45	−4.36	−4.27	−3.41
7. Dec-1-ene	−2.99	−3.75	−3.13	−3.08	−3.29
8. Undec-1-ene	−3.26	−3.83	−3.38	−3.33	−3.34
9. Dodec-1-ene	−3.66	−3.88	−3.63	−3.58	−3.40
10. Tridec-1-ene	−3.94	−3.90	−3.88	−3.82	−3.46
11. Tetradec-1-ene	−4.12	−3.90	−4.13	−4.07	−3.51
12. Hexadec-1-ene	−4.70	−3.83	−4.63	−4.57	−3.62
13. $R_{f8}CH{=}CH_2$	2.67	2.36	2.27	2.14	2.82
14. Cyclohexanone	−3.79	−3.93	−2.88	−2.92	−3.96
15. Cyclohexenone	−4.06	−3.90	−3.15	−3.22	−4.25
16. Cyclohexanol	−4.12	−3.92	−3.81	−3.66	−4.74
17. Trifluoroethanol	−1.77	−1.78	−1.50	−1.59	−1.37
18. $(CF_3)_2CHOH$	−1.02	−0.78	−0.80	−0.92	−0.70
19. $R_{f6}(CH_2)_2OH$	0.10	0.56	0.18	−0.01	0.47
20. $R_{f6}(CH_2)_3OH$	−0.24	0.25	−0.08	−0.26	0.50
21. $R_{f8}(CH_2)_2OH$	1.02	1.53	0.95	0.73	0.72
22. $R_{f8}(CH_2)_3OH$	0.59	1.20	0.70	0.48	0.80
23. $R_{f10}(CH_2)_3OH$	1.42	2.10	1.47	1.21	1.25
24. Pentafluorobenzene	−1.24	−0.95	−0.86	−0.95	−0.58
25. Hexafluorobenzene	−0.94	−0.50	−0.54	−0.64	−0.12
26. Ethylbenzene	−4.41	−3.65	−2.95	−2.98	−4.23
27. Dodecylbenzene	−4.70	−4.78	−5.45	−5.46	−4.79
28. $R_{f8}(CH_2)_3C_6H_5$	−0.02	0.19	0.28	0.24	0.38
29. o-$R_{f6}(CH_2)_3C_6H_4(CH_2)_3R_{f6}$	1.03	1.14	1.31	1.04	1.37
30. o-$R_{f8}(CH_2)_3C_6H_4(CH_2)_3R_{f8}$	2.34	2.37	2.85	2.50	2.32
31. o-$R_{f10}(CH_2)_3C_6H_4(CH_2)_3R_{f10}$	3.62	3.39	4.39	3.96	3.23
32. m-$R_{f8}(CH_2)_3C_6H_4(CH_2)_3R_{f8}$	2.28	2.37	2.85	2.50	2.32
33. p-$R_{f8}(CH_2)_3C_6H_4(CH_2)_3R_{f8}$	2.33	2.37	2.85	2.50	2.32
34. $R_{f8}(CH_2)_3Cl$	0.03	1.01	1.27	1.32	0.37
35. $R_{f8}(CH_2)_3NH_2$	0.85	1.15	1.26	1.20	1.29
36. $R_{f8}(CH_2)_3NH(CH_2)_3R_{f8}$	3.32	3.77	3.79	3.74	3.34
37. $(R_{f6}(CH_2)_2)_3P$	4.41	4.75	4.31	4.36	3.75
38. $(R_{f8}(CH_2)_3)P$	4.41	5.27	5.87	5.81	4.79
39. $(R_{f8}(CH_2)_4)_3P$	4.50	4.43	5.12	5.07	4.53
40. $(R_{f8}(CH_2)_5)_3P$	4.50	3.81	4.37	4.33	4.27
41. $(R_{f6}(CH_2)_2)_2PC_{10}H_{19}$ (menthyl)	1.29	0.68	−0.67	−0.29	1.11
42. $(R_{f8}(CH_2)_2)_2PC_{10}H_{19}$ (menthyl)	2.70	1.83	0.87	1.18	2.10
43. $(p$-$R_{f6}C_6H_4)_3P$	−1.32	−0.25	−0.46	−0.71	−0.57
44. $(p$-$R_{f8}C_6H_4)_3P$	0.76	0.73	1.85	1.49	0.78
45. $Ph(CH_2)_2SiH_3$	−3.29	−3.09	−3.29	−3.29	−4.53
46. $Ph(CH_2)_2SiOC_8H_{15}$	−5.11	−5.16	−5.22	−5.21	−5.56
47. $Ph(CH_2)_2SiOC_6H_{11}$ (cyclohexyl)	−4.82	−4.9	−4.72	−4.72	−5.56
48. $R_{f6}I$	1.31	1.08	1.29	1.33	0.34
49. $R_{f8}I$	2.04	2.08	2.06	2.06	0.93
50. $R_{f10}I$	2.84	3.03	2.83	2.80	1.48
51. $R_{f8}CH{=}CH_2$	2.67	2.36	2.27	2.14	2.82
52. $R_{f8}(CH_2)_3SH$	0.24	1.12	0.25	0.25	1.23
53. $R_{f8}N(CH_2CH_2)_2O$	0.86	0.18	1.15	0.99	1.48
54. $R_{f6}S(CH_2)_2CO_2Et$	−0.67	−0.41	0.10	−0.35	−0.05
55. $R_{f8}S(CH_2)_2CO_2Et$	0.04	0.50	0.87	0.39	0.49
56. CF_3SPh	−2.45	−2.76	−2.30	−2.32	−2.01
57. m-$CF_3SC_6H_4CF_3$	−1.58	−2.03	−2.04	−1.59	−0.85
58. $R_{f8}SPh$	0.59	0.39	−0.05	0.30	−0.15
59. $R_{f7}CH_2NHMe$	1.07	0.99	1.00	1.13	1.49
60. $R_{f7}CH_2NMe_2$	1.53	0.69	0.63	0.92	1.63
61. $R_{f7}CH_2N(CH_2CH_2)_2O$	0.14	−0.57	0.52	0.38	0.60
62. $R_{f7}CH_2NHCH(Me)Ph(+)$	−0.87	−0.96	−1.34	−1.31	−0.65

Table 2 (*Continued*)

Molecule	ln P_{exp}	DS_4 (99)	DS_3 (99)	DS_3 (91)	Huque et al. (91)
63. $R_{f7}C(O))Ph$	0.48	0.44	0.40	–	–
64. $R_{f7}C(O)OCH_2Ph$	2.14	0.08	0.15	–	–
65. *p*-$R_{f7}C(O)OCH_2CH_4OCF_3$	3.15	2.96	1.59	–	–
66. $R_{f7}C(O)Sme$	1.16	0.36	0.50	0.65	0.57
67. $R_{f7}C(O)NHMe$	0.15	0.36	0.34	0.16	−0.23
68. $R_{f7}C(O)NMe_2$	0.34	0.03	−0.03	−0.05	0.66
69. $R_{f7}C(O)N(CH_2CH_2)_2O$	−0.62	0.13	−0.14	−0.59	−0.38
70. $R_{f7}C(S)Me$	1.08	0.36	1.48	0.89	0.19
71. $R_{f7}C(S)NMe_2$	−0.66	−0.06	−0.79	−0.37	−0.20
72. $R_{f7}C(S)N(CH_2CH_2)_2O$	−1.56	−1.35	−0.90	−0.91	−1.18
73. $R_{f7}C(S)NHCH(Me)Ph(+)$	−1.84	−1.77	−2.76	−2.60	−3.18
74. C_6H_6	−2.77	−3.12	−2.45	−2.49	−4.12
75. CF_3Ph	−1.96	−2.17	−1.75	−1.81	−1.82
76. $R_{f6}Ph$	0.54	0.30	0.18	0.02	0.24
77. $R_{f8}Ph$	1.24	1.20	0.95	0.75	0.78
78. $Rf_{10}Ph$	1.77	2.05	1.72	1.48	1.28
79. *o*-$R_{f8}C_6H_4CF_3$	1.50	2.03	1.65	1.42	1.37
80. *m*-$R_{f8}C_6H_4CF_3$	2.37	2.03	1.65	1.42	1.37
81. *p*−$R_{f8}C_6H_4CF_3$	2.13	2.03	1.65	1.42	1.37
82. *p*-$R_{f8}C_6H_4R_{f8}$	4.98	4.67	4.35	–	–
83. $[p\text{-}CF_3C_6H_4(CF_2)_4]_2$	−0.56	0.36	−0.05	−0.40	−0.18
84. *o*-$R_{f6}(CH_2)_2C_6H_4Cl$	−0.64	−1.29	−1.35	−1.35	−0.63
85. *p*-$R_{f6}(CH_2)_2C_6H_4Cl$	−1.02	−1.29	−1.35	−1.36	−0.63
86. *p*-$R_{f8}(CH_2)_2C_6H_4Cl$	−0.37	−0.43	−0.58	−0.62	−0.04
87. *o*-$R_{f6}(CH_2)_2C_6H_4Br$	−1.05	−1.33	−1.33	−1.33	−1.22
88. *m*-$R_{f6}(CH_2)_2C_6H_4Br$	−1.44	−1.33	−1.33	−1.33	−1.22
89. *p*-$R_{f6}(CH_2)_2C_6H_4Br$	−1.49	−1.33	−1.33	−1.33	−1.22
90. *o*-$R_{f8}C_6H_4CO_2Me$	−0.39	0.50	0.53	−0.60	−0.18
91. *m*-$R_{f8}C_6H_4CO_2Me$	0.12	0.50	0.53	−0.60	−0.18
92. *p*-$R_{f8}C_6H_4CO_2Me$	−0.01	0.50	0.53	−0.60	−0.18
93. 1,3,5-$R_{f8}C_6H_3(CF_3)_2$	4.05	2.86	2.36	–	–
94. 1,3,5-$(R_{f8})_2C_6H_3CO_2Me$	4.41	3.82	3.93	–	–
95. 1,3,5−$(R_{f8})_2C_6H_3CH_2OH$	3.62	3.51	2.54	–	–
96. 1,3,5-$(R_{f8})_2C_6H_3CHO$	4.25	3.60	3.44	–	–
97. 2-$R_{f8}C_5H_4N$ (pyridine)	0.54	0.78	0.74	0.74	0.64
98. 3-$_{f8}C_5H_4N$ (pyridine)	0.88	0.78	0.74	0.74	0.64
99. 4-$R_{f8}C_5H_4N$ (pyridine)	0.80	0.78	0.74	0.74	0.64
Average absolute deviation	–	0.40	0.46	0.34	0.44

The number of molecules appears between parenthesis.

3. Results and discussion

In order to determine to which extent the conclusions depend on the chosen set of molecules we applied the procedure outlined in the preceding section to the whole set of 99 molecules and to subsets of 50 and 49 molecules chosen randomly.

So as to have a better understanding of the modelling process we first consider the effect of each set of descriptors separately. For brevity we name the descriptor sets DS_j, where $j = 1, 2, \ldots$. When we choose only atom descriptors (DS_1) we conclude that the optimum one-descriptor models are always given by *F*, the two-descriptor models are given by *F* and *C* and the three-descriptor models are given by *F*, *C*, and *P* or by *F*, *C*, and *S* (depending on the set of molecules chosen). At present it is not clear if such results have any physical meaning or if they are merely mathematical artefacts; more research on this subject is yet needed to draw satisfactory conclusions. The correlation matrix exhibits correlation between the following pairs of descriptors: {*C*, *F*}, {*P*, *F*}, {*C*, *P*}. However, the ratio of the largest to the smallest eigenvalue of the correlation matrix (with all the molecules) $\lambda_{max}/\lambda_{min} = 80.1$ suggests that correlation is not a serious problem in this case [22].

The best model for the whole set of molecules is given by the descriptors {1, 2, 3, 5, 6, 8, 10, 11} yielding $S = 0.832$ and a correlation coefficient $R = 0.951$. The ''leave-one-out'' validation test for this model yields $R_v = 0.940$ and $S_v = 0.882$. The optimum model depends on the set of molecules chosen. An additional difficulty faced in our tests is that the descriptor vectors become linearly dependent for some subsets of molecules. Later on we discuss this problem in more detail.

If we consider only bond descriptors (DS_2) the best one-descriptor models are given either by C–C or C–H bonds. The best two-descriptor models are given by C–C and C–H

bonds, and the best three-descriptor models are given by C–C, C–H and C–C(aromatic) bonds. In this case the eigenvalue ratio is even smaller $\lambda_{max}/\lambda_{min} = 10.40$ suggesting a less serious correlation problem.

The *N*-dimensional vector descriptors for all atoms and chemical bonds (DS_3) are linearly dependent. In order to obtain a linearly independent set of regressors we arbitrarily start with the first descriptor and add the remaining ones one by one removing those that result to be linearly dependent. We are thus left with the set of atom plus bond descriptors with labels {1–17, 19, 20, 22, 23}.

When trying the analysis discussed above on subsets of molecules we conclude that the optimal models with one descriptor are always given by *F*. Atoms F and C contribute to most two-descriptor models but also atom H and bond C–C appear in some cases. The analysis of three-descriptor models is more complicated but we also find that atoms F and C contribute to the optimal models in most of the tests.

The optimum model is given by the descriptors {1–11, 13, 15, 16, 19, 20} with statistical parameters $R = 0.967$ and $S = 0.715$. The leave-one-out validation yields $R_v = 0.952$ and $S_v = 0.790$.

When we consider only atom descriptors and their squares (DS_4) we are again faced to linear dependency. Proceeding as before we arrive at a linearly independent subset of descriptors with labels {1–15}. The best model according to regression is given by the descriptors {1–13, 15} with $R = 0.979$ and $S = 0.571$. The leave-one-out validation test yields $R_v = 0.970$ and $S_v = 0.623$.

Finally we construct a set of regressors with all atoms and bonds plus their squares (DS_5). Only those with labels {1–17, 19, 20, 22–27, 35, 36, 39, 40, 42} are linearly independent. The best model in this case is {1–5, 7–11, 13–15, 17, 19, 24–26, 35, 36, 39, 40} with $R = 0.986$ and $S = 0.493$. The leave-one-out validation tests yields $R_v = 0.979$ and $S_v = 0.534$.

The analysis of the dominant descriptors in the cases DS_4 and DS_5 is not of relevance because of the lack of physical interpretation of the squares of the number of atoms and bonds.

In order to investigate the presence of outliers we just searched for those molecules satisfying the condition $|r_i| > 3AAV$ (C_1) or $|r_i| > 2.5AAV$ (C_2) using the best model for each set of descriptors. Here AAV stands for average absolute deviation

$$AAV = \frac{1}{N}\sum_{i=1}^{N}|r_i| \tag{4}$$

We thus obtained:

$$DS_1: \begin{cases} C_1: \{65, 64\} \\ C_2: \{65, 64, 93, 41, 42\} \end{cases}$$

$$DS_2: \begin{cases} C_1: \text{none} \\ C_2: \{64, 93, 65, 15, 14, 34, 52, 60, 16\} \end{cases}$$

$$DS_3: \begin{cases} C_1: \text{none} \\ C_2: \{64, 65, 41, 42, 93\} \end{cases}$$

$$DS_4: \begin{cases} C_1: \{64\} \\ C_2: \{64\} \end{cases}$$

$$DS_5: \begin{cases} C_1: \{64\} \\ C_2: \{64, 93\} \end{cases}$$

where each list of molecule labels indicates decreasing values of $|r_j|$. We appreciate that molecule 64 appears in 8 sets, molecule 65 appears in four sets and molecule 93 appears in four sets. These molecules were already identified by Huque et al. [16] as outliers. It is not surprising that the agreement is not complete because our descriptors are completely different from theirs.

In order to compare our results with those of Huque et al. [16] more closely we choose the set of 91 molecules obtained after removing the outliers found by those authors. In what follows we show our results in a compact form indicating the best model and its statistical parameters for each descriptor set:

$$DS_1: \begin{cases} \{1, 2, 3, 4, 5, 6, 8, 10, 11\} \\ R = 0.967, \quad S = 0.654 \\ R_v = 0.956, \quad S_v = 0.713 \end{cases}$$

$$DS_2: \begin{cases} \{12, 13, 15, 16, 17, 18, 22, 23\} \\ R = 0.958, \quad S = 0.732 \\ R_v = 0.946, \quad S_v = 0.787 \end{cases}$$

$$DS_3: \begin{cases} \{2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, \\ \quad 19, 20, 23\} \\ R = 0.979, \quad S = 0.561 \\ R_v = 0.956, \quad S_v = 0.720 \end{cases}$$

$$DS_4: \begin{cases} \{1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15\} \\ R = 0.984, \quad S = 0.475 \\ R_v = 0.975, \quad S_v = 0.542 \end{cases}$$

$$DS_5: \begin{cases} \{1, 2, 3, 4, 5, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 19, \\ \quad 20, 24, 25, 26, 35, 36, 40\} \\ R = 0.993, \quad S = 0.344 \\ R_v = 0.985, \quad S_v = 0.423 \end{cases}$$

On the other hand, the five descriptor model proposed by Huque et al. [16] yields $S = 0.321$ and $R_v = 0.970$ [16].

Finally, we have randomly divided the sets of 99 and 91 molecules into subsets of 50 and 49 in the former case and 45 and 46 in the latter one, choosing one subset to be a test set and the remaining one to be a training set, and then reverting their roles. The values of R_v obtained by means of this test oscillate roughly between 0.9 and 0.98 and those of S_v between 0.7 and 1.3. A few cases give worse results probably because of an unfortunate effect of linear dependence. Results are better for the set of 91 molecules than for the set of 99 ones suggesting that the molecules removed by Huque et al. [16] can already be considered to be outliers.

Although the results of this test are not impressive they strongly suggest that the simple descriptors proposed in this paper exhibit a reasonable predictive power.

4. Conclusions

It is clear from the results above that only our best model for the descriptor set DS_5 yields statistical parameters comparable to the LFER model of Huque et al. [16]. If we adopt the criterion that the best model should give the best statistical parameters with the smallest number of descriptors there is no doubt that the five descriptor LFER model is considerably better than our best 23-descriptor model. However, the construction of our descriptors is remarkably simpler because it reduces to just counting atoms and bonds. Thus, present results support previous studies about the use of naive molecular descriptors to predict physicochemical properties and biological activities [8–14].

It is also our purpose to investigate which descriptors are most important to describe a property or activity. Present discussion suggests that some atoms and bonds may be more relevant than others to predict fluorophilicity. However, one should be cautious about such statements because it is well known that the fact that a set of descriptors exhibits a good correlation for a given property or activity does not mean a direct causal connection between the former and the latter.

Since the elements of the vector descriptors are positive integers one may think that one obtains a reasonable correlation by chance. In order to investigate this point further we built sets of linearly independent vector regressors generating random integers between zero and the maximum number of atoms that appeared in the actual descriptors. None of the regression models constructed this way gave statistical parameters as good as those obtained by the actual descriptors. This test suggests that the least-squares fitting produced by such simple topological indices as number of atoms and bonds is not fortuitous.

Acknowledgements

We thank Professor Ricardo Maronna for useful discussions about the application of statistical methods. Authors thank the useful comments made by the reviewer that have been helpful to improve the final version of this paper.

References

[1] Z. Mihalic, N. Trinajstiæ, J. Chem. Educ. 69 (1992) 701.
[2] A. Sablijc, N. Trinajstiæ, Acta Pharm. Jugosl. 31 (1981) 189.
[3] P.J. Hansen, P.C. Jurs, J. Chem. Educ. 65 (1988) 574.
[4] P.G. Seybold, M. May, V.A. Bagal, J. Chem. Educ. 64 (1987) 189.
[5] E. Estrada, O. Ivanciuc, I. Gutman, A. Gutiérrez, I. Rodríguez, New J. Chem. 22 (1998) 819.
[6] A.T. Balaban (Ed.), Chemical Applications of Graph Theory, Academic Press, New York, 1976.
[7] D. Bonchev, Information Theoretic Indices for Characterization of Chemical Structures, UMI, Bell & Howell Company, Ann Harbor, 1998.
[8] P.R. Duchowicz, E.A. Castro, J. Korean Chem. Soc. 43 (1999) 621.
[9] P.R. Duchowicz, E.A. Castro, J. Korean Chem. Soc. 44 (2000) 501.
[10] P.R. Duchowicz, E.A. Castro, Acta Chem. Slov. 47 (2000) 281.
[11] P.R. Duchowicz, E.A. Castro, Arkivoc 2 (2001) 227.
[12] P.R. Duchowicz, E.A. Castro, J. Indian Chem. Soc. 78 (2001) 192.
[13] P.R. Duchowicz, E.A. Castro, Russ. J. Gen. Chem. 72 (2002) 1867.
[14] P.R. Duchowicz, C.C. Chen, E.A. Castro, J. Korean Chem. Soc. 47 (2003) 1.
[15] L.E. Kiss, I. Kövesdi, J. Rábai, J. Fluorine Chem. 108 (2001) 95.
[16] F.T.T. Huque, K. Jones, R.A. Saunders, J.A. Platts, J. Fluorine Chem. 115 (2002) 119.
[17] L.E. Kiss, J. Rábai, L. Varga, I. Kövesdi, Synletters (1998) 1243.
[18] C. Rocaboy, D. Rutherford, B.L. Bennet, J.A. Gladysz, J. Phys. Org. Chem. 13 (2000) 596.
[19] http://education.ti.com/us/product/software/derive/features/features.html, 1998.
[20] http://www.maplesoft.com/, 2000.
[21] B. Luèiæ, S. Nikoliæ, N. Trinajstiæ, D. Juretiæ, J. Chem. Inf. Comput. Sci. 35 (1995) 532.
[22] D.C. Montgomery, E.A. Peck, Introduction to Linear Regression Analysis, Wiley, New York, 1992.