Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 3711–3719
Bioorganic & Medicinal Chemistry

Application of descriptors based on Lipinski's rules in the QSPR study of aqueous solubilities

Pablo R. Duchowicz,[a,*] Alan Talevi,[a,b] Carolina Bellera,[b] Luis E. Bruno-Blanch[b] and Eduardo A. Castro[a]

[a]*Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas (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*
[b]*Cátedra de Química Medicinal, Departamento de Ciencias Biológicas, Facultad de Ciencias Exactas, 47 y 115, (1900) La Plata, Argentina*

Received 13 January 2007; revised 11 March 2007; accepted 14 March 2007
Available online 18 March 2007

Abstract—We complement new physically interpretable descriptors inspired by the Lipinski's rules of drug bioavailability with others obtained from the Dragon 3.0 software, in order to find the best QSPR relationship for aqueous solubilities of 100 structurally heterogeneous organic, drug-like compounds. The simultaneous linear regression analyses of 1367 variables lead to a six-parameter model containing two of the new proposed descriptors and which also possess good predictive ability given by $R = 0.8798$ and cross-validated $R_{1-10\%-o} = 0.8199$. We further validate the model found with an external test set composed of 48 compounds.
© 2007 Elsevier Ltd. All rights reserved.

1. Introduction

At present, it is accepted that, besides being pharmacologically active, an ideal drug should gather some features regarding its bioavailability and its toxicological profile. In silico ADME/Tox filters are nowadays widely used to determine whether it is or not probable for a drug candidate to reach its site of action or to elicit toxic effects at its therapeutic dose. Moreover, modern approaches developed in the pharmaceutical industry for a rational molecular design have moved the ADME/Tox in silico evaluations to the early stages of drug development, where an optimal activity of the compound is searched.[1]

It is known that the degree of absorption of a substance depends simultaneously on dose, solubility and permeability. The exploration of large databases containing orally bioavailable drugs led to the formulation of the widely used Lipinski's 'rule of five' in compounds absorbed through passive diffusion through the gastrointestinal barrier.[2] These simple rules state that oral bio-availability is likely to occur if at least three of the following rules are obeyed: molecular weight below 500; no more than five hydrogen bond donors and less than 10 hydrogen bond acceptors; and calculated octanol–water partition coefficient ($c\log P$) below 5.

The empirical conditions to satisfy Lipinski's rule and manifest a good oral bioavailability involve a balance between the aqueous solubility of a compound and its ability to diffuse passively through the different biological barriers. Aqueous solubility governs both the rate of dissolution of the compound and the maximum concentration reached in the gastrointestinal fluid. However, excessively polar compounds would result problematic at the stage of passing through the various biological barriers.

Furthermore, many other reasons make aqueous solubility an important parameter in Medicinal Chemistry: soluble compounds are associated to shorter metabolization and elimination times, thus leading to lower toxicity and side effects, and most preclinic tests involve solubilization of the drug being tested in hydrophilic solvents.[3,4] Therefore, a high number of theoretical models have been proposed in the past to predict aqueous solubilities, ranging from the early studies of Amidon et al. in 1975[4] to several approaches including

Keywords: QSPR theory; Molecular descriptors; Replacement method; Aqueous solubility.
* Corresponding author. Fax: +54 221 425 4642; e-mail addresses: prduchowicz@yahoo.com.ar; duchow@inifta.unlp.edu.ar

0968-0896/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2007.03.044

thermodynamic calculations, group contribution approximations and quantitative structure–property relationships (QSPR).[5–9]

The main purpose of present research is to apply in a QSPR study for aqueous solubilities of 100 heterogeneous organic chemicals a new set of physically interpretable descriptors introduced in a previous study,[10] characterized for contemplating in a single number several of the parameters involved in the Lipinski's rules. In order to achieve good predictive QSPR models, we complement these new molecular descriptors with others obtained from the software Dragon,[11] thus leading to a total pool containing $D = 1367$ variables including definitions of all types. The best models are searched by means of the replacement method (RM) variable subset selection technique.[12–15]

2. Results and discussion

We have verified, first, the correlation between the descriptors based on Lipinski's rule and the aqueous solubility of the 148 structures of the dataset. The scatter plot for the descriptor that showed best correlation ($D/B^{1/3}$, $R = 0.7236$) is shown in Figure 1. Note that this is a very good correlation coefficient for a single variable and 148 diverse organic structures, The R value increases to 0.7673 if we do not take into consideration the two most obvious outliers: 85 (dimorpholamine) and 94 (Etofenprox).

We proceeded then to search for a QSPR solubility model that minimizes the S parameter subjected to the condition of combining at least one of the proposed molecular descriptors reflecting the Lipinski's rules together with those calculated with the Dragon software. The application of RM to the available pool with $D = 1367$ descriptors leads to an optimal relationship over 100 compounds that, in terms of the best predictive power of the equation (measured via the calibration and the *1–n%-o* parameters) and the least number of variables involved, contain six molecular descriptors of different type:

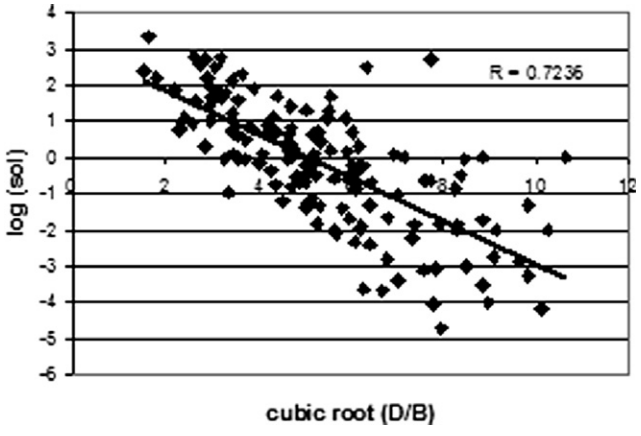

Figure 1. The best correlation between experimental aqueous solubilities given by descriptor $D/B^{1/3}$ for 148 diverse compounds.

$$\begin{aligned}\log(\mathrm{Sol}) &= 2.786(\pm0.3) + 0.0479(\pm0.02)\ RDF040e \\ &\quad + 0.285(\pm0.07)\ C\text{-}006 \\ &\quad - 5.639(\pm0.7)\ H3p + 0.00389(\pm0.001)\ \mathrm{D/A} \\ &\quad - 0.231(\pm0.04)\ \mathrm{D/B^{1/2}} \\ &\quad + 0.00988(\pm0.002)\ QXXe \\ N &= 100\ R = 0.8798\ S = 0.858 \\ F &= 53.091\ p < 10^{-4} \\ R_{\mathrm{loo}} &= 0.8527\ S_{\mathrm{loo}} = 0.911\ R_{1\text{–}10\%\text{–}o} \\ &= 0.8199\ S_{1\text{–}10\%\text{–}o} = 1.006 \end{aligned} \quad (1)$$

where the absolute errors of the regression coefficients are given in parentheses and R is the correlation coefficient, F is the Fisher ratio and p is the significance of the model. Table 1 shows the predicted aqueous solubilities for the training set and the corresponding residuals between parentheses, suggesting that good estimations can be achieved in many cases considering the heterogeneous nature of the training set of molecules. Only five compounds (representing 5% of the training set) are predicted with a residual exceeding the value 2.*S*: 15 (acibenzolar-*S*-methyl), 74 (diazepam), 64 (cyclizine), 40 (azintamide) and 27 (amicarbalide). This may be a statistical consequence of the selected model for predicting the variation of the log(*Sol*) values. The predicted and experimental aqueous solubilities plotted in Figure 2 suggest that the organic compounds tend to follow a line. Figure 3 displays the residuals in terms of the experimental data and demonstrates that the best molecular descriptors given by Eq. 1 lead to estimations that follow a normal distribution trend.

Eq. 1 involves different molecular descriptors that can be classified as follows: two of the proposed absorption-based descriptors: D/A and $D/B^{1/2}$; a radial distribution function (RDF): *RDF040e*, RDF-4.0/weighted by atomic Sanderson electronegativities; a GETAWAY descriptor: *H3p*, H autocorrelation of lag 3/weighted by atomic polarizabilities; an Atom-Centred Fragment: *C*-006, the number of CH_2RX functional groups (X: heteroatom (O, N, S, P, Se or halogens), R: any group linked through carbon); and a geometrical descriptor: *QXXe*, Q_{xx} COMMA2 value/weighted by atomic Sanderson electronegativities. The correlation matrix for the descriptors of the model (indicated in Table 2) reveals that there exists some degree of intercorrelation between D/A and $D/B^{1/2}$, although the model proposed exhibits good 1–10%-*o* cross-validation parameters measured on 90,000 randomly generated cases of compounds exclusion. Randic has indicated that sometimes two highly intercorrelated descriptors may be included in a model if they differ in some parts that are significant to the structure–property relationship.[16,17] In our case, D/A and $D/B^{1/2}$ are included in the model with coefficients of opposite signs, while both descriptors take positive values and the normalized coefficients (see below) indicate that $D/B^{1/2}$ has relatively more importance than D/A in the QSPR equation. Therefore, D/A might be indicating that the contribution of the number of donors in the D/B definition is more important to the aqueous

Table 1. Experimental and predicted values of aqueous solubility [mg ml^{-1}] for training and test sets of organic compounds

No.	Compound name	log(*Sol*)	
		Exp.	Pred. (dif.)
Training set			
1	2,4,5-Trichlorophenol	0.079	−0.601 (0.680)
2	2,4-DB	−1.337	−0.864 (−0.473)
3	2,6-Dibromoquinone-4-chlorimide	−1.231	−1.735 (0.505)
4	2-Cyclohexyl-4,6-dinitrophenol	−1.824	−0.805 (−1.018)
5	2-Ethyl-1-hexanol	−0.056	0.422 (−0.477)
6	3,4-Dinitrobenzoic acid	0.826	0.424 (0.402)
7	4-Amino-2-sulfobenzoic acid	0.477	1.176 (−0.699)
8	Acequinocyl	−4.174	−3.851 (−0.323)
9	Acetamide	3.352	2.358 (0.994)
10	Acetamiprid	0.623	−0.478 (1.101)
11	Acetanilide	0.806	0.487 (0.319)
12	Acetazolamide	−0.009	0.150 (−0.159)
13	Acetochlor	−0.652	−0.771 (0.120)
14	Acetylacetone	2.221	1.205 (1.017)
15	Acibenzolar-*S*-methyl	−2.114	−0.164 (−1.950)
16	Aconitic acid	2.699	1.746 (0.953)
17	Acrylamide	2.806	1.837 (0.969)
18	Acrylonitrile	1.872	2.107 (−0.235)
19	Adenine	0.013	1.355 (−1.342)
20	Adipic acid	1.415	0.303 (1.112)
21	Alanine	2.215	2.182 (0.033)
22	Aldicarb	0.780	0.332 (0.449)
23	Allidochlor	1.294	0.871 (0.423)
24	Allobarbital	0.258	0.472 (−0.214)
25	Alochlor	−0.620	−0.843 (0.223)
26	α-Acetylbutyrolactone	2.301	1.547 (0.754)
27	Amicarbalide	0.700	−1.110 (1.810)
28	Aminopromazine	−3.240	−2.208 (−1.032)
29	Amitraz	−3.000	−3.132 (0.132)
30	Amobarbital	−0.220	0.231 (−0.451)
31	Ancymidol	−0.187	−0.684 (0.497)
32	Aniline	1.556	1.220 (0.336)
33	ANTU	−0.222	0.242 (−0.464)
34	Arabinose	2.699	2.207 (0.492)
35	Ascorbic acid	2.522	1.542 (0.980)
36	Aspartic acid	0.732	1.407 (−0.676)
37	Aspirin	0.663	0.564 (0.099)
38	Asulam	0.699	−0.229 (0.928)
39	Azidamfenicol	1.301	0.286 (1.015)
40	Azintamide	0.699	−1.257 (1.956)
41	Azoxystrobin	−2.000	−2.208 (0.208)
42	Badische acid	−0.222	−0.326 (0.105)
43	Barban	−1.959	−0.648 (−1.310)
44	Barbital	0.873	1.361 (−0.488)
45	Bendiocarb	−0.585	−0.450 (−0.135)
46	Benzidine	−0.495	−0.646 (0.151)
47	Bifenox	−3.398	−3.000 (−0.397)
48	Bifentrhin	−4.000	−3.172 (−0.828)
49	Biotin	−0.658	−0.612 (−0.046)
50	Capric acid	−1.209	−1.307 (0.098)
51	Caproic acid	1.013	0.596 (0.417)
52	Carbofuran	−0.495	−1.020 (0.526)
53	Carbosulfan	−3.523	−3.171 (−0.352)
54	Carboxin	−0.701	−0.038 (−0.663)
55	Carfentrazone-ethyl	−1.658	−1.907 (0.250)
56	Carisoprodol	−0.523	0.271 (−0.794)
57	Carmustine	0.602	1.306 (−0.704)
58	Carnosine	1.914	0.378 (1.537)
59	Cleve's acid	0.000	−0.412 (0.412)
60	Crotonic acid	1.881	1.841 (0.041)
61	Cumic acid	−0.821	0.206 (−1.027)
62	Cyanazine	−0.767	0.533 (−1.300)
63	Cyanuric acid	0.301	1.729 (−1.428)

(*continued on next page*)

Table 1 (*continued*)

No.	Compound name	log(*Sol*)	
		Exp.	Pred. (dif.)
64	Cyclizine	0.000	−1.963 (1.963)
65	Cyclobarbital	0.204	0.128 (0.076)
66	Cycloleucine	1.699	1.323 (0.376)
67	Cymoxanil	−0.051	0.908 (−0.958)
68	Cyproconazole	−0.854	−0.708 (−0.146)
69	Cyprodinil	−1.886	−0.952 (−0.934)
70	Cystine	−0.951	−0.424 (−0.526)
71	Dehydroacetic acid	−0.161	0.919 (−1.081)
72	Dexamethasone	−0.051	−0.646 (0.596)
73	Diallate	−1.854	−2.210 (0.356)
74	Diazepam	−1.301	−3.415 (2.114)
75	Diazinon	−1.398	−1.033 (−0.365)
76	Dicamba	−0.080	0.073 (−0.153)
77	Dichlobenil	−1.674	−1.315 (−0.359)
78	Dichlofenthion	−3.611	−2.338 (−1.273)
79	Diclofop-methyl	−3.097	−3.246 (0.149)
80	Difenoconazole	−1.824	−2.304 (0.481)
81	Digallic acid	−0.301	−1.632 (1.331)
82	Dimethenamid	0.079	−0.413 (0.492)
83	Dimethirimol	0.079	−0.579 (0.658)
84	Dimethomorph	−1.728	−1.685 (−0.043)
85	Dimorpholamine	2.699	3.196 (−0.497)
86	Diniconazole	−2.398	−1.551 (−0.847)
87	EPTC	−0.426	0.791 (−1.217)
88	Equilin	−2.851	−1.562 (−1.288)
89	Ethinamate	0.398	0.556 (−0.158)
90	Ethirimol	−0.699	0.001 (−0.700)
91	Ethofumesate	−1.301	−0.887 (−0.414)
92	Ethohexadiol	1.623	0.965 (0.658)
93	Ethoprop	−0.125	0.208 (−0.333)
94	Etofenprox	0.000	−1.211 (1.211)
95	Fenbuconazole	−4.699	−3.417 (−1.282)
96	Fenbufen	−3.656	−2.350 (−1.305)
97	Fenoxaprop-ethyl	−4.046	−4.011 (−0.035)
98	Fenpiclonil	−2.319	−2.293 (−0.026)
99	Fludrocortisone	−0.854	−0.693 (−0.161)
100	Flufenacet	−0.252	−0.839 (0.587)
Test set			
101	Flufenamic acid	−2.041	−0.953 (−1.088)
102	Flumioxazin	−2.747	−3.255 (0.508)
103	Fluspirilene	−2.000	−1.309 (−0.691)
104	Fluthiacet-methyl	−3.071	−3.298 (0.228)
105	Folic acid	−2.796	−1.575 (−1.221)
106	Fumaric acid	0.845	1.760 (−0.915)
107	Furametpyr	−0.648	−1.483 (0.835)
108	Furazolidone	−1.398	0.047 (−1.445)
109	Ganciclovir	0.633	0.355 (0.279)
110	Gluconolactone	2.771	1.581 (1.190)
111	Glutamic acid	0.933	1.233 (−0.300)
112	Glycine	2.396	2.482 (−0.086)
113	Glyphosate	1.079	1.704 (−0.625)
114	Guaifenesin	1.699	0.730 (0.968)
115	Haloperidol	−1.854	−2.293 (0.439)
116	Heptabarbital	−0.602	0.082 (−0.684)
117	Hexazinone	2.519	−1.069 (3.588)
118	Histidine	1.659	1.405 (0.254)
119	Hydrocortisone	−0.495	−0.606 (0.111)
120	Hydroflumethiazide	−0.523	0.207 (−0.730)
121	Hydroquinone	1.857	1.156 (0.701)
122	Hydroxyphenamate	1.398	0.303 (1.095)
123	Hydroxyproline	2.558	1.592 (0.965)
124	Hymexazol	1.929	1.420 (0.509)
125	Idoxuridine	0.301	−1.304 (1.605)
126	Imazapyr	1.053	−1.140 (2.193)

Table 1 (*continued*)

No.	Compound name	log(*Sol*)	
		Exp.	Pred. (dif.)
127	Imazaquin	−1.046	−1.802 (0.756)
128	Imazethapyr	0.146	−1.399 (1.545)
129	Iridomyrmecin	0.301	−0.366 (0.667)
130	Isoflurophate	1.188	0.560 (0.628)
131	Isoleucine	1.537	1.667 (−0.130)
132	Isoniazid	2.146	1.060 (1.086)
133	Isophorone	1.079	−0.444 (1.523)
134	Ketanserin	−2.000	−1.503 (−0.497)
135	Khellin	0.017	−0.689 (0.706)
136	Lenacil	−2.222	−1.779 (−0.443)
137	Linuron	−1.125	−1.824 (0.699)
138	Methomyl	1.763	1.318 (0.446)
139	PABA	0.769	0.755 (0.015)
140	*p*-Fluorobenzoic acid	0.079	0.689 (−0.610)
141	Phthalazine	1.699	−0.097 (1.796)
142	Phthalic acid	0.846	0.984 (−0.138)
143	Phthalimide	−0.444	0.138 (−0.582)
144	*p*-Hydroxybenzoic acid	0.699	0.805 (−0.106)
145	Picloram	−0.367	−0.731 (0.364)
146	Picric acid	1.104	0.674 (0.430)
147	Pirimicarb	0.431	1.014 (−0.583)
148	Thionazin	0.057	1.784 (−1.727)

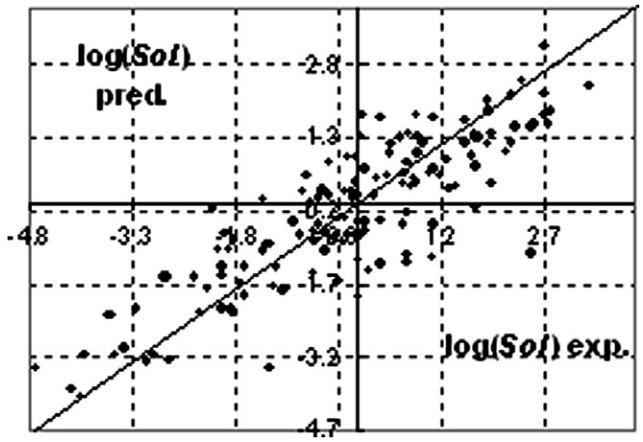

Figure 2. Predicted and experimental aqueous solubilities with Eq. 1 ($N = 100$).

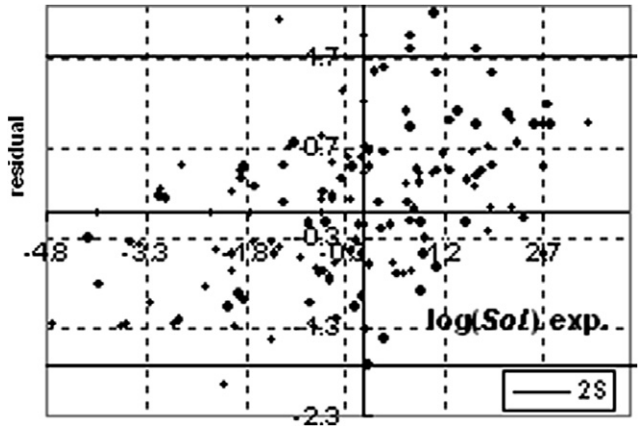

Figure 3. Dispersion plot of the residuals for Eq. 1 ($N = 100$).

solubility than the contribution of the number of acceptors. Figure 4 includes the histograms of the 100 training-compounds for each of the descriptors appearing in the optimal QSPR equation.

The radial distribution function[18] of an ensemble of atoms can be interpreted as the probability distribution of finding an atom in a spherical volume of certain radius, employing different atomic properties such as atomic polarizabilities, volumes, masses, or atomic electronegativities in order to differentiate the contribution of atoms to the property being analysed. For the case of *RDF040e*, the sphere radius is of 4.0 angstroms and the atomic Sanderson electronegativities are employed to distinguish their nature.

GETAWAY type of descriptors[18] match the 3D-molecular geometry and are derived from the elements h_{ij} of the molecular influence matrix (H), obtained through the numerical values of atomic cartesian coordinates. The diagonal elements of H (h_{ii}) are called leverages, and are considered to represent the influence of each molecule atom in determining the whole shape of the molecule. For example, the mantle atoms always have higher h_{ii} values than atoms near the molecule centre. Each off-diagonal element h_{ij} represents the degree of accessibility of the *j*th atom to interactions with the *i*th atom. Descriptor *H3p* employs the off-diagonal elements h_{ij} in its calculation and the atomic polarizabilities in order to quantify the atomic contributions to the aqueous solubilities.

The functional group descriptor *C*-006 takes into consideration the number of CH_2RX fragments present in the molecular structure, therefore including both the effect of size (number of CH_2 groups) and electronegativity

Table 2. Correlation matrix between the descriptors appearing in the QSPR solubility model ($N = 100$)

	RDF040e	*C*-006	*H3p*	D/A	$D/B^{1/2}$	*QXXe*
RDF040e	1	0.4430	0.7243	0.6230	0.6412	0.7504
C-006		1	0.4176	0.2185	0.3036	0.5725
H3p			1	0.7783	0.8621	0.8805
D/A				1	0.9400	0.7007
$D/B^{1/2}$					1	0.7852
QXXe						1

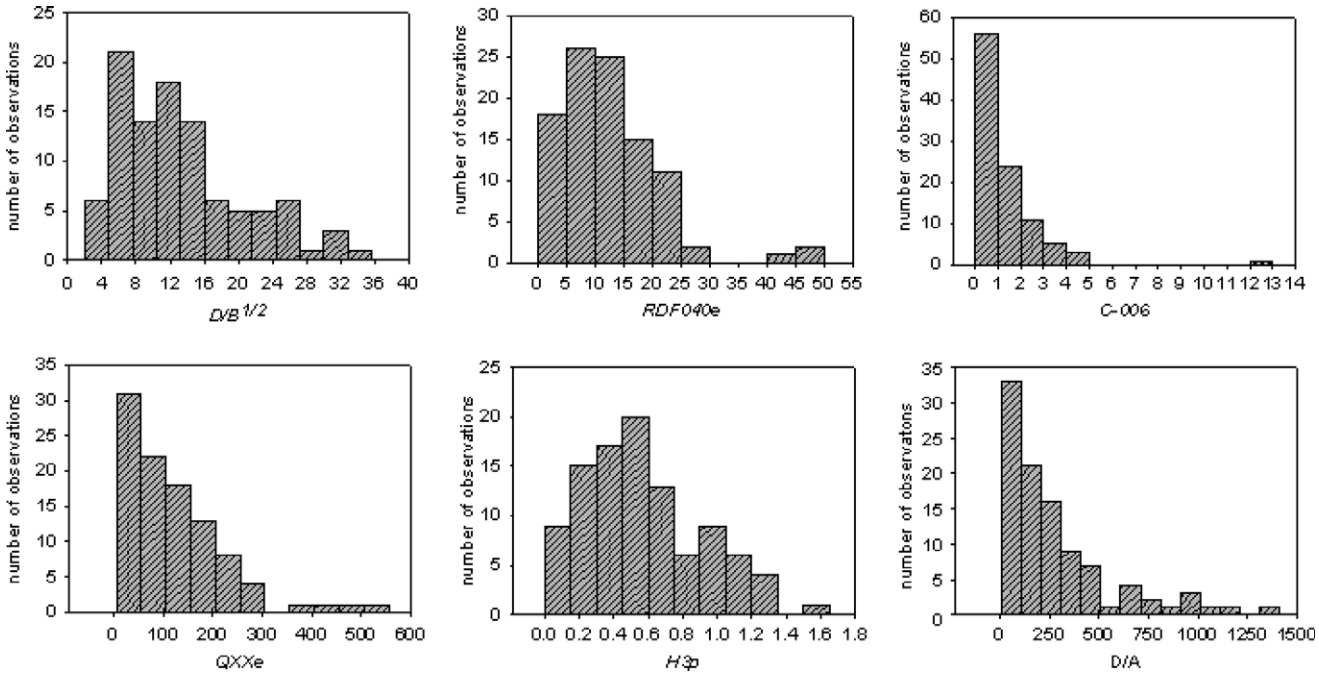

Figure 4. Histograms for the molecular descriptors appearing in the QSPR solubility model ($N = 100$).

(number of X atoms) on the solubility values of the analysed compounds.

Finally, the geometrical parameter *QXXe* (Q_{xx} COMMA2 value) captures three-dimensional molecular features and constitutes a second order-based moment of geometry of the property field density.[19] In present case, this field density is obtained by mapping the Sanderson electronegativities at the atomic positions. The principal component *QXXe* is defined to be invariant with reference to the principal geometric *X*-axis located at the molecular centroid, therefore allowing to properly characterize the shape of each structure along this orientation.

The standardization[20] of the regression coefficients of Eq. 1 allows to assign more importance to the variables having larger absolute standardized coefficients, leading to the following ranking of contributions to log(*Sol*):

$$H3p > \mathrm{D/B^{1/2}} > \mathrm{D/A} > QXXe > C-006 > RDF040e \quad (2)$$

All the molecular descriptors appearing in the QSPR model take positive numerical values. Analysing the more relevant descriptors from Eq. 2, it is expected that increasing numerical values of D/A or *QXXe* (possessing positive regression coefficients) or decreasing values of *H3p* or $D/B^{1/2}$ (with negative coefficients) would tend to predict higher aqueous solubilities (the remaining variables of the model being held invariant). This last conclusion is in complete agreement with the fact that it is possible to correlate the molecular weight of the studied compounds with *H3p* ($R = 0.9121$) or $D/B^{1/2}$ ($R = 0.7962$). It is known that the greater are the molecular weights in homologous series of compounds, the greater will be the solubilities. Also, the degree of accessibility of an atom for inter/intramolecular interactions with another is contemplated through the *H3p* descriptor, being probably this another factor that makes the descriptor more important. $D/B^{1/2}$ and D/A also consider, as well as the other parameters appearing in the QSPR equation such as *C*-006 and *RDF040e*, the influence of the molecular size and the contributions of electronegative atoms in describing the solubilities. Therefore, the greater the number of electronegative atoms in the molecule, the higher the probability of hydrogen-bonding with water and therefore higher predicted aqueous solubilities are achieved.

A next step in present analysis is to further validate the predictive power of the QSPR solubility model found by means of predicting the log(*Sol*) values in a test set containing 48 organic compounds. The predictions are inserted in Table 1 together with the experimental data, demonstrating that it is possible to achieve good estimations in many situations. The statistics achieved in the test set is of $R = 0.7733$, $S = 1.050$.

3. Conclusions

We succeeded in establishing a QSPR model for describing the variation of the aqueous solubilities for 148 heterogeneous organic compounds by considering new descriptors based on the generally accepted Lipinski's rules of drug absorption. The relationship found incorporates the effects of the molecular size, shape and their capability for hydrogen-bonding on the solubility of the compounds. In order to employ the new set of descriptors inspired by the Lipinski's rule to predict oral bioavailability, this study should be complemented by a similar analysis on drug permeability in the future.

4. Dataset and methodology

The experimental aqueous solubilities (*Sol*) (measured at 25 °C and expressed in mg ml^{-1}) for 148 structurally diverse organic compounds were extracted from Merck Index 13th.[21] About 99% of these compounds are 'drug-like', satisfying Lipinski's rule. For modelling purposes, these data are converted into logarithmic units as log(*Sol*) and are presented in Table 1. This table also includes the partitioning of the compounds into a training set (the first 100 chemicals) and a test set (the remaining 48 compounds), selected in such a way that the same structural characteristics appear in both sets. Figure 5 includes a histogram representing the distribution of the 148 aqueous solubilities under study, which suggests that the experimental sample is normally distributed over more than four logarithmic units and can be employed in a regression analysis.

The molecular descriptors derived taking into consideration the Lipinski's rule[2] are based on combinations of the detour index dd from the Chemical Graph Theory (derived as the half sum of the elements of the Detour matrix—DD)[22] together with molecular features such as the number of H donors (D), the number of H acceptors (A) and the number of heteroatoms (H) present in the structure:

$$\mathrm{D/D} = \mathrm{dd}/(\mathrm{D} + 0.1) \qquad \mathrm{D/A} = \mathrm{dd/A}$$
$$\mathrm{D/B} = \mathrm{dd/A} + \mathrm{D} \qquad \mathrm{D/H} = \mathrm{dd/H} \qquad (3)$$

The 0.1 term in the D/D definition is introduced only to prevent dividing by zero, considering that several of the studied compounds do not have any H donor functional group.

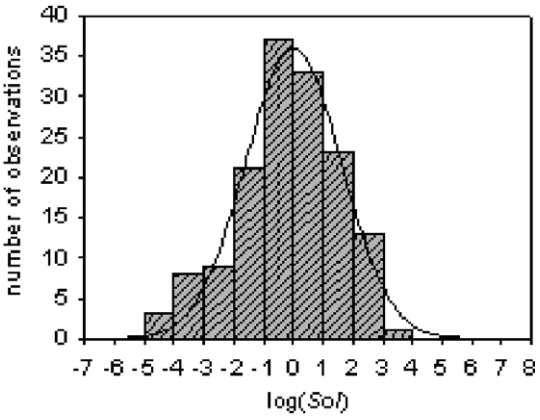

Figure 5. Normal distribution of the experimental log (*Sol*) values under analysis ($N = 148$).

Descriptors' definition took into consideration many reports in the literature which demonstrate linear, polynomial and exponential correlations between dd and the boiling point of alkanes, cycloalkanes and aromatic compounds. Good correlations between combinations of the Detour index and the Wiener index and boiling points were also found in the literature.[23–27] Since the boiling point of compounds from homologous series usually correlates well with molecular weight (MW), we have investigated the relationship between the dd and the MWs of the 148 compounds used for the present study. Inspection of the correlation between dd and MW pushed us to explore possible relationships between the square and cubic roots of dd and the MW. Results of this analysis can be viewed in Figure 6. It is noticeable that cubic root of dd, in the first place, and square root of dd, in the second, are quite better linearly correlated with the molecular weight of the 148 structures, and that there are strong linear correlations between MW and the squares and cubic roots of dd, for the 148 structures ($R = 0.9177$ and $R = 0.9320$, in that order). These indicate a very good correlation, specially noticing the structural diversity of the dataset (the structures of the dataset are available on request). It is clear then that the Detour index may be an appropriate descriptor to explain the differences in the aqueous solubility values that could be explained through the molecular weight of compounds (related to the first parameter in the Lipinski's rule). It can also characterize other molecular properties such as the degree of ramification and cyclation.

However, there are a lot of examples of compounds that, although sharing the same graph and therefore the same dd value, have very different solubilities because of the other three parameters included in Lipinski's rule (number of H donor and acceptors and log *P*). To answer this issue we have included A, B and H in the new descriptors' definition. We also considered the square and cubic roots of the four descriptors above ($\mathrm{D/D^{1/2}}$, $\mathrm{D/D^{1/3}}$, $\mathrm{D/A^{1/2}}$, $\mathrm{D/A^{1/3}}$, $\mathrm{D/B^{1/2}}$, $\mathrm{D/B^{1/3}}$, $\mathrm{D/H^{1/2}}$ and $\mathrm{D/H^{1/3}}$), based on the better correlation between the squares and cubic roots of dd and MW compared to that between dd and MW.

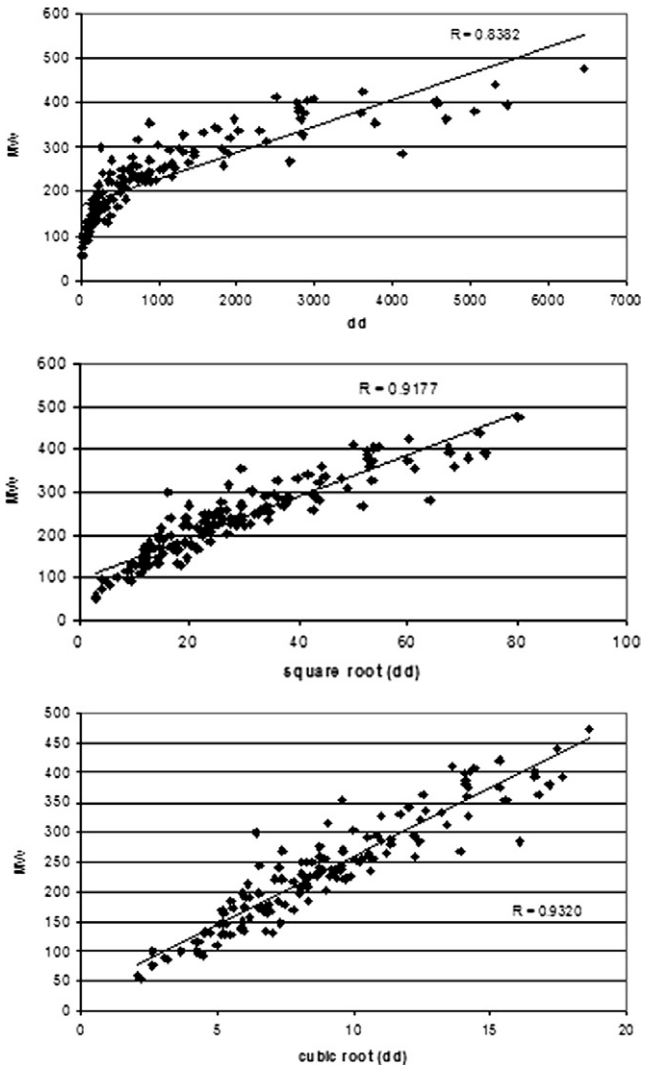

Figure 6. Correlation between MW and dd, square root of dd and cubic root of dd for 148 diverse structures.

The values for D and A were extracted from the Pubchem database developed by US NLM.[28] Pubchem count of H donors and acceptors is based on a substructure- and partial-charge (Gasteiger sigma charges) classification of the acidity/basicity of the hydrogens in the structure. The method was derived from a standard force field atom type classification scheme.

The physicochemical sense of these descriptors is immediate. MW is directly correlated with dd and the solubility tends to decrease, in homologous series, when MW increases. The more H donor and acceptors present in the molecule the more water soluble the compound will be. If no H donor or acceptor is present in the molecule, the water solubility would be jeopardized or non-existent (as is the case of alkanes). Therefore, the defined descriptors will take high values in compounds with slight aqueous solubility, while they will tend to infinity in non-soluble compounds.

For the calculation of the descriptors using the Dragon software, the structures of the compounds are first pre-optimized with the Molecular mechanics force field (MM+) procedure included in Hyperchem 6.03,[29] and the resulting geometries are further refined by means of the semiempirical method PM3 (Parametric Method-3). We chose a gradient norm limit of 0.01 kcal/Å for the geometry optimization. The descriptors derived include several types of variables: constitutional, topological, geometrical, charge, GETAWAY (GEometry, Topology and Atoms-Weighted AssemblY), WHIM (Weighted Holistic Invariant Molecular descriptors), 3D-MoRSE (3D-Molecular Representation of Structure based on Electron diffraction), molecular walk counts, BCUT descriptors, 2D-Autocorrelations, aromaticity indices, Randic molecular profiles, radial distribution functions, functional groups and atom-centred fragments.[11] We excluded from our calculations the empirical and property-based descriptors provided by the software, and also added more quantum-chemical type of descriptors to the pool such as HOMO and LUMO energies, and HOMO–LUMO gap ($\Delta_{\mathrm{HOMO-LUMO}}$).

It is our purpose to search in a large set of D descriptors for an optimal subset of d ones that minimize the standard deviation (S) of the model, defined as follows:

$$S = \sqrt{\frac{\sum_{i=1}^{N} \mathrm{res}_i^2}{N - d - 1}} \tag{4}$$

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 value for i). More precisely, we want to obtain the global minimum of $S(\mathbf{d})$ where $\mathbf{d}$ is a point in a space of $D!/[d!(D-d)!]$ ones. A full search (FS) of optimal variables requires $D!/[d!(D-d)!]$ linear regressions. Some time ago we proposed the Replacement Method (RM)[12–15] that produces QSPR models that are quite close to the FS ones with much less computational work. The RM gives better statistical parameters than the Forward Stepwise Regression (FSR)[20] and similar ones to the more elaborated Genetics Algorithms.[30] The RM 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 $\mathbf{d}$ descriptors $\mathbf{d} = \{X_1, X_2, \ldots, X_d\}$.

As it is common practice in this sort of studies the predictive power of the model has to be validated. The theoretical validation carried out on our models is based on the leave-more-out cross validation procedure (*1–n%-o*),[31] with $n\%$ representing the number of molecules removed from the training set. The number of cases analysed for random exclusion in *1–n%-o* is 90,000. The model was further validated through external validation with a set of 48 compounds extracted from Merck index 13th.

Acknowledgments

We thank the Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET) institution. This research was supported by Agencia de Promoción

Científica y Tecnológica (PICT 06-11985/2004) and University of La Plata, Argentina.

References and notes

1. Yu, H.; Adedoyin, A. *Drug Discov. Today* **2003**, *8*, 852.
2. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3.
3. Hansch, C.; Bjorkroth, J. P.; Leo, A. *J. Pharm. Sci.* **1987**, *76*, 663.
4. Amidon, G. L.; Yalkowsky, S. H.; Anik, S. T.; Valvani, S. C. *J. Phys. Chem.* **1975**, *79*, 2239.
5. Smith, C. J.; Hansch, C. *Food Chem. Toxicol.* **2000**, *38*, 637.
6. Katritzky, A. R.; Maran, U.; Lobanov, V. S.; Karelson, M. *J. Chem. Inf. Model.* **2000**, *40*, 1.
7. Klopman, G.; Wang, S.; Balthasar, D. M. *J. Chem. Inf. Model.* **1992**, *32*, 474.
8. McFarland, J. W. *J. Chem. Inf. Model.* **2001**, *41*, 1355.
9. Pogliani, L. *J. Chem. Inf. Model.* **1996**, *36*, 1082.
10. Talevi, A.; Castro, E. A.; Bruno-Blanch, L. E. *J. Arg. Chem. Soc.*, in press.
11. Dragon 5.0 Evaluation Version, http://www.disat.unimib.it/chm.
12. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M.; González, M. P. *Chem. Phys. Lett.* **2005**, *412*, 376.
13. Duchowicz, P. R.; Castro, E. A.; Fernández, F. M. *MATCH Commun. Math. Comput. Chem.* **2006**, *55*, 179.
14. Duchowicz, P. R.; Fernández, M.; Caballero, J.; Castro, E. A.; Fernández, F. M. *Bioorg. Med. Chem.* **2006**, *16*, 5876.
15. Helguera, A. M.; Duchowicz, P. R.; Pérez, M. A. C.; Castro, E. A.; Cordeiro, M. N. D. S.; González, M. P. *Chemometr. Intell. Lab.* **2006**, *81*, 180.
16. Randic, M. *Croat. Chem. Acta* **1993**, *66*, 289.
17. Randic, M.; Pompe, M. *J. Chem. Inf. Model.* **2001**, *41*, 631.
18. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Model.* **2002**, *42*, 693.
19. Silverman, D. B. *J. Chem. Inf. Model.* **2000**, *40*, 1470.
20. The Merck Index *An Encyclopedia of Chemicals, Drugs, and Biologicals*; 13th ed., Merck& Co.: New Jersey, 2001.
21. Harary, F. *Graph Theory*; Addison-Wesley, 1969.
22. Trinajstić, N.; Nikolić, S.; Lučić, B. *J. Chem. Inf. Model.* **1997**, *37*, 631.
23. Lukovits, I. *Croat. Chem. Acta* **1996**, *69*, 873.
24. Firpo, M.; Gavernet, L.; Castro, E. A.; Toropov, A. A. *J. Mol. Struc.-Theochem.* **2000**, *419*, 501.
25. Castro, E. A.; Tueros, M.; Toropov, A. A. *Comput. Chem.* **2000**, *24*, 571.
26. Devillers, J.; Balaban, A. T. *Topological Indices and Related Descriptors in QSAR and QSPR*; Gordon and Breach Science Publishers, 1999, pp 296–301.
27. U.S. National Library of Medicine, http://pubchem.ncbi.nlm.nih.gov/Pubchem.
28. Hyperchem 6.03 (Hypercube), http://www.hyper.com.
29. Draper, N. R.; Smith, H. *Applied Regression Analysis*; John Wiley& Sons: New York, 1981.
30. So, S. S.; Karplus, M. *J. Med. Chem.* **1996**, *39*, 1521.
31. Hawkins, D. M.; Basak, S. C.; Mills, D. *J. Chem. Inf. Model.* **2003**, *43*, 579.