

*Dedicated to the memory of
Professor Maria Brezeanu (1924–2005)*

NEW MOLECULAR DESCRIPTORS FOR HYDROGEN BONDING POTENTIAL

Laszlo TARKO

Center of Organic Chemistry “C.D.Nenitzescu” – Roumanian Academy, Roumania, Bucharest, Sector 6, Spl. Independenței 202B,
PO Box 35-108, MC 060023, Fax 312160, E-mail: ltarko@cco.ro

Received February 3, 2006

The computation procedure of the HBx (HBf, HBn and HBa) descriptors is presented. Their values do not depend on the size of the molecule, and are not influenced by the difference between intra- and inter-molecular hydrogen bonds. They are proportional to the number of single bonds $Y - H^{\delta}$ ($Y = O, N, S$) and with the value of net charges $+\delta$, as well as with the number of heteroatoms $Z^{-\delta}$ and with the absolute value of the net charges $-\delta$. Some QSPR equations for $\log K_{ow}$ obtained with PRECLAV software are presented. The training set includes molecules with at least one single bond $Y - H^{\delta}$. After the elimination of “outlier by residue” molecules we have found $N = 110$, $p = 6$, $s = 0.4426$, $r^2 = 0.8943$, $F = 146.7$, $r^2_{CV} = 0.8790$. From the group of the six used predictors HBa has the greatest influence on the value of $\log K_{ow}$. Working with a training set of 88 molecules and a testing set of 22 molecules the equation with the highest predictive value contains the HBf descriptor.

INTRODUCTION

The capacity/incapacity of organic molecules to form intra-/inter- molecular hydrogen bonds has a great influence on the physical properties such as the melting point, density, solubility, etc. In the case of some bio-macro-molecules (nucleic acids, proteins, polysugars, etc.) it is the hydrogen bonds that fundamentally determine the shape of the molecules and the way they change in certain phases of the biochemical processes. Therefore, the literature is flooded with papers¹⁻³³ defining, computing and using “hydrogen bonds”, “hydrogen bonding potential”, “hydrogen bonding properties”, etc.

To estimate the capability of molecules to form hydrogen bonds the following molecular descriptors are most frequent:

- a) the number of hydrogen donors-acceptors defined as the total number of hetero atoms included in the OH, NH and NH_2 groups;
- b) the number of hydrogen acceptors defined as the total number of oxygen and nitrogen atoms which were not included at a);
- c) the ration between the number of hydrogen acceptors/donors defined above at a) and b) and the atom number / mass / surface / volume of the molecule.

A priori, it is considered that the ability of molecules to form hydrogen bonds $Y-H\dots Z$ is proportional to the number of heteroatoms Y (bonded to a hydrogen) and the number of heteroatoms Z (bonded or not bonded to a hydrogen atom) present in the molecule. The descriptors at point c) also take into account the size of the molecule.

The molecular descriptors proposed here – called HBx - alike the a) and b) descriptors – relate to the number of hydrogen acceptors/donors but moreover relate to the net charges of the involved atoms. The hydrogen atoms in the molecule to be analyzed, having a net positive charge, are taken into account only if they belong to YH groups, and the heteroatoms Z in the molecule to be analyzed are taken into account only if they have a net negative charge. Moreover, we also consider the difference between the positive charges

(of the H atoms) and the negative charges (of the Z atoms). It has been estimated that this difference can give a measure of the strength of the hydrogen bond. So the molecular descriptors proposed here are considered a measure of the global capacity of the analyzed molecule to form hydrogen bonds $Y - H^{+\delta} \dots Z^{-\delta}$ ($Y = O, N, S$ and $Z^{-\delta} =$ heteroatom). The value of the proposed descriptors is proportional to the number of $H^{+\delta}$ from the $Y - H^{+\delta}$ groups, with the number of heteroatoms $Z^{-\delta}$ and the difference $\delta^+ - \delta^-$.

METHODS AND FORMULAS

For testing the utility of HBx descriptors for describing certain macroscopic properties various QSPR (Quantitative Structure Property Relationship) are computed. The descriptors presence in the final QSPR equation – the equation with the highest predictive power for the values of a certain dependent property – has been considered a valid proof of “utility”. Some QSPR equations are presented. These equations reflect the utility of HBx descriptors for the dependent property “logK_{ow}” – the logarithm of the partition coefficient in the octanol-water system.

As a first step, the computations need the virtual building of the analyzed molecules and the determination of their structures. Here the “structure” is the position in space of the atoms of the analyzed molecule considering the conformer with the minimum potential energy. The virtual construction of molecules and the determining of the minimum energy conformer geometry (the geometry optimization) were realized by PCModel³⁴ molecular mechanics software.

Then the geometry was optimized more rigorously by the Mopac³⁵ quantum mechanics software using the following key words: “pm3 pulay gnorm=0.01 shift=50 geo-ok camp-king bonds vectors mmok nointer”. The net charges were computed using the semi-empirical PM3 method.³⁶

The output file produced by Mopac was the input file for PRECLAV^{37, 38} software which conducts QSPR/QSAR computations.

The last version of PRECALV includes the computation procedure for the HBx descriptors presented below.

For computing the values of the HBx descriptors for the analyzed molecule one has to use a matrix M having $n_{yh} + n_x$ rows and two columns, where n_{yh} is the number of hydrogen atoms with a positive net charge from the $Y - H^{+\delta}$ ($Y = O, N, S$) groups, and n_x is the total number $Z^{-\delta}$ of heteroatoms with a negative net charge.

If the product $n_{yh} \cdot n_x \neq 0$ then the molecule is labeled “able to form hydrogen bonds”. If the product $n_{yh} \cdot n_x = 0$ (*i.e.*, at least one of the two terms is zero) then the molecule is labeled “unable to form hydrogen bonds” and the value of HBx descriptors is null.

The steps of computing the HBx descriptors are:

- the values for the net charges S^{yh} of the hydrogen atoms from the YH groups are placed in the first column of the matrix M if they are positive (there are n_{yh} such values);
- the values for the net charges S^x of all heteroatoms are placed in the second column of the matrix M if they are negative (there are n_x such values);
- the items in the second column are ordered (the most negative value will be on the first line);
- the HBx descriptors are computed using the following formula:

$$HBx = \sum_{i=1}^p \sum_{j=1}^q S_i^{yh} - S_j^x \quad (1)$$

Formula (1) computes the sum of the differences between the net charges $+\delta$ of the hydrogen atoms from YH groups and the net charges $-\delta$ of the heteroatoms.

All the values from the first column are utilized (*i.e.*, in formula (1) $p = n_{yh}$) and some or all the values from the second column (*i.e.*, in formula (1) $q = 1, n_{yh}$ or n_x)

If $q = 1$ then only the first (the largest negative) value from the second column was utilized and the HBf (first) descriptor was computed.

If $q = n_{yh}$ then n_{yh} values from the second column were utilized and the HBn (number) descriptor was computed.

If $q = n_x$ then all the values from the second column were utilized and the HBa (all) descriptor was computed.

If $n_{yh} = n_x = 1$ (e.g., monohydroxylic compounds) then $HB_f = HB_n = HB_a$.

If $n_{yh} < n_x$ (e.g., amides) then $HB_f \neq HB_n \neq HB_a$.

If $n_{yh} \geq n_x$ (e.g., aliphatic amines) then $HB_f \neq HB_n = HB_a$.

The values of the HBx descriptors:

- do not depend on the size of the molecule;
- do not reflect any differences between molecular areas;
- do not reflect any differences between intra- and inter- molecular hydrogen bonds;
- are proportional to the number of single bonds $Y - H^{+\delta}$ ($Y = O, N, S$) and to the value of net charges $+\delta$;
- are proportional to the number of heteroatoms $Z^{-\delta}$ and to the absolute value of net charges $-\delta$.

The HBx descriptors are only three descriptors from a group of 400 whole molecule descriptors computed by PRECLAV. The HBx descriptors are useful only if they win the mathematical competition with the other descriptors. Using only the “significant” descriptors PRECLAV computes thousands of QSPR equations, *i.e.*, multilinear formulas of the dependent property P:

$$P_{cal} = c_0 + \sum c_k \cdot p_k \quad (2)$$

where c_k coefficients (weighted factors) are computed by the Ordinary Least Squares Method. The program computes successively equations with $k = 2, 3, \dots, 10$ predictors. At the same time with the k value enhancement, the r^2 correlation between the observed and the computed values of P continuously rises. On the other hand, the value of the quality function Q increases until it reaches a maximum and then it diminishes:

$$Q = K_{CV} \cdot (N - k) / N \quad (3)$$

where: N is the number of the training set molecules;

k is the number of predictors of equation (2);

K_{CV} is the cross-validated Kendall ranks correlation between the observed and the computed values of P property; the program uses LOO (leave one out) cross-validation procedure.^{39, 40}

From the thousands of the computed QSPR equations, PRECLAV uses for prediction only the type (2) equations with the highest value of quality function (3). The equation used for prediction usually includes less than eight predictors. The HBx descriptors are obviously correlated. Therefore, any QSPR equation will contain only one of them.

The relative influence I of a certain predictor on dependent property values was computed by the following formula:

$$I = (R^2 - r^2) / (1 - r^2) \quad (4)$$

where: R^2 is the square of Pearson correlation between the P_{obs} values and the P_{calc} values (computed by the k predictors QSPR);

r^2 is the square of Pearson correlation between the P_{obs} values and the P_{calc} values (computed by the $k-1$ predictors QSPR, *i.e.*, the equation without the analyzed predictor).

After I computations – one for each predictor of the final QSPR – the values of I are normalized by the highest of them (the highest value for I becomes 1000). The predictors with large enough values of I ($I > 400$) may be considered having “large relative influence on dependent property”.

The computations were conducted on a Pentium4 / 2400 MHz / 1024 RAM.

RESULTS AND DISCUSSION

Table 1 presents the HBx values for five simple molecules. The computed (positive) values of S^{yh} , the net charges of the hydrogen atoms in YH groups, are much larger if Y is O than if Y is N. The computed (negative) values of S^x , the net charges of the oxygen atoms, are much more negative for NO_2 and CO groups than for OH group. The value of S^x , the net charge of the nitrogen atom of the NH group, for 4-Methyl-imidazole is very positive (0.3083). Thus, it is not present in matrix M. The value of the S^x , the net charge of the nitrogen of the NH group, for succinimide is weakly negative. Thus, it is present in matrix M.

To obtain the QSPR equations we used – first analysis – a training set of 113 molecules (Table 2, column 1, 2 and 3). The observed values of $\log K_{ow}$ have been taken from literature and various databases.⁴¹⁻⁴³ All the molecules from Table 2 have Y – H^δ ($n_{yh} > 0$) chemical bonds. The HBx descriptors are null in case of molecules without any heteroatoms with negative charges ($n_x = 0$) (some aromatic amines, thiols and pyrrole).

We have noticed that in various computation stages the HBf, HBn and HBa descriptors appear as predictors in type (2) equations with the highest r^2 value. These equations contain 2, 3, 4 and 6 descriptors.

Table 1
Computed values of HBx descriptors

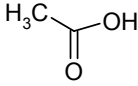
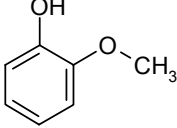
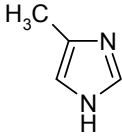
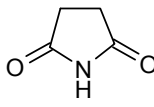
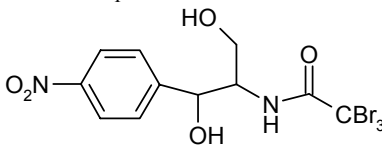
Molecule	M	matrix	HBf	HBn	HBa
	S _{vh}	S _x			
Acetic acid 	0.1935	-0.3345 0.2718	0.5280	0.5280	0.9932
Guaiacol 	0.2073	-0.2242 0.2066	0.4315	0.4315	0.8453
4-Methyl-imidazole 	0.0694	-0.1130	0.1825	0.1825	0.1825
Succinimide 	0.1229	-0.3319 0.3319 -0.0628	0.4548	0.4548	1.0954
Tribromoamphenicol 	0.1962 0.1876 0.0916	-0.5969 0.5966 -0.3317 -0.3035 -0.2995 -0.0077	2.2663	6.0021	9.2608

Table 2
Observed/computed values of $\log K_{ow}$

Molecule	CAS	Obs. $\log K_{ow}$	Comp. $\log K_{ow}$		
			by #1	by #2	by #3
Formic acid	64-18-6	-0.54	-0.868	-0.807	2.731
1, 1-Dimethyl-3-(<i>p</i> -bromo-phenyl)urea	20940-43-6	2.19	2.582	2.784	
Adipic acid	124-04-9	0.08	0.229	-0.045	
Ethylene Bromohydrin	540-51-2	0.23	0.366	1.128	
Hydroxy acetic acid	79-14-1	-1.11	-1.123	-1.262	
Diethanolamine	111-42-2	-1.43	-0.251	-	
O-Methyl carbamate	598-55-0	-0.66	-1.270	-1.196	
2-Chloroacetamide	79-07-2	-0.53	-0.390	-0.105	

Table2 (continues)

Table2 (continued)

Crotonic acid	3724-65-0	0.72	0.463	0.643	
Cinnamic acid	621-82-9	2.13	2.001	2.029	
Benzanilide	93-98-1	2.62	2.841	2.942	
N-Phenylbenzylamine	103-32-2	3.13	3.119	3.159	
Imidazole	288-32-4	-0.08	0.377	-0.068	0.325
2-Phenylimidazole	670-96-2	1.88	2.087	2.141	
Benzimidazole	51-17-2	1.32	1.435	1.589	1.707
Pyrrole	109-97-7	0.75	1.499	0.743	
5-bromo-indole	10075-50-0	3.00	2.321	2.600	
Uracil	66-22-8	-1.07	-0.387	-0.076	
Barbituric acid	67-52-7	-1.47	-1.562	-1.004	
Formanilide	103-70-8	1.12	1.429	1.496	1.570
Benzoyl hydrazine	613-94-5	0.19	0.523	1.023	
Thiabendazole	148-79-8	2.47	2.513	2.132	
6-Nitrobenzimidazole	94-52-0	1.50	1.207	1.149	1.207
4-Chlorophenoxyacetic acid	122-88-3	2.25	1.811	1.977	
<i>m</i> -Dimethylaminobenzamide	33225-17-1	0.95	1.364	1.279	
Tribromoamphenicol	49648-42-2	2.17	2.044	2.024	
Dexpropranolol	5051-22-9	3.48	3.515	2.984	
Propanol	71-23-8	0.25	0.706	0.562	0.598
2-Propanol	67-63-0	0.05	0.460	0.154	
Butanol	71-36-3	0.88	1.033	0.995	
2-Methyl-1-propanol	78-83-1	0.76	0.747	0.656	
2-Butanol	78-92-2	0.61	0.800	0.647	
Pentanol	71-41-0	1.51	1.338	1.377	
2-Methyl-2-butanol	75-85-4	0.89	0.783	0.647	
Hexanol	111-27-3	2.03	1.634	1.723	1.634
2-Hexanol	626-93-7	1.76	1.474	1.403	1.358
3-Hexanol	623-37-0	1.65	1.366	1.263	
Heptanol	111-70-6	2.62	1.934	2.055	
Octanol	111-87-5	3.00	2.253	2.384	
Cyclopentanol	96-41-3	0.71	1.325	1.358	0.768
<i>m</i> -Cresol	108-39-4	1.96	1.799	1.758	
Phenol	108-95-2	1.46	1.554	1.445	
Ethanol	64-17-5	-0.31	0.382	0.108	
<i>tert</i> -Butanol	75-65-0	0.35	0.425	0.229	
Isoamyl Alcohol	123-51-3	1.16	1.116	1.116	
1, 3-dichloroisopropanol	96-23-1	0.20	0.697	1.194	
Thymol	89-83-8	3.30	2.180	2.280	
(-)-Diethyl D-tartrate	13811-71-7	-0.29	-0.675	-0.465	
O-Ethyl carbamate	51-79-6	-0.15	-0.657	-0.496	
O-Phenyl carbamate	622-46-8	1.08	0.907	1.081	
Acetaldoxime	107-29-9	-0.13	-0.198	-0.456	
Trichloroacetamide	594-65-0	1.04	0.040	0.679	
Ethane thiol	75-08-1	1.18	0.852	1.260	
Pinacone	76-09-5	-0.68	0.188	-0.011	
Chlorohydrin	96-24-2	-1.00	-0.101	0.055	0.111
Valeramide	626-97-1	0.35	0.459	0.526	
Phenylurea	64-10-8	0.83	0.791	0.642	
Phenylthiourea	103-85-5	0.71	1.367	1.668	
1, 3-Diacetylurea	638-20-0	-0.68	-0.846	-0.657	
N, N'-diethylthiourea	105-55-5	0.57	0.644	0.880	
Menthol	89-78-1	3.30	2.255	2.465	1.889
Urea	57-13-6	-2.11	-2.687	-2.741	
Acetamide	60-35-5	-1.26	-1.114	-1.179	-1.455
Succinimide	123-56-8	-1.89	0.109	-	
Ethylene glycol	107-21-1	-1.36	-0.280	-0.682	
<i>o</i> -Cresol	95-48-7	1.95	1.744	1.694	
<i>p</i> -Cresol	106-44-5	1.94	1.810	1.737	
1-naphthol	90-15-3	2.85	2.254	2.184	
Resorcinol	108-46-3	0.80	1.185	0.883	
Hydroquinone	123-31-9	0.59	1.193	0.835	
Guaiacol	90-05-1	1.32	1.528	1.311	

Table2 (continues)

Table2 (continued)

Eugenol	97-53-0	2.27	1.916	1.767	
Phloroglucinol	108-73-6	0.16	0.393	0.015	0.067
Pyrogallol	87-66-1	0.14	0.280	-0.055	
Vanillin	121-33-5	1.21	1.386	1.468	
Acetanilide	103-84-4	1.16	1.704	1.540	
Methacetin	51-66-1	1.03	1.903	1.312	
Phenacetin	62-44-2	1.58	2.239	2.033	
Diphenyl amine	122-39-4	3.50	2.803	2.744	
Aniline	62-53-3	0.90	1.478	1.316	
Morphine	57-27-2	0.89	1.882	1.897	
Acetic acid	64-19-7	-0.17	-0.429	-0.515	-0.620
<i>p</i> -(<i>tert</i> -amyl)phenol	80-46-6	3.83	2.422	-	
Pyrocatechol	120-80-9	0.88	1.166	0.869	1.002
Coniine	458-88-8	2.13	1.810	2.111	
Butanoic acid	107-92-6	0.79	0.479	0.590	0.692
Amyl amine	110-58-7	1.49	1.075	1.261	
Benzenethiol	108-98-5	2.52	1.944	2.102	
4-methyl-imidazole	822-36-6	0.23	0.668	0.459	
<i>n</i> -Butyl amine	109-73-9	0.97	0.692	0.848	1.361
Ethyl amine	75-04-7	-0.13	-0.280	-0.136	
Hexyl amine	111-26-2	2.06	1.426	1.625	
Propyl amine	107-10-8	0.48	0.250	0.379	0.823
Monoacetyl hydrazine	1068-57-1	-1.58	-1.632	-1.553	
1, 3-Dimethyl-2-nitroguanidine	101250-97-9	-0.70	-1.632	-1.599	
3-Methylindole	83-34-1	2.60	2.171	2.124	2.174
Propanoic acid	79-09-4	0.33	0.047	0.070	
Furan-2-carboxylic acid	26447-28-9	0.64	0.884	1.049	
5-Formyluracil	1195-08-0	-1.03	-1.136	-0.651	
Uric acid	69-93-2	-2.17	-2.398	-2.399	
Furan-3-carboxylic acid	488-93-7	1.03	0.839	0.914	
Furan-3-carboxamide	609-35-8	0.09	0.242	0.350	
Pyrrrole-2-carboxylic acid	634-97-9	0.85	0.614	0.484	
Thiophene-2-carboxylic acid	527-72-0	1.57	1.316	1.566	
Thiophene-3-carboxylic acid	88-13-1	1.50	1.331	1.599	
Glycerin	56-81-5	-1.76	-1.088	-1.453	
LSD	50-37-3	2.95	3.877	3.706	3.476
Thiourea	62-56-6	-1.08	-1.319	-1.088	
5-Hydroxy-1-naphthalenesulfonic acid	117-59-9	-0.17	0.274	0.038	
Hymexazol	10004-44-1	0.46	0.506	0.727	
Vitamin C	50-81-7	-1.85	-2.427	-2.262	-0.823
1-(2-Chloroethyl)-3-cyclohexyl-1-nitrosourea	13010-47-4	2.83	1.889	2.184	
Methamidophos	10265-92-6	-0.80	-1.479	-1.080	

The type (2) equation with the highest quality function (3) contains five predictors:

QSPR Equation #1

$$c_0 = 2.2182$$

$$c_1 = -.0347$$

$$p_1 - \text{percent of nitrogen (I = 676)}$$

$$c_2 = 0.0005$$

$$p_2 - \text{moment of inertia C (I = 611)}$$

$$c_3 = -3.4117$$

$$p_3 - \text{Balaban topologic index / heavy atom number ratio (I = 599)}$$

$$c_4 = -.0812$$

$$p_4 - \text{percent of carbon} \cdot \text{average charge for C atoms product (I = 394)}$$

$$c_5 = -0.3786$$

$$p_5 - \text{HBa descriptor (I = 1000)}$$

From the five predictors, p_1 has the weakest correlation with the observed values of $\log K_{ow}$ ($r^2 = 0.1214$ – here r^2 is the square of Pearson correlation between the p_1 and the $\log K_{ow}$ values), and p_2 and p_3 are the most intercorrelated ($r^2 = 0.3639$ – here r^2 is the square of Pearson correlation between the p_2 and the p_3 values).

There is good agreement between the observed / computed values of $\log K_{ow}$ ($s = 0.5430$, $r^2 = 0.8514$, $F = 123.7$, $r^2_{CV} = 0.8345$, $Q = 0.7579$). Here r^2_{CV} is the square of cross-validated Pearson correlation between observed / computed values of $\log K_{ow}$ and s is standard error of estimation.

Hba has the largest influence (inverse proportional) on the values of $\log K_{ow}$. If molecules form hydrogen bonds more easily the value of $\log K_{ow}$ is lower.

In case of diethanolamine, succinimide and *p*-(*t*-amyl)-phenol the difference D between the observed and the computed value of $\log K_{ow}$ is significant ($D > 2 \cdot s$, Table 2, column 3 and 4). We consider these molecules to be outliers by residue.

When we eliminate them a training set of 110 molecules is obtained. Using this set we obtain a different QSPR equation for the highest quality function Q . This equation has six predictors.

QSPR Equation #2

$$c_0 = 3.9088$$

$$c_1 = -.0425$$

p_1 – percent of nitrogen ($I = 878$)

$$c_2 = -.0276$$

p_2 – percent of oxygen ($I = 502$)

$$c_3 = 0.0004$$

p_3 – moment of inertia B ($I = 524$)

$$c_4 = -4.638$$

p_4 – Balaban topologic index / heavy atom number ratio ($I = 959$)

$$c_5 = -.2972$$

p_5 – QSPR of molecular orbital energies ($I = 158$)

$$c_6 = -.3706$$

p_6 – HBa descriptor ($I = 1000$)

From the six predictors, p_1 has the weakest correlation with the observed values of $\log K_{ow}$ ($r^2 = 0.1149$), and p_2 and p_5 are the most intercorrelated ($r^2 = 0.3082$).

There is a better agreement between the observed / computed values of $\log K_{ow}$ ($s = 0.4426$, $r^2 = 0.8943$, $F = 146.7$, $r^2_{CV} = 0.8790$, $Q = 0.7751$, Table 2, column 3 and 5).

Hba continues to have the largest influence (inversely proportional) on the values of $\log K_{ow}$.

The predictors with a high value of I ($I > 400$) may be considered very useful in calculating the value of $\log K_{ow}$. These predictors are useful as they correlate well enough with $\log K_{ow}$ and do not correlate with the other predictors – see formula (4). Each “useful” predictor explains (quite) a lot of the $\log K_{ow}$ variation and, at the same time, a different thing as the other predictors. Consequently, the presence in the final equation of some descriptors only slightly correlated with $\log K_{ow}$ (for instance the percentage of nitrogen) is not that surprising. Using for prediction of a set of descriptors relatively superficially correlated with the dependent property but slightly intercorrelated is frequent in QSAR practice.

We will also mention that there have been QSAR studies – that do not make the subject of this article – where the dependent property was “melting point” and “Gibbs free energy of hydration”. In these situations, the descriptors HBx have lost the mathematical competition in favor of some type a) and b) (see the INTRODUCTION) and topological descriptors.

We also wanted to see what happens – from the discussed point of view – if a testing set is used. We arrange the molecules of the training set used in study #2 (110 molecules) after the value of $\log K_{ow}$, starting with the lowest value. We chose for the testing set the ranked molecules 3, 8, 13, 18, 23, ... , 103, 108 (22 molecules, Table 2, column 6). The remaining molecules (88 molecules) were the training set.

When one has a testing set, PRECLAV selects the “significant” descriptors *via* a different procedure. From the point of view of the descriptors now identified as “significant” the training set has to be representative sample for the testing set + training set joint.³⁷ From this point of view the most “performant” descriptor is “Shannon index of topologic distances”. The equation obtained in such conditions and used for making predictions has six descriptors.

QSPR Equation #3

$$c_0 = .8603$$

$$c_1 = -.0360$$

p_1 – percent of nitrogen
 $c_2 = .0004$
 p_2 – moment of inertia B
 $c_3 = .6079$
 p_3 – Shannon index of topologic distances
 $c_4 = -1.2982$
 p_4 – positive area - negative area gap / Molecular surface area ratio
 $c_5 = .0868$
 p_5 – heat of formation / bond number ratio
 $c_6 = -1.1285$
 p_6 – Hbf descriptor

In this case the value of influence I is not relevant. From presented point of view, it is significant that Hbf is present. Descriptor p_1 (Equation #1), descriptors p_1 and p_2 (Equation #2) and descriptor p_1 (Equation #3) – percents of nitrogen/oxygen – may be considered descriptors of type c) (see INTRODUCTION section).

For the testing set molecules the agreement between the observed/computed values of $\log K_{ow}$ is reasonable ($r^2 = 0.8259$, $s = 0.5838$). The computed values are also ordered reasonably well ($r_{Kendall} = 0.8268$).

According to PRECLAV statistical formulas, the “large” observed testing set values of $\log K_{ow}$ are $\log K_{ow} > 1.50$, and the “low” are $\log K_{ow} < 0$. The “large” computed values of $\log K_{ow}$ are $\log K_{ow} > 1.500$, and the “low” are $\log K_{ow} < 0.200$. Thus, using the Hbf descriptor, PRECLAV correctly identifies five out of seven of the molecules with “large” $\log K_{ow}$ values (Table 2, columns 3 and 6, bold) and five out of six of the molecules with “low” $\log K_{ow}$ values (Table 2, columns 3 and 6, italics).

CONCLUSIONS

The way they are defined, the HBx descriptors measure the capacity of a molecule to form intramolecular hydrogen bonds and/or intermolecular hydrogen bonds $Y - H^{+\delta} \dots Z^{-\delta}$ ($Y = O, N, S$ and $Z^{-\delta} =$ heteroatom).

In case the dependent property is $\log K_{ow}$ and all the molecules of the training set have single bonds $Y - H^{+\delta}$, the HBa descriptor is present in the QSPR equations having the highest value of the quality function Q. Other HBx descriptors appear in the QSPR equations with the highest quality function r^2 . When we used a testing set the equations with the highest predictive power includes the Hbf descriptor.

The presence of HBx descriptors in the equations having the highest predictive power is a proof of their utility for estimating the value of $\log K_{ow}$ for molecules containing single bonds $Y - H^{+\delta}$.

The HBa descriptor has the highest influence (inverse proportional) on the values of $\log K_{ow}$. The physical meaning of HBa is correct: the molecules more inclined to form hydrogen bonds have lower values of $\log K_{ow}$.

REFERENCES

1. J. Cheney, B. V. Cheney and W. G. Richards, *Biochim. Biophys. Acta*, **1988**, 954, 137.
2. H. van de Waterbeemd and M. Kansy, *Chimia*, **1992**, 46, 299.
3. N. el Tayar, B. Testa and P. – A. Carrupt, *J. Phys. Chem.*, **1992**, 96, 1455.
4. K. H. Kim, *Quant Struct-Act Relat.*, **1993**, 12, 232.
5. E. G. Chikhale, K. Y. Ng, P. S. Burton and R. T. Borchardt, *Pharm. Res.*, **1994**, 11, 412.
6. A. M. ter Laak, R. – S. Tsai, G. M. D.- O. den Kelder, P.- A. Carrupt and B. Testa, *Eur. J. Pharm. Sci.*, **1994**, 2, 373.
7. R. O. Potts and R. H. Guy, *Pharm. Res.*, **1995**, 12, 1628.
8. W. J. Pugh and M. S. Roberts, *Int. J. Pharm.* **1996**, 138, 149.
9. A. Douhal, *Science*, **1997**, 276, 221.
10. M. H. Abraham, F. Martins and R. C. Mitchell, *J. Pharm. Pharmacol.*, **1997**, 49, 858.
11. J. P. M. Lommerse, S. L. Price and R. Taylor, *J. Comput. Chem.*, **1997**, 18, 757.
12. M. Masella and J. P. Flament, *Bull. Soc. Chim. France*, **1997**, 134, 439.
13. I. Nobeli, S. L. Price, J. P. M. Lommerse and R. Taylor, *J. Comput. Chem.*, **1997**, 18, 2060.
14. R. P. Apaya, M. Bondi and S. L. Price, *J. Comput-Aided Mol. Design*, **1997**, 11, 479.
15. L. J. Bain, J. B. McLachlan, and G. A. LeBlanc, *Environ. Health Persp.*, **1997**, 105, 812.
16. C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, **1997**, 23, 3.

17. P. S. Kushwaha and P. C. Mishra, *Int. J. Quant. Chem.*, **2000**, 76, 700.
18. R. Vargas, J. Garza, D. A. Dixon and B. P. Hay, *J. Am. Chem. Soc.*, **2000**, 122, 4750.
19. J. Du Plessis, W. J. Pugh, A. Judefeind and J. Hadgraft, *Eur. J. Pharm. Sci.*, **2001**, 13, 135.
20. M. Buck, and M. Karplus, *J. Phys. Chem. B*, **2001**, 105, 11000.
21. E. Gancia, J. G. Montana and D. T. Manallack, *J. Mol. Graph. Modell.*, **2001**, 19, 349.
22. H. Kubinyi, *Helv. Chim. Acta*, **2001**, 84, 513.
23. J. S. Barker, C. K. Hattotuwigama and M. G. B. Drew, *Pure Appl. Chem.*, **2002**, 74, 1207.
24. X. – Q. Chen, S. J. Cho, Y. Li and S. Venkatesh, *J. Pharm. Sci.*, **2002**, 91, 1838.
25. D. F. Veber, S. R. Johnson, H. Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *J. Med. Chem.*, **2002**, 45, 2615.
26. T. E. Exner, M. Keil and J. Brickmann, *J. Comp. Chem.* **2002**, 23, 1176.
27. R. Perkins, H. Fang, W. Tong and W. J. Welsh, *Environ. Toxicol. Chem.*, **2002**, 22, 1666.
28. I. Zamora, T. Oprea, G. Cruciani, M. Pastor and A. – L. Ungell, *J. Med. Chem.*, **2003**, 46, 25.
29. C. Subhash, S. C. Basak, D. Mills, D. M. Hawkins and H. A. el-Masri, *Risk Analysis*, **2003**, 23, 1173.
30. H. Guo, R.F. Beahm and H. Guo, *J. Phys. Chem. B*, **2004**, 108, 1806.
31. A. V. Morozov, T. Kortemme, K. Tsemekhman and D. Baker, *Proc. Natl. Acad. Sci. USA.*, **2004**, 101, 6946.
32. A. Crammers and S. Parkin, *Cryst. Eng. Comm.*, **2004**, 6, 168.
33. A. T. Lithoxidou and E. G. Bakalbassis, *J. Phys. Chem. A*, **2005**, 109, 366.
34. PCModel v. 9.0, Serena Software, Box 3076, Bloomington, IN, USA.
35. MOPAC93, J. J. P. Stewart, Fujitsu Limited, Tokyo, Japan, 1993.
36. J. J. P. Stewart, *J. Comput. Chem.* **1989**, 10, 209.
37. L. Tarko, *Rev. Chim. (București)*, **2005**, 56, 639.
38. PRECLAV program is available from Center of Organic Chemistry (CCO) - Bucharest, Roumanian Academy; Director CCO
E-mail address pfilip@cco.ro
39. L. Eriksson, E. Johansson, M. Muller and S. Wold, *J. Chemometrics*, **2000**, 14, 599.
40. M. Stone, *J. Roy. Stat. Soc. B*, **1977**, 38, 44.
41. KowWin v. 1.67, included in EPI suite v. 3.12, developed by the U.S. Environmental Protection Agency.
42. D. Eros, I. Kovcsdi, L. Orfi, K. Takacs-Novak, G. Acsady and G. Keri, *Curr. Med. Chem.*, **2002**, 20, 1819.
43. R. F. Rekker and H. M. de Kort, *Eur. J. Med. Chem.*, **1979**, 14, 479.