

A NOVEL QSAR APPROACH IN MODELING HYDROPHOBICITY OF A SET OF FLAVONOIDS

ALEXANDRA MARIA HARSA^a

ABSTRACT. A novel QSAR approach, based on correlation weighting and alignment over a hypermolecule, that mimics the investigated correlational space, was performed on a set of 40 flavonoids, downloaded from the PubChem database. The best models describing log P of this set of flavonoids were validated in the external test set and in a new version of prediction by using similarity clusters.

Key-words: QSAR, log P, flavonoids, correlation weighting, similarity clustering.

1. INTRODUCTION

Quantitative structure-activity relationships (QSAR) mathematically relate descriptors of a molecular structure to a biological activity (or a physico-chemical property involved in that activity). A structure-activity relationship can indicate which features of a given molecule are responsible for its activity, thus making possible to synthesize new and more potent compounds with enhanced biological activity [1,2]. QSAR analysis is based on the assumption that the activity of compounds is a function of their structural characteristics [3].

Molecular similarity was extensively and successfully used in drug discovery, often to compare molecules in the absence of other mechanistic information [4-6]. Reasons for the increasing popularity of similarity based methods include technological advances in high throughput screening and synthesis in the last decade and the need of applications of computer based methods in compound selection and activity evaluation [7]. Similarity search [8,9] and clustering methods [9,10] can be used to classify compounds into structural groups [11] and in prediction of biological activities as well [12]. The paradigm of similarity-based QSAR approaches was explicitly enounced by Johnson and Maggiora [13,14]: "*molecules that are structurally similar likely will have similar properties*". Thus, when the activity of a set of molecules is

^a Babes-Bolyai University, Department of Chemistry, Faculty of Chemistry and Chemical Engineering, Arany Janos 11, 400028, Cluj, Romania, maria_sanda_13@yahoo.com

unknown, one can predict that activity by taking into account the similarity values between the molecules under study and the molecules of a data set whose activities are known.

Studies of similarity [15] in chemical structures can be overtaken by using topological indices [16]. Among thousands of such topological descriptors, the Cluj indices have been defined by Diudea [17,18], as follows.

A Cluj fragment $CJ_{i,j,p}$ collects vertices v lying closer to i than to j , the endpoints of a path $p(i,j)$. Such a fragment collects the vertex proximities of i against any vertex j , joined by the path p , with the distances measured in the subgraph $D_{(G-p)}$, as shown in the following equation:

$$CJ_{i,j,p} = \left\{ v \mid v \in V(G); D_{(G-p)}(i, v) < D_{(G-p)}(j, v) \right\}$$

In graphs containing rings, more than one path could join the pair (i, j) , thus resulting more than one fragment related to i (with respect to j and a given path p). The entries in the Cluj matrix are taken, by definition, as the maximum cardinality among all such fragments:

$$[UCJ]_{i,j} = \max_p |CJ_{i,j,p}|$$

Indices I_e and I_p are calculated, from the Cluj topological matrices UCJ_e , and UCJ_p , respectively (see above), as half sum of matrix entries. In the above symbols, e refers to edge-calculated matrix while p refers to the path-calculated ones.

Correlation weighting [19] was used as a weighting scheme applied to local descriptors. Within this paper, we used the correlation weighting in the frame of a hypermolecule built on the overall set of structures taken in study (see below).

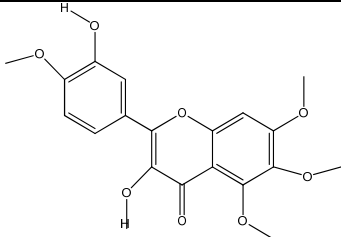
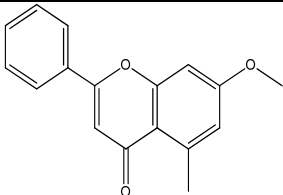
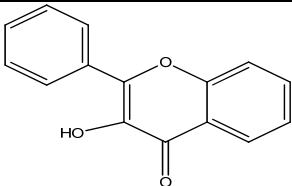
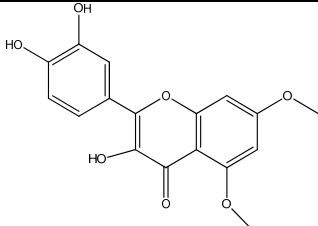
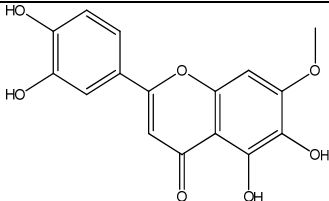
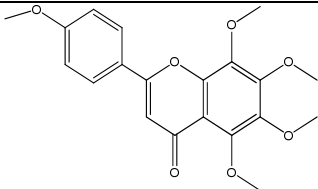
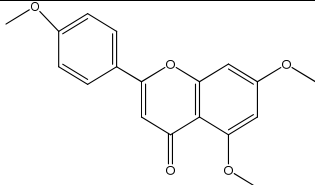
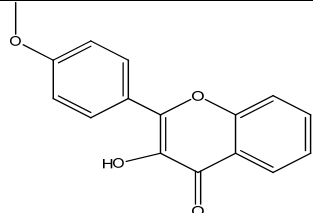
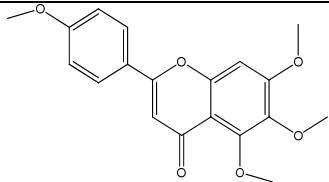
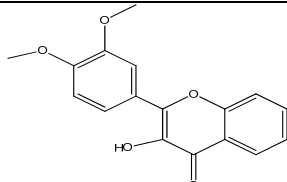
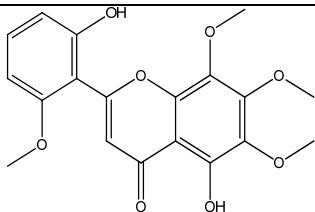
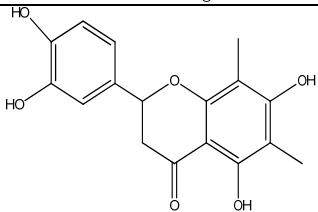
The article proposes a new approach called "direct prediction" which develops clusters of similar structures aimed to be quasi-congeneric subsets in predicting of a biological activity.

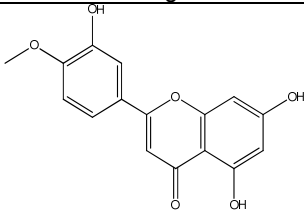
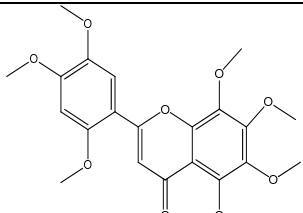
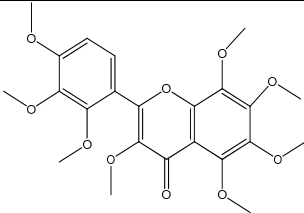
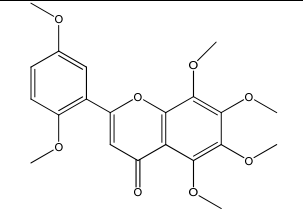
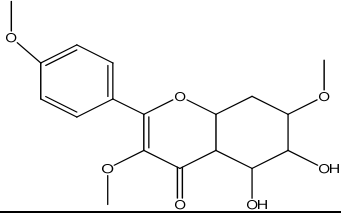
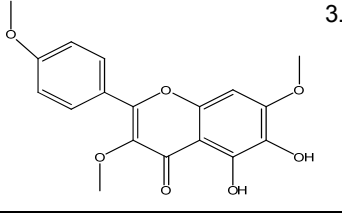
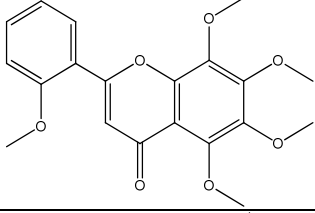
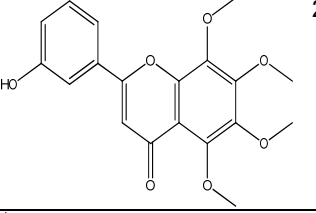
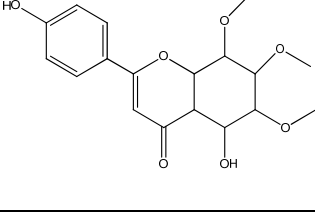
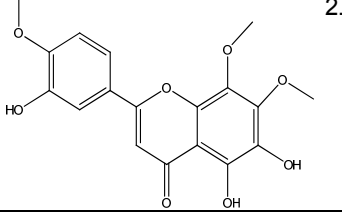
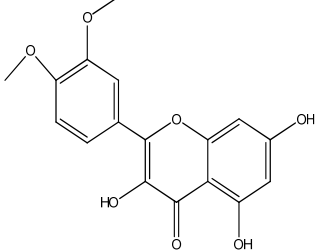
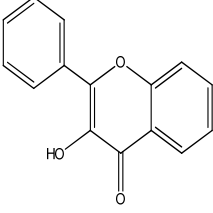
2. STRUCTURAL DATA

A set of 40 flavonoids were taken from PubChem Database (Table 1) and were divided into a training set (30 molecules) and a test set (ten molecules), taken randomly. The property chosen for modeling was log P (see Table 1), the (calculated) partition coefficient between *n*-octanol and water, a measure of hydrophobicity, involved in the passive transport of a drug molecule through cell membrane.

A hypermolecule (Figure 1) was built up as the union of all structural features in all 40 molecules under study. The hypermolecule is considered to mimics the investigated statistical hyperspace [20].

Table 1. Flavonoid molecular structures and their log P (from PubChem)

Structures		log P	Structures		log P
Training set					
1		2.8	2		3.5
3		3.4	4		1.6
5		2.3	6		3
7		2.2	8		3.4
9		3.1	10		3.3
11		2.9	12		2.8

Structures		log P	Structures		log P
Training set					
13		1.7	14		3
15		3.2	16		3
17		2.9	18		3.1
19		3	20		2.7
21		3.1	22		2.6
23		2.2	24		2.6

Structures		log P	Structures		log P
Training set					
25		2.6	26		3.8
27		1.5	28		2.3
29		1.8	30		3.1
Test set					
31		2.6	32		1.5
33		1.7	34		3.2
35		2.1	36		2.8

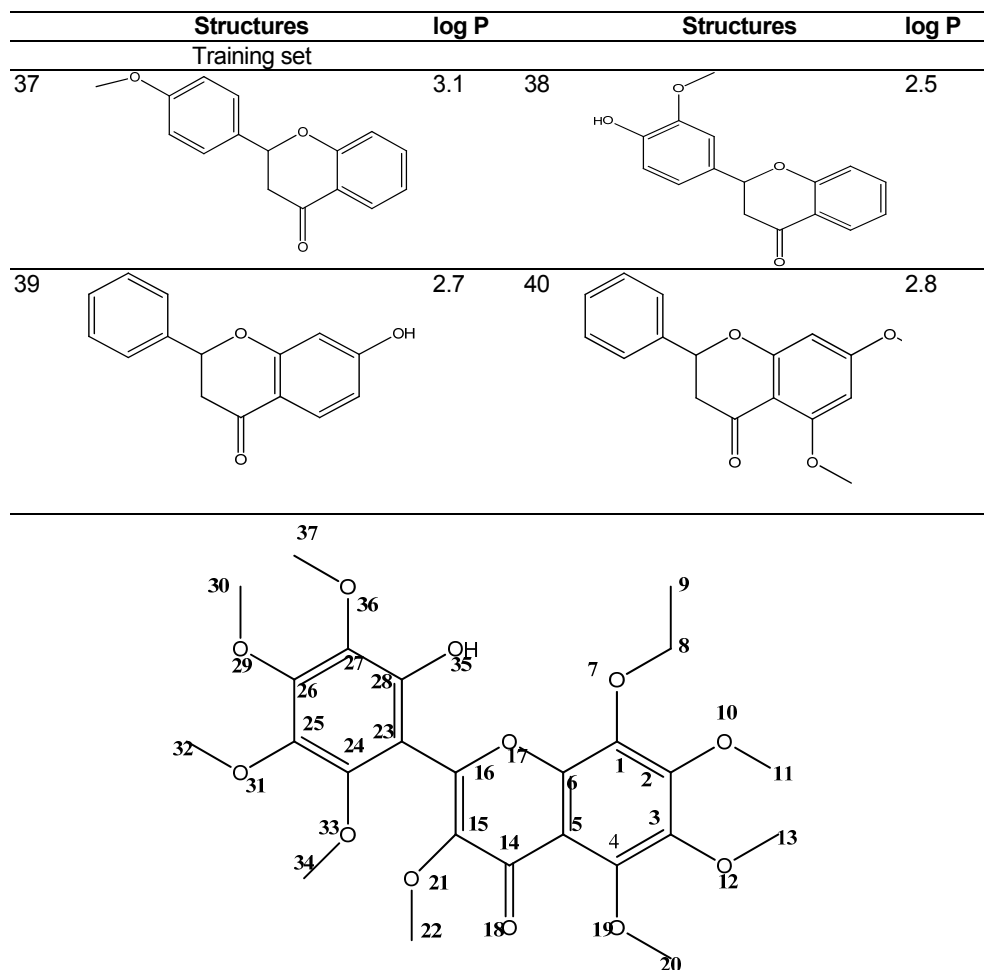


Figure 1. The hypermolecule comprising the common features of the dataset

3. METHOD

The structures have been optimized in Hyperchem, at PM3 level of theory. Topological indices implemented in TOPOCLUJ software [21] have been computed for all the structures. A selection of these indices is listed in Table 2.

3.1. Alignment over the hypermolecule

By aligning all the molecular structures over the hypermolecule, a binary vector (Table 3) was assigned to each molecule: 1-for a common feature in a given position of the hypermolecule and 0- for an empty position.

Next, the binary vector was weighted by the mass of “hydride” fragments composing each molecule. The weighted vector was used in the data-reduction step and correlation weighting procedure [22].

Table 2. Topological descriptors computed for the flavonoids in Table 1.

Structure	logP	SD	Detour	IE _{max} .
1	2.8	0.167	3197	325
2	3.5	0.462	1781	138.5
3	3.4	0.244	1431	107
4	1.6	-0.907	2690	249.5
5	2.3	-0.109	2424	214.5
6	3	0.204	3480	369.5
7	2.2	-0.326	2444	232.5
8	3.4	0.244	1800	158.5
9	3.1	0.323	3203	334
10	3.3	0.157	2225	218
11	2.9	0.170	3564	386
12	2.8	0.399	2422	204.5
13	1.7	-1.117	2200	192
14	3	0.204	4802	570.5
15	3.2	0.292	5546	672.5
16	3	0.204	4180	475.5
17	2.9	0.245	2936	294.5
18	3.1	0.327	2936	294.5
19	3	0.204	3556	388.5
20	2.7	0.072	3205	322.5
21	2.8	0.148	3170	309
22	2.6	0.041	3191	314.5
23	2.2	-0.419	2695	263
24	2.6	-0.457	1431	107
25	2.6	0.095	1986	164
26	3.8	0.846	2966	308.5
27	1.5	-1.396	2212	187.5
28	2.3	-0.279	1789	138
29	1.8	-0.977	1990	162
30	3.1	0.492	3522	367
31	2.6	-0.429	2420	222
32	1.5	-1.396	2225	194
33	1.7	-0.791	1779	135.5
34	3.2	0.069	1254	85
35	2.1	-0.429	1602	114
36	2.8	-0.281	1414	103.5
37	3.1	-0.018	1592	131
38	2.5	-0.544	1789	153
39	2.7	0.279	1594	114
40	2.8	0.267	1994	169.5

Table 3. The binary vectors, cf. hypermolecule, for the 40 flavonoids.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	0
2	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	1	1	0	0
3	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
4	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1
5	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	0
6	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1
8	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
9	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
11	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
12	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
13	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
14	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	0
18	1	1	1	1	1	1	0	0	0	1	1	1	0	1	1	1	1	1	1	0
19	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1
21	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	0
22	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	0
23	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
24	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
25	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	1	1	1	0
26	1	1	1	1	1	1	0	1	1	1	0	0	0	1	1	1	1	1	0	0
27	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
28	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
29	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
30	1	1	1	1	1	1	1	1	0	1	1	0	0	1	1	1	1	1	1	0
31	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	1	1	1	0	0
32	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
33	1	1	1	1	1	1	0	0	0	1	0	1	0	1	1	1	1	1	1	0
34	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
35	1	1	1	1	1	1	1	0	0	0	0	0	0	1	1	1	1	1	1	0
36	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
37	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
38	1	1	1	1	1	1	0	0	0	0	0	0	0	1	1	1	1	1	0	0
39	1	1	1	1	1	1	0	0	0	1	0	0	0	1	1	1	1	1	1	0
40	1	1	1	1	1	1	0	0	0	1	1	0	0	1	1	1	1	1	1	1

3.2. Data reduction and correlation weighting

In the step of data reduction, all the descriptors with the variance $\text{Var} < 30\%$ and those with intercorrelation larger than 0.80 have been discarded.

Correlation weighting was performed on all the positions in the hypermolecule: the correlating coefficients of the statistically significant positions of the hypermolecule were used to multiply the local descriptors, actually the mass fragments, thus resulting new weighted vectors CD_{ij} . Next,

the local correlating descriptors are summed to give a global descriptor, $SD_i = \sum_j CD_{ij}$. This new descriptor is a linear combination of the local correlating descriptors for the significant positions in the hypermolecule (i.e. H1, H3, H7, H8, H10, H11, H12, H19 – Table 4). It correlates with log P as below:

$$\log P = 2.783 + 0.999 \times SD$$

$$N=40; R^2=0.845; s=0.230; F = 206.616$$

The summative descriptor SD will be used as the basis for modeling log P.

Table 4. Correlation weighted descriptors (see text)

Structure	SD	H1	H3	H7	H8	H10	H11	H12	H19
1	0.167	-0.464	0.671	0	0	0.222	-0.119	-0.062	-0.080
2	0.462	-0.586	0.918	0	0	0.280	-0.151	0	0
3	0.244	-0.431	0.676	0	0	0	0	0	0
4	-0.907	1.449	-2.270	0	0	-0.693	0.372	0	0.234
5	-0.109	0.309	-0.447	0	0	-0.148	0.080	0.044	0.053
6	0.204	-0.661	1.036	0.149	-0.266	0.342	-0.184	-0.096	-0.116
7	-0.326	0.521	-0.816	0	0	-0.249	0.134	0	0.084
8	0.244	-0.431	0.676	0	0	0.000	0	0	0
9	0.323	-0.871	1.259	0	0	0.416	-0.224	-0.117	-0.141
10	0.157	-0.277	0.434	0	0	0	0	0	0
11	0.170	-0.571	0.894	0.129	-0.230	0.296	-0.159	-0.083	-0.106
12	0.399	-0.428	0.671	0	0	0.236	0	0	-0.080
13	-1.117	1.238	-1.939	0	0	-0.629	0	0	0.213
14	0.204	-0.661	1.036	0.149	-0.266	0.342	-0.184	-0.096	-0.116
15	0.293	-0.946	1.483	0.214	-0.381	0.490	-0.264	-0.138	-0.166
16	0.204	-0.661	1.036	0.149	-0.266	0.342	-0.184	-0.096	-0.116
17	0.245	-0.667	0.969	0	0	0.296	-0.159	-0.088	-0.106
18	0.327	-0.928	1.342	0	0	0.444	-0.239	-0.132	-0.160
19	0.204	-0.661	1.036	0.149	-0.266	0.342	-0.184	-0.096	-0.116
20	0.072	-0.233	0.365	0.053	-0.094	0.121	-0.065	-0.034	-0.041
21	0.148	-0.464	0.727	0.097	-0.172	0.222	-0.119	-0.062	-0.080
22	0.041	-0.143	0.224	0.032	-0.057	0.074	-0.040	-0.022	-0.027
23	-0.419	0.464	-0.727	0	0	-0.236	0	0	0.080
24	-0.457	0.806	-1.263	0	0	0	0	0	0
25	0.095	-0.155	0.242	0	0	0.074	-0.039	0	-0.027
26	0.846	-0.969	1.645	0	-0.364	0.534	0	0	0
27	-1.396	1.547	-2.423	0	0	-0.786	0	0	0.266
28	-0.279	0.309	-0.485	0	0	-0.157	0	0	0.053
29	-0.977	1.083	-1.696	0	0	-0.550	0	0	0.186
30	0.492	-0.856	1.454	0.193	-0.345	0.444	-0.239	0	-0.160
31	-0.429	0.806	-1.165	0	0	-0.385	0.207	0.108	0
32	-1.396	1.547	-2.423	0	0	-0.786	0	0	0.266
33	-0.791	1.238	-1.789	0	0	-0.629	0	0.176	0.213
34	0.069	-0.122	0.191	0	0	0	0	0	0
35	-0.429	0.571	-0.969	-0.137	0	0	0	0	0.106
36	-0.281	0.497	-0.778	0	0	0	0	0	0
37	-0.019	0.033	-0.051	0	0	0	0	0	0
38	-0.544	0.961	-1.505	0	0	0	0	0	0
39	0.279	-0.309	0.485	0	0	0.157	0	0	-0.053
40	0.267	-0.407	0.638	0	0	0.207	-0.105	0	-0.066

4. RESULTS AND DISCUSSION

4.1. QSAR models

The models were performed on the training set (the first 30 structures in Table 1) and the best results (in decreasing order of R^2) are listed below and in Table 5.

(i) Monovariate regression

$$\log P = 2.739 + 1.028 \times SD$$

(ii) Bivariate regression

$$\log P = 3.011 + 0.995 \times SD + 0.033 \times HOMO$$

(iii) Three-variate regression

$$\log P = 4.153 + 0.970 \times SD - 0.002 \times Detour + 0.013 \times IE_{\max}$$

(iv) Four-variate regression

$$\log P = 4.149 + 0.969 \times SD - 0.00001 \times D3D - 0.002 \times Detour + 0.013 \times IE_{\max}$$

Table 5. Best models in describing log P in the training set of flavonoids in Table 1

	Descriptors	R^2	Adjust. R^2	St. Error	F
1	SD	0.882	0.878	0.200	209.213
2	IE max	0.240	0.213	0.508	8.863
3	Detour	0.213	0.185	0.517	7.592
4	SD, HOMO	0.885	0.877	0.201	104.767
5	SD, IP max	0.882	0.874	0.203	101.818
6	SD, D3D	0.882	0.873	0.204	100.883
7	SD, Distance	0.882	0.873	0.204	100.930
8	SD, Detour	0.882	0.873	0.204	100.908
9	SD, IE max	0.882	0.873	0.203	101.567
10	SD, Detour, IE max	0.934	0.926	0.156	122.133
11	SD, Distance, IE max	0.906	0.895	0.185	83.746
12	SD, IE max, D3D	0.904	0.893	0.188	81.449
13	SD, C, HOMO	0.888	0.875	0.203	68.442
14	SD, IE max, IP max	0.887	0.873	0.204	67.791
15	SD, IP max, HOMO	0.886	0.873	0.205	67.317
16	SD, Distance, HOMO	0.886	0.873	0.204	67.417
17	SD, Detour, D3D	0.885	0.872	0.205	67.046
18	SD, IE max, HOMO	0.885	0.872	0.204	67.267
19	SD, D3D, Detour, IE max	0.933	0.923	0.159	88.080
20	SD, IE max, IP max, HOMO	0.892	0.874	0.203	51.509

4.2. Model Validation

An essential factor related to QSAR development is represented by the model validation. In this respect, Y-randomization and external validation Golbraikh-Tropsha procedure are required in order to confirm the statistical significance and predictive abilities of the obtained QSAR models [23].

(a) External Validation

The values $\log P$ for the test set of flavonoids were calculated by using equation cf. entry 10, Table 5. Data are listed in Table 6 and the monovariate correlation: $\log P = 0.125 + 0.976 \times \log P_{calc.}$; $n=10$; $R^2=0.887$; $s=0.217$; $F=62.525$ is plotted in Figure 2. One can see that Golbraikh-Tropsha criteria are fulfilled ($R^2_{pred}>0.8$) [23].

Table 6. Calculated values of $\log P$ for themolecules in the test set (Table 1)

Molecules	$\log P_{calc.}$	$\log P$
2	3.195	3.5
3	3.204	3.4
12	2.840	2.8
24	2.524	2.6
27	1.255	1.5
34	3.066	3.2
35	2.333	2.1
36	2.679	2.8
37	2.973	3.1
39	3.035	2.7

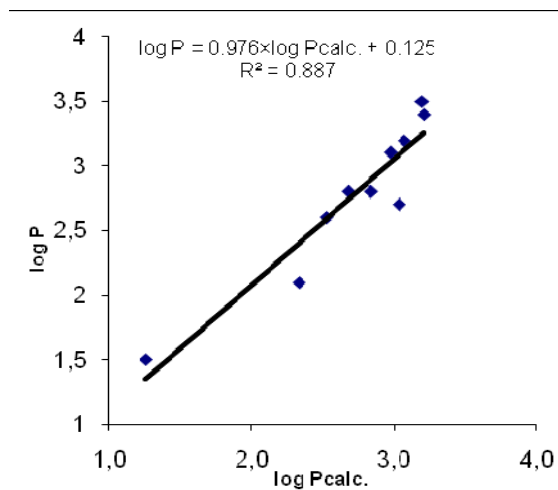


Figure 2. The plot $\log P$ vs. $\log P_{calc.}$ for the test set (external validation)

(b) Similarity Cluster Validation

Validation can also be performed by calculating $\log P$ for the molecules in the test set by using clusters of similarity: each of the 10 molecules is the leader of its own cluster, selected by 2D similarity among the 30 structures of the initial learning set. The values $\log P_{calc.}$ were computed by 10 new equations (the leader being left out) with the same descriptors as in eq. 10, Table5. Data are listed in Table 7 and the monovariate correlation:

$$\log P = -0.002 + 1.028 \times \log P_{calc.}; n=10; R^2=0.909; s=0.194; F= 80.033$$

is plotted in Figure 3.

One can see that the prediction of log P by the similarity clusters is better than that obtained in the external validation. This is because the similarity procedure provides a set of quasi-congeners, thus making possible the basic paradigm of QSAR: similar structures show similar properties.

Table 7. Calculated values of log P by similarity clusters, for themolecules in the test set

Molecules	log P calc.	log P
2	3.132	3.5
3	3.169	3.4
12	2.772	2.8
24	2.585	2.6
27	1.305	1.5
34	3.057	3.2
35	2.268	2.1
36	2.693	2.8
37	3.023	3.1
39	2.955	2.7

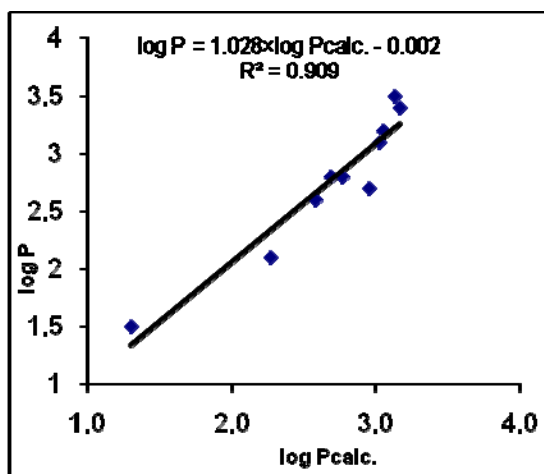


Figure 3. The plot log P vs. log P calc. by similarity clusters

CONCLUSIONS

A novel QSAR approach, based on correlation waighting within the hypermolecule, considered to mimic the investigated correlational space, was performed on a set of 40 flavonoids, downloaded from the PubChem database. The set was split into a learning set and a test set, the last one being used for the validation of the models, in the so-called „external set validation”. Also, the validation was made by a new version of prediction by using similarity clusters. The similarity clustering permitted realization of „quasi-congeneric” set of structures, thus providing a better prediction than the classical external validation procedure.

ACKNOWLEDGEMENTS

I am thankful fo Professor Mircea V. Diudea for helpful discussion.

REFERENCES

1. W.E. Dismukes, *Clinical Infectious Diseases*, **2000**, 30, 653.
2. T.A. Özlem, *Turk J Med Sci*, **2001**, 31, 493.
3. D. Rogers, and A.J. Hopfinger, *J. Chem. Inf. Comput. Sci.*, **1994**, 34, 854.
4. R.P. Sheridan, S.K. Kearsley, *Drug Discov. Today*, **2002**, 7, 903.
5. H.J. Bohm, G. Schneider, *Wiley-VCH*, Weinheim, **2000**.
6. G. Schneider, H.J. Bohm, *Drug Discov. Today* **2002**, 7, 64.
7. J.A. DiMasi, R.W. Hansen, H.G. Grabowski, *J. Health Econ.*, **2003**, 22, 151.
8. P.M. Dean, *Blackie Academic*, London, **1995**.
9. P. Willet, *Research Studies Press Ltd*: Letchworth, U.K., **1987**.
10. J.M. Barnard, G.M. Downs, *J. Chem. Inf. Comput. Sci.*, **1992**, 32, 644.
11. D.J. Wild, C.J. Blankey, *J. Chem. Inf. Comput. Sci.*, **2000**, 40(1), 155.
12. R.D. Brown, Y.C. Martin, *J. Chem. Inf. Comput. Sci.*, **1996**, 36, 572.
13. M.A. Johnson, G.M. Maggiora, Eds., John Wiley & Sons, New York, **1990**.
14. C.D. Moldovan, A. Costescu, G. Katona and M.V. Diudea, *MATCH Commun. Math. Comput. Chem.*, **2008**, 60, 977.
15. P. Willett, *J. Chem. Inf. Comput. Sci.*, **1998**, 38, 983.
16. M. Randić, *J. Chem. Inf. Comput. Sci.* **1995**, 35, 373.
17. M.V. Diudea, *MATCH Commun. Math. Comput. Chem.* **1997**, 35, 169.
18. M.V. Diudea, *J. Chem. Inf. Comput. Sci.*, 1997, 37, 300.
19. A.A. Toropov, and A.P. Toropova, *Internet El. J. Molec. Design*, **2002**, 1, 108.
20. A.T. Balaban, A. Chiriac, I. Motoc, and Z. Simon, *Steric Fit in QSAR (Lectures Notes in Chemistry*, Vol. 15), Springer, Berlin, **1980**, Chap. 6.21.
21. O. Ursu, M.V. Diudea, "TOPOCLUJ software program", Babes-Bolyai University, Cluj, **2005**.
22. A. A. Toropov, A. P. Toropova, *J. Mol. Struct. (Theochem)* **2001**, 538, 287.
23. A. Golbraikh, A. Tropsha, *J Comp Aided Mol Des*, **2002**, 16, 357.