J Math Chem (2014) 52:948–965
DOI 10.1007/s10910-013-0302-0

ORIGINAL PAPER

The effect of the diversity of molecules in sets and similarity of sets on the quality of prediction in QSAR studies

Laszlo Tarko

Received: 16 July 2013 / Accepted: 12 December 2013 / Published online: 31 January 2014

Abstract We report here: (a) formulas/procedures for calculating the similarity of molecules, considering their chemical structure, size, shape and hydrophilicity (b) a procedure for clusterization of the sets of molecules, according to similarity (c) formulas/procedures for calculating the diversity of molecules in clusterized sets as well as similarity of clusterized sets, based on Shannon Entropy formalism The paper analyses the influence of the diversity of molecules and similarity of calibration/prediction sets on the quality of prediction for prediction set molecules. The calculated influence of certain molecular feature (chemical structure, size, shape and hydrophilicity) on toxicity depends on the structure of the database, specifically the number of molecules and diversity of molecules having analyzed molecular feature. A QSAR analysis of 49 phenol derivatives revealed the effect of the diversity of molecules in sets and of the similarity of sets on the quality of prediction for prediction set molecules: (a) a direct correlation with the similarity of sets, regardless of analyzed molecular feature (b) an inverse correlation with the diversity of molecules in the calibration set, from the point of view of chemical structure, size and shape (c) a direct correlation with the diversity of molecules in calibration set, from the point of view of hydrophilicity (d) a direct correlation with the diversity of molecules in prediction set, regardless of analyzed feature.

Keywords Molecular diversity · Molecular similarity · Shannon Entropy · QSAR

L. Tarko (✉)
Centre of Organic Chemistry, Romanian Academy, Sector 6, Spl. Independentei 202B,
PO Box 35-108, 060023 Bucharest, Romania
e-mail: ltarko@cco.ro; tarko_laszlo@yahoo.com

1 Introduction

In QSPR (*Quantitative Structure-Property Relationship*) studies the *dependent property* P can be *biochemical activity* (QSPR ≡ QSAR), *toxicity* (QSPR ≡ QSTR), *chromatographic retention time* (QSPR ≡ QSRR), *viscosity* (QSPR ≡ QSVR), *molecular aromaticity* (QSPR ≡ QSArR) etc. About 90 % of QSPR studies quoted in literature are QSAR studies.

In QSPR studies one uses two groups of molecules. Here, these groups are named *calibration set* and *prediction set*.

In QSAR studies the calibration set includes molecules having known (observed) value of biochemical activity. Applying specific methodology, these molecules are used to identify *the best* QSAR, i.e. the mathematical formula which gives the minimum difference between the observed and computed values of activity. The QSARs include few descriptors, i.e. computable features of molecules. The descriptors in the best QSAR are named *predictors*.

The QSAR is used to compute the value of activity for prediction set molecules, which are not used in QSAR computation.

In *academic* QSAR studies the prediction set includes molecules having known value of activity. Consequently, the computed values of activity can be compared with observed values. In this case the comparison is an *external validation test*, the prediction set is named *validation set* and the agreement between observed and computed values is a measure of the quality of QSAR. A certain external validation test uses, as validation set, molecules extracted from the initial calibration set. Any QSAR obtained using *all* available molecules having known value of activity (*all* included in calibration set) cannot be validated by external validation, because the validation set is non-existent. Therefore, the computation using *all* molecules and the computation in the validation test use different calibration sets and the obtained QSARs are different. Hence, any external validation test says nothing regarding the predictive power for molecules in the initial prediction set of QSAR obtained using all molecules in the initial calibration set. Consequently, the utility of external validation tests is a debatable subject in the literature [1–8]. Some authors presented *rational* procedures for the selection of calibration and validation sets [9].

In *practical* QSAR studies, the prediction set includes new, not yet synthesized molecules, as imaginary result of (rational) drug design. The real activity for molecules in the prediction set is unknown and the agreement between observed and computed values cannot be computed. As a rule, this (unknown) agreement is good only if the (known) best QSAR for the calibration set and the (unknown) best QSAR for the prediction set are *similar*. There are no studies in literature concerning the similarity of QSARs and it is not subject of this paper. In principle, this similarity should be the effect of the similarity of the calibration set (as a whole) and the prediction set (as a whole) from the points of view of various features F (chemical structure, molecular size, molecular shape, lipophilicity etc.). This similarity should emphasize the similarity of the structure-activity relationship in the calibration set and the prediction set, respectively.

Frequently, the observed values of activity are within a large enough range, more than two logarithmic units, and the diversity of these values is high. Consequently,

another problem is the diversity of the molecules in the calibration and prediction sets, from the points of view of F. If the diversity of molecules in the calibration set, from the point of view of F, is too low, the QSAR methodology cannot identify the descriptors regarding F as *significant* and the best QSAR cannot include, as predictors, the descriptors regarding F. On the contrary, if the diversity is too high these sets can be non-homogeneous, i.e. can include two or more classes, different from the point of view of structure-activity relationship.

The chemical similarity of molecules is often described as an inverse of Euclidean/ non-Euclidean distance in a certain descriptor space [10–22]. Our paper presents some formulas/procedures for the calculation of the similarity of molecules from the point of view of chemical structure, size, shape and hydrophilicity, a procedure for the clusterization of the sets of molecules, according to their similarity and some formulas/procedures for the calculation of the diversity of molecules in clusterized sets and the similarity of clusterized sets.

To increase the chance for a good prediction of activity for the prediction set molecules in *practical* QSAR studies, one should uses certain calibration and prediction sets, suitable from the point of view of similarity and diversity. In theory, the similarity of calibration and prediction sets (as a whole) should be as high as possible and the diversity of molecules, from the point of view of F, should be either high or low, depending on the nature of F. The goal of paper is to verify this idea.

2 Methods and formulas

Sometimes the databases (calibration set + prediction set) used in the vast QSPR/QSAR field include quite similar molecules from the point of view of the chemical structure and non-similar from the point of view of size and shape [23,24]. Other databases include very non-similar molecules from all points of view [25,26].

2.1 The used database

To verify the effect of similarity and diversity on the quality of prediction for prediction set molecules, we analyzed, to give an example, a small number of phenol derivatives in Table 1, ordered by the value of activity. The *activity* is the toxicity against *Tetrahymena pyriformis* protozoan. The values of toxicity are quoted in literature [27,28]. Here, the values of toxicity are weighted within the [0, 2.638] range, in order to use in computations only positive values of the dependent property.

2.2 Geometry optimization and computation of descriptors

The virtual building of the molecules in Table 1 and the geometry optimization were done using the molecular-mechanics program PCModel [29]. The more rigorous geometry optimization was subsequently performed by quantum-mechanics program MOPAC [30] using the keywords *pm6 pulay gnorm = 0.2 geo-ok bonds vectors*. These keywords fixed the parameters of geometry optimization.

Table 1 The chemical structure and the toxicity of analyzed phenols

No.	Substituent(s)	Toxicity
1	None	0.000
2	4-Methyl	0.239
3	3-Methyl	0.369
4	2,5-Dimethyl	0.440
5	2-Fluoro	0.448
6	3,5-Dimethyl	0.544
7	2,3-Dimethyl	0.553
8	3,4-Dimethyl	0.553
9	2,4-Dimethyl	0.559
10	2-Ethyl	0.607
11	3-Ethyl	0.660
12	3-Fluoro	0.679
13	2-Chloro	0.708
14	2-Fluoro	0.827
15	2,3,6-Trimethyl	0.849
16	3-Fluoro	0.904
17	4-*iso*-propyl	0.904
18	2-Bromo	0.935
19	4-Chloro	0.976
20	3-*iso*-propyl	1.040
21	2-Chloro-5-methyl	1.071
22	4-Bromo	1.112
23	2-Methyl-4-chloro	1.131
24	3-*tert*-butyl	1.161
25	3-Methyl-4-chloro	1.226
26	2-*iso*-propyl	1.234
27	3-chloro-4-fluoro	1.273
28	4-iodo	1.285
29	4-*tert*-butyl	1.344
30	3,4,5-trimethyl	1.361
31	2,4-dichloro	1.467
32	2-phenyl	1.525
33	3-iodo	1.549
34	2.5-dichloro	1.559
35	3,5-dimethyl-4-chloro	1.634
36	2,4-dimethyl-6-*tert*-butyl	1.676
37	2,3-dichloro	1.702
38	2-methyl-4-bromo-6-chloro	1.708
39	2,6-dimethyl-4-bromo	1.709

Table 1 continued

No.	Substituent(s)	Toxicity
40	2-*tert*-butyl-4-methyl	1.728
41	2,4-dibromo	1.834
42	3,5-dichloro	1.993
43	2,6-dichloro-4-bromo	2.210
44	2,6-di-*tert*-butyl-4-methyl	2.219
45	2-*iso*-propyl-4-chloro-5-methyl	2.293
46	2,4,6-tribromo	2.481
47	2,4,5-trichloro	2.531
48	2,6-diphenyl	2.544
49	2,4-dibromo-6-phenyl	2.638

Based on the output files created by MOPAC, the PRECLAV software [7,31] calculated, for each molecule, more than 1,000 *whole molecule* and 3D descriptors, specific to this program. The same software was used in the identification of molecular fragments, similarity/diversity and statistical computations.

2.3 Identification of the *significant* molecular fragments

These calculations use the result of the identification of molecular fragments, according to a previously described algorithm [32]. Actually, two bonded (by a chemical bond having B bond order value) heavy atoms (different from hydrogen) are included within the same fragment if $B > 1.051$ (an empirical limit value specific to computation by quantum-mechanics method PM6). For instance, if the conjugation of the OH group with the cycle, in phenol derivatives, is strong enough (high enough value of B), the OH fragment is missing and the C_6H_xO fragment is present. Using this idea the program cuts off the fragments in the analyzed molecule.

After the identification of the fragments the program computes, for each molecule, some descriptors of fragments and the percentage in weight of the fragments. For each fragment, the sign of the Pearson linear correlation r, within $[-1, +1]$ range, between the value of percentages and the value of activities, can be positive or negative and the square of correlation r^2, within [0, 1] range, can be high or low. A positive sign of correlation means *high mass percent of this fragment increases the activity value*. Consequently, a negative sign of correlation means *high mass percent of this fragment decreases the activity value*. The *significant* molecular fragments fulfill condition (1).

$$r^2 > \text{Ln (N)} / \text{N} \tag{1}$$

where N is the number of molecules in the calibration set

2.4 The algorithm for chemical similarity

The chemical similarity calculation also uses the result of the identification of molecular fragments.

All identified fragments are classified according to following criteria #1 and #2.

If the number of heavy atoms included is 1 the value of criterion #1 is 1. If the number of heavy atoms included is 2 or 3 the value of criterion #1 is 2. If the number of heavy atoms included is >3 the value of criterion #1 is 3.

Criterion #2 is the string of symbols of included elements, in alphabetical order.

If the value of criterion #1 *and* criterion #2 is the same the analyzed fragments are considered *in the same class*.

Exceptionally, fragment C is considered *in the same class* with fragments CH, CH_2 and CH_3. Fragments F, Cl, Br, I are identified as *different* fragments and are included in different classes. Also, fragments B, N, O, S, P, As, Si, Se and Te are identified as *different* fragments and are included in different classes. Fragments NO and SO are *different*. Fragments NH, OH and SH are *different* also. The fragments in the pairs NO/NO_2, SO/SO_2 and NH/NH_2 are included in the same class. According to this algorithm the fragments NCO (disubstituted amide), OCN (cyanate) and NCO (*iso*-cyanate) are included into the same class also. The fragments NHCO (monosubstituted amide) and N_2H_nCO (substituted urea) are different because of the different values of criterion #1. All aromatic fragments C_nH_m are included into the same class (however, if $m = 0$ the fragment is considered *different*).

We computed, within the [0, 1] range, the ratios (in weight) p_i of each class of fragments and we used these ratios in the calculation of Shannon Entropy SE [33] of the analyzed molecule.

$$\mathrm{SE} = -\sum_{i=1}^{k} p_i \cdot \mathrm{Log}\ (p_i) \tag{2}$$

where k is number of class of fragments

If the molecule includes just one fragment (for instance PAHs and azines) or just one class of fragments (for instance alkanes) the value of SE is zero because $k = 1$ and $p_1 = 1$.

The Chemical Structure similarity $\mathrm{SIM_{CS}}$ of two molecules is

$$\mathrm{SIM_{CS}} = (\mathrm{SE_1} + 1)/(\mathrm{SE_{12}} + 1) \cdot (\mathrm{SE_2} + 1)/(\mathrm{SE_{12}} + 1) \cdot \mathrm{k_1}/\mathrm{k_{12}} \cdot \mathrm{k_2}/\mathrm{k_{12}} \tag{3}$$

where $\mathrm{SE_1}$—the Shannon Entropy of the molecule #1; $\mathrm{SE_2}$—the Shannon Entropy of the molecule #2; $\mathrm{SE_{12}}$—the Shannon Entropy for the aggregate #1 + #2; $\mathrm{k_1}$—the number of classes in molecule #1; $\mathrm{k_2}$—the number of classes in molecule #2; $\mathrm{k_{12}}$—the number of classes in aggregate #1 + #2.

If the ratios in formula (3) are > 1 the program uses the inverse of these ratios. Consequently, the value of $\mathrm{SIM_{CS}}$ is within the [0, 1] range.

Two molecules are very similar from the point of view of chemical structure if they include the same classes of molecular fragments. If the value of $\mathrm{SIM_{CS}}$ is high enough the molecules can be included into the same *chemical cluster*.

Ideas presented here regarding the calculation of chemical similarity are an updated version of a previously described algorithm [34].

2.5 Formulas for similarity of shapes

To compute the similarity SIM_{SH} of the shapes for two molecules the program uses the value of descriptors GSI and PAX.

The general shape, 1D, 2D or 3D, of the circumscribed ellipsoid of the analyzed molecule is the value of function GSI, within the range [1,3].

$$GSI = (md_1 + md_2)/md_3/k + k \tag{4}$$

The descriptors md_1, md_2 and md_3, $md_1 \leq md_2 \leq md_3$, are the molecular dimensions computed by MOPAC software [30]. If the ratio md_1/md_3 is small enough ($\leq$0.25) then the factor $k = 1$, else $k = 2$.

If GSI $\sim$ 1 the shape of circumscribed ellipsoid is very elongated (dicyane, triacetylene etc.). If 1.8 $\leq$ GSI $\leq$ 2.2 the shape is somehow planar (benzene, 1,3,5-trinitro-benzene, pyrene etc.). If GSI $\geq$ 2.7 the shape of the circumscribed ellipsoid is almost spherical (methane, cubane, adamantane, fullerenes etc.).

PAX is the variation coefficient of distances to geometric center, computed for peripheral atoms. A certain atom is considered *peripheral* if it is bonded with maximum two other atoms. For instance, dimethyl-ether includes 7 *peripheral* atoms, i.e. 6 hydrogen atoms and 1 oxygen atom.

$$SIM_{SH} = \text{minimum } (r_1,\ r_2) \tag{5}$$

where $r_1 = GSI_1/GSI_2$ $r_2 = PAX_1/PAX_2$

If $r_1 > 1$ and/or $r_2 > 1$ the program uses the inverse of these ratios and the value of SIM_{SH} is within the range [0, 1].

Two molecules are similar from the point of view of molecular shape if the general shape indices GSI *and* the unevenness of molecular surfaces PAX are similar. If the value of SIM_{SH} is high enough the molecules can be inserted into the same *shape cluster*.

2.6 Formula for the similarity of size

To compute the similarity SIM_{SZ} of size for two molecules the program uses the value of descriptor *COSMO volume* CVO, computed by MOPAC software [30].

$$SIM_{SZ} = CVO_1/CVO_2 \tag{6}$$

If $SIM_{SZ} > 1$ the program uses the inverse of this ratio and the value of SIM_{SZ} is within the range [0, 1]. If the value of SIM_{SZ} is high enough the molecules can be inserted into the same *size cluster*.

2.7 Formulas for the similarity of hydrophilicity

To compute the similarity SIM_{HP} of hydrophilicity for two molecules the program uses the value of descriptors AHY (average hydrophilicity of molecular fragments) and XHY (maximum hydrophilicity of molecular fragments).

The hydrophilicity of a certain molecular fragment is the difference Δ between the maximum value S_{max} of the net charges of hydrogen atoms and the minimum value S_{min} of the net charges of heteroatoms.

$$\Delta = S_{max} - S_{min} \tag{7}$$

If the hydrogen atoms in the fragment are missing $S_{max} = 0$. If the heteroatoms in the fragment are missing $S_{min} = 0$. Therefore, $\Delta = 0$ for fragments which includes only carbon atoms (C in carbon tetrachloride, C_2 in tetrachloroethylene, C_6 in total substituted benzene etc.).

$$SIM_{HP} = \text{ minimum } (r_1, r_2) \tag{8}$$

where $r_1 = AHY_1/AHY_2\, r_2 = XHY_1/XHY_2$

If $r_1 > 1$ and/or $r_2 > 1$ the program uses the inverse of these ratios and the value of SIM_{HP} is within the range [0, 1].

Two molecules are similar from the point of view of hydrophilicity if the values of the AHY *and* XHY descriptors are similar. Dodecane and dodecanol present close value of the AHY descriptor but very different value of the XHY descriptor. Methanol and dodecanol present different value of the AHY descriptor but very close value of the XHY descriptor.

If the value of SIM_{HP} is high enough the molecules can be inserted into the same *hydrophilicity cluster*.

2.8 Algorithm for clusterization

To include a certain molecule into a certain cluster the algorithm uses criteria K#1 and K#2.

$$K\#1 = SIM_{max}/\ SIM_{min} \tag{9}$$

where SIM_{max} is maximum similarity with not yet included (*clusterized*) molecules; SIM_{min} is minimum similarity with not yet *clusterized* molecules

$$K\#2 = (SIM_{max} - k)\ /SIM_{min} \tag{10}$$

where

SIM_{max} is maximum similarity with molecules included in analyzed cluster

k is empirical limit value for similarity which depends on similarity criterion (chemical structure, molecular size, molecular shape, hydrophilicity)

SIM_{min} is minimum similarity with not yet included molecules

The first molecule, included into the first cluster, is the molecule having the maximum value of ratio K#1. This molecule is *the seed* of the first cluster.

Then, the algorithm computes, for each not yet clusterized molecule and for each existent cluster, the value of K#2. The maximum value of K#2 indicates which molecule will be clusterized and which cluster will include this molecule.

If the maximum value of K#2 is negative, because $SIM_{max} < k$, the algorithm computes, for all not yet clusterized molecules, the value of K#1. Therefore, the algorithm identifies a *seed* for a new cluster.

The program uses clusterization procedure for the calibration set, the prediction set and the entire database (calibration set + prediction set aggregate).

2.9 The diversity of molecules in sets and the similarity of sets

After the clusterization of the molecules in the calibration and prediction set, the program computes, using formula (2), the diversity of molecules in the calibration and prediction set.

In these computations *the classes* in formula (2) are the identified clusters. Always, $n > k$, where n is the total number of clusterized molecules, and the maximum value of entropy is Log (n) [34]. Actually, *the diversity* is the Shannon Entropy SE weighted by Log (n). Consequently, the value of diversity is within the range [0, 1].

Then the program computes, using formula (3), the similarity of calibration and prediction sets.

2.10 Statistical calculations

The molecules in Table 1 were divided into four groups. The group G#1 includes molecules 2, 4, 6, …, 38, 40, 42, 43, 45, 47 and 49. The group G#2 includes the molecules in Table 1 which are not included in G#1. The group G#3 includes the molecules 1–25. The group G#4 includes the molecules in Table 1 which are not included in G#3.

In QSAR study #1 the calibration set is the group G#1 and the prediction set is the group G#2. In QSAR study #2 the calibration set is the group G#2 and the prediction set is the group G#1. In QSAR study #3 the calibration set is the group G#3 and the prediction set is the group G#4. In QSAR study #4 the calibration set is the group G#4 and the prediction set is the group G#3.

In QSAR studies #1 and #2 the average toxicity in the calibration set is close to average toxicity in the prediction set (1.22 vs. 1.31). Therefore, the prediction for the prediction set molecules is, as a rule, the effect of interpolation.

On the contrary, in QSAR studies #3 and #4 the average toxicity in the calibration set and the average toxicity in the prediction set are very different (1.81 vs. 0.74). Therefore, the prediction for the prediction set molecules is, as a rule, the effect of extrapolation.

In QSAR study #5 the calibration set includes all molecules in Table 1 and the prediction set is missing. This QSAR study allows an useful comparison with the results in QSAR studies #1–#4, from the point of view of structure-toxicity relationship.

The program PRECLAV computes type (11) multilinear QSARs.

$$T = C_0 + \sum_{i=1}^{k} C_i \cdot D_i \tag{11}$$

where T is (the value of) toxicity; C_0 is the free term (intercept); C_i are coefficients (weighting factors); D_i are (the value of) *significant* descriptors; k is the number of descriptors in the analyzed set.

The algorithm of QSAR computation, specific to PRECLAV software, statistical formulas included, was previously described [7], including the identification and step-by-step elimination of outliers.

The square of Pearson linear correlation r^2 of observed/computed values, the Fisher function F, the standard error of estimation *SEE*, the relative standard error of estimation *RSEE* and the quality function Q are criteria for the quality of prediction for the calibration/prediction set molecules.

$$\mathrm{RSEE} = 1 - \mathrm{SEE} / \mathrm{A} \tag{12}$$

where A is the average of the observed values of toxicity

$$\mathrm{Q} = \mathrm{r}^2 \cdot (1 - p/\mathrm{N}) \tag{13}$$

where p is number of predictors; N is number of molecules in the calibration set

3 Commented results

Table 2 includes the identified clusters in database.

Some molecules can be considered *atypical* because they are included in clusters having just one molecule. These molecules are non-similar with all other molecules. The molecules 5, 19, 27, 38, 39, 43, 48 and 49 are atypical from the point of view of chemical structure. The molecules 2, 3, 19, 26, 32 and 42 are atypical from the point of view of shape. The molecules 1 and 49 are atypical from the point of view of size. The molecule 44 is atypical from the point of view of hydrophilicity.

The diversity of molecules in sets and the similarity of sets are presented in Tables 3 and 4.

The similarity of G#1 and G#2 groups is higher than similarity of G#3 and G#4 groups, in average 0.584 ± 0.174 versus 0.498 ± 0.126. The diversity of molecules in G#1 and G#2 groups and the diversity of molecules in G#3 and G#4 groups, are somehow similar, in average 0.602 ± 0.084 versus 0.579 ± 0.114.

According to PRECLAV algorithm, there are no outliers in the calibration sets used in QSAR studies #1–#5.

3.1 QSAR study #1

Calibration set: G#1 Prediction set: G#2
The *significant* molecular fragments in the calibration set:

C_6H_3O	$r = 0.6144$
Cl	$r = 0.5984$
OH	$r = -0.4524$
C_6H_3	$r = -0.3686$

Table 2 Clusters in database

According to	Index of cluster	Molecules in Table 1
Chemical structure	1	12, 14, 16
	2	28, 33
	3	18, 22, 41, 46
	4	13, 31, 34, 37, 42, 47
	5	2, 3, 4, 6, 7, 8, 9, 10, 11, 15, 17, 20, 24, 26, 29, 30, 36, 40, 44
	6	21, 25, 35
	7	23, 45
	8	1, 32
	9	48
	10	39
	11	27
	12	5
	13	43
	14	19
	15	49
	16	38
Molecular shape	1	6, 7, 9, 25, 30, 38, 44
	2	27, 41, 43, 47
	3	8, 10, 21, 36, 40, 45
	4	11, 17, 20, 24, 29, 48, 49
	5	1, 5, 12, 16, 18, 31, 34, 46
	6	13, 14, 37
	7	4, 23, 35
	8	22, 28, 33
	9	15, 39
	10	26
	11	32
	12	42
	13	2
	14	3
	15	19
Molecular size	1	2, 3, 5, 12, 13, 14, 16
	2	18, 19, 22, 27
	3	4, 6, 7, 8, 9, 10, 11, 21, 23, 25, 28, 31, 33, 34, 37, 42
	4	24, 29, 32, 40, 46
	5	15, 17, 20, 26, 35, 36, 38, 39, 41, 43, 47
	6	44, 48
	7	36, 45
	8	1

Table 2 continued

According to	Index of cluster	Molecules in Table 1
	9	49
Molecular hydrophilicity	1	24, 29, 36, 40, 43, 45, 46, 47
	2	12, 13, 18, 22, 28, 33
	3	2, 3, 5, 19, 25, 32, 48
	4	15, 17, 20, 26, 30, 31, 34, 37, 38, 39, 41, 49
	5	1, 16
	6	4, 6, 7, 8, 9, 10, 11, 14, 21, 23, 35, 27, 42
	7	44

Table 3 Similarity and diversity of G#1 and G#2

According to	SIM	$DIV_{G\#1}$	$DIV_{G\#2}$
Chemical structure	0.4110	0.5731	0.6773
Molecular shape	0.5344	0.7147	0.7092
Molecular size	0.5661	0.5528	0.5375
Molecular hydrophilicity	0.8224	0.5131	0.5386

Table 4 Similarity and diversity of G#3 and G#4

According to	SIM	$DIV_{G\#3}$	$DIV_{G\#4}$
Chemical structure	0.3683	0.5102	0.7172
Molecular shape	0.5336	0.6868	0.7363
Molecular size	0.4317	0.4627	0.5319
Molecular hydrophilicity	0.6570	0.4868	0.5010

The best type (11) QSAR:

$C_0 = -5.4002$
$C_1 = 4.3714$
D_1 is Flexibility index no. 1 [35]
$C_2 = 0.3811$
D_2 is Molecular volume/number of atoms ratio

Quality of prediction for the calibration set:

$$r^2 = .9113 \ F = 118.2 \ SEE = .1989 \ Q = .8384$$

Quality of prediction for the prediction set: $r^2 = .9104$ RSEE = 0.805

3.2 QSAR study #2

Calibration set: G#2 Prediction set: G#1

The *significant* molecular fragments in the calibration set:

OH	r = −0.6891
Br	r = 0.4824
C_6H_4	r = −0.4135

The best type (11) QSAR:

C_0 =−1.6727
C_1 = 1.9989
D_1 is Harary topological index/heavy atoms number
C_2 = −0.0259
D_2 is Percentage of carbon

Quality of prediction for the calibration set:

$$r^2 = .9650 \ F = 303.2 \ SEE = .1242 \ Q = .8846$$

Quality of prediction for the prediction set: $r^2 = .8213$ RSEE = 0.786

3.3 QSAR study #3

Calibration set: G#3 Prediction set: G#4
The *significant* molecular fragments in the calibration set:

OH	r = −0.5792
C_6H_5	r = −0.4923
Cl	r = 0.4476

The best type (11) QSAR:

C_0 = −13.7314
C_1 = 0.1521
D_1 is LUMO–HOMO gap weighted molecular volume
C_2 = 12.7054
D_2 is Maximum bond order in C-A or A-A bonds (A is any heteroatom)
C_3 = 442.8112
D_3 is Resultant electrostatic force on probe atom no. 64 (a 3D descriptor) [36]

Quality of prediction for the calibration set:

$$r^2 = .9327 \ F = 101.7 \ SEE = .0812 \ Q = .8208$$

Quality of prediction for the prediction set: $r^2 = .7387$ RSEE = 0.748

3.4 QSAR study #4

Calibration set: G#4 Prediction set: G#3
The *significant* molecular fragments in the calibration set:

Br	r = 0.4086
C_6H_3O	r = 0.3988
C_6H_4	r = −0.3948

The best type (11) QSAR:
C_0 = −11.3063
C_1 = 67.1448
D_1 is Resultant electrostatic force on probe atom no. 65 (a 3D descriptor) [36]
C_2 = 78.6486
D_2 is 1/[E(lumo + 1) − E(homo − 1)] ratio
C_3 = 0.5708
D_3 is Molecular lipophilicity #1 [36]
C_4 = 16.2655
D_4 is Maximum free valence of C atoms

Quality of prediction for the calibration set:

$$r^2 = .9196 \ F = 57.2 \ SEE = .1257 \ Q = .7663$$

Quality of prediction for the prediction set: $r^2 = .3434$ RSEE = 0.528

3.5 QSAR study #5

Calibration set: all molecules in Table 1 Prediction set: missing
The *significant* molecular fragments in the calibration set:

OH	r = −0.5731
C_6H_3O	r = 0.4728
Br	r = 0.3931
C_6H_4	r = −0.3788
Cl	r = 0.3418
C_6H_3	r = −0.3241

The best type (11) QSAR:

C_0 = −1.9823
C_1 = 0.0214
D_1 is COSMO area
C_2 = 38.8059
D_2 is Resultant electrostatic force on probe atom no. 62 (a 3D descriptor) [36]
C_3 = 11.1615
D_3 is Resultant electrostatic force on probe atom no. 78 (a 3D descriptor) [36]

Table 5 Observed and computed values of toxicity

No. in Table 1	T_{obs}	T_{calc} by QSAR				
		#1	#2	#3	#4	#5
1	0.000	**−0.159**	0.030	0.044	**0.136**	0.015
2	0.239	0.409	**0.317**	0.279	**0.334**	0.292
3	0.369	**0.367**	0.330	0.360	**0.827**	0.397
4	0.440	0.664	**0.626**	0.565	**0.917**	0.650
5	0.448	**0.453**	0.666	0.589	**0.509**	0.600
6	0.544	0.689	**0.619**	0.555	**1.309**	0.783
7	0.553	**0.575**	0.656	0.512	**1.107**	0.565
8	0.553	0.668	**0.626**	0.537	**0.903**	0.624
9	0.559	**0.654**	0.626	0.430	**0.643**	0.508
10	0.607	0.740	**0.571**	0.695	**0.958**	0.576
11	0.660	**0.844**	0.541	0.668	**1.144**	0.738
12	0.679	0.552	**0.700**	0.621	**0.175**	0.618
13	0.708	**0.564**	0.913	0.653	**0.653**	0.760
14	0.827	1.046	**1.259**	0.970	**0.723**	0.893
15	0.849	**0.859**	0.940	0.734	**1.294**	1.030
16	0.904	0.628	**0.679**	0.836	**0.771**	0.800
17	0.904	**0.904**	0.787	0.983	**1.049**	1.013
18	0.935	0.946	**1.286**	0.956	**0.756**	1.003
19	0.976	**0.679**	0.880	0.819	**0.857**	0.902
20	1.040	1.167	**0.810**	1.023	**1.366**	1.062
21	1.071	**1.073**	1.137	1.069	**1.119**	1.147
22	1.112	1.053	**1.253**	1.121	**0.703**	1.092
23	1.131	**1.064**	1.137	1.075	**1.111**	1.110
24	1.161	1.388	**1.121**	1.248	**1.651**	1.286
25	1.226	**1.079**	1.137	1.152	**1.227**	1.180
26	1.234	1.054	**0.847**	**1.128**	1.333	1.026
27	1.273	**1.396**	1.390	**1.360**	1.491	1.412
28	1.285	1.228	**1.484**	**1.345**	1.148	1.207
29	1.344	**1.509**	1.094	**1.245**	1.355	1.253
30	1.361	0.924	**0.923**	**0.752**	1.321	0.955
31	1.467	**1.651**	1.519	**1.532**	1.521	1.580
32	1.525	1.536	**1.497**	**1.929**	1.586	1.789
33	1.549	**1.185**	1.496	**1.216**	1.443	1.204
34	1.559	1.649	**1.519**	**1.584**	1.578	1.666
35	1.634	**1.270**	1.390	**1.350**	1.698	1.468
36	1.676	1.819	**1.664**	**1.817**	1.696	1.414
37	1.702	**1.564**	1.549	**1.438**	1.929	1.508
38	1.708	1.800	**1.997**	**1.861**	1.823	1.741
39	1.709	**1.325**	1.742	**1.427**	1.496	1.760

Table 5 continued

No. in Table 1	T_{obs}	T_{calc} by QSAR				
		#1	#2	#3	#4	#5
40	1.728	1.589	**1.402**	**1.549**	1.639	1.568
41	1.834	**1.920**	1.924	**1.862**	1.748	2.004
42	1.993	1.674	**1.512**	**1.661**	2.090	1.865
43	2.210	2.485	**2.209**	**2.290**	2.292	2.195
44	2.219	**2.715**	2.314	**2.947**	2.353	2.468
45	2.293	1.955	**1.754**	**2.166**	2.055	2.204
46	2.481	**2.758**	2.416	**2.660**	2.533	2.583
47	2.531	2,382	**2.025**	**2.163**	2.386	2.284
48	2.544	**2.848**	2.597	**3.956**	2.491	2.423
49	2.638	2.737	**3.019**	**3.617**	2.494	2.768

Quality of prediction for the calibration set:

$$r^2 = .9484 \;\; F = 282.0 \;\; SEE = .1501 \;\; Q = .8904$$

The set of significant molecular fragments identified in QSAR study #5 includes all significant molecular fragments identified in QSAR studies #1–#4, with the exception of fragment C_6H_5.

According to the physical meaning of predictors in QSAR studies #1–#4 the influence of the chemical structure on toxicity is higher than the influence of molecular size and shape. On the contrary, according to the physical sense of predictors in QSAR study #5, the influence of the chemical structure on toxicity is missing and the toxicity depends on molecular size and shape only.

Table 5 includes the observed values versus calculated values of toxicity in QSAR studies #1–#5. The calculated values for prediction set molecules are bolded.

4 Conclusions

The calculated influence of certain molecular feature (chemical structure, size, shape, hydrophilicity) on the toxicity of phenol derivatives depends on the structure of the database (number of molecules and diversity of molecules having analyzed molecular feature).

We observed a direct or inverse correlation between the quality of prediction for the prediction set (according to r^2 and RSEE), the similarity of sets and the diversity of molecules in sets (according to Tables 3 and 4).

More precisely:

- there is a direct correlation with the similarity of sets, regardless of analyzed feature (chemical structure, size, shape, hydrophilicity)

- there is an inverse correlation with the diversity of molecules in the calibration set, from the point of view of chemical structure, size and shape
- there is a direct correlation with the diversity of molecules in the calibration set, from the point of view of hydrophilicity
- there is a direct correlation with the diversity of molecules in the prediction set, regardless of analyzed feature

The analysis of much greater number of databases should permit identification of the formula for the suitability (adequacy), function of number of molecules in sets, diversity of molecules in sets and similarity of sets, from the point of view of some significant molecular features. This formula will be, in principle, a measure of the adequacy of a database for following QSAR studies.

References

1. P. Gramatica, P. Pilutti, E. Papa, SAR QSAR Environ. Res. **13**, 743 (2002)
2. P. Gramatica, P. Pilutti, E. Papa, QSAR Comb. Sci. **22**, 364 (2003)
3. D.M. Hawkins, S.C. Basak, D. Mills, J. Chem. Inf. Comp. Sci. **43**, 579 (2003)
4. C. Helma, SAR QSAR Environ. Res. **15**, 367 (2004)
5. P. Gramatica, QSAR Comb. Sci. **26**, 694 (2007)
6. P.P. Roy, S. Paul, I. Mitra, K. Roy, Molecules **14**, 1660 (2009)
7. L. Tarko, C.T. Supuran, Bioorg. Med. Chem. **21**, 1404 (2013)
8. P. Gramatica, P. Pilutti, Report in Joint Research Centre (European Commission), contract ECVA-CCR.496576-Z
9. A. Golbraikh, M. Shen, Z. Xiao, Y.-D. Xiao, K.-H. Lee, A. Tropsha, J. Comp-Aid, Mol. Des. **17**, 241 (2003)
10. R. Carbo, M. Arnau, L. Leyda, Int. J. Quantum. Chem **17**, 1185 (1980)
11. R. Carbo, B. Calabuig, *Concepts and Applications of Molecular Similarity* (Wiley, New- York, 1990), pp. 147–171
12. R. Carbo-Dorca, P.G. Mezey, Advances in Molecular Similarity, vol. 1 (JAI Press, Greenwich, 1996), pp. 89–120
13. H. Kubinyi, Persp. Drug Discov. Des. **9**, 225 (1998)
14. Y.C. Martin, J.L. Kofron, L.M. Traphagen, J. Med. Chem. **45**, 4350 (2002)
15. J.L. Durant, B.A. Leland, D.R. Henry, J.G. Nourse, J. Chem. Inf. Comput. Sci. **42**, 1273 (2002)
16. N. Nikolova, J. Jaworska, QSAR Comb. Sci. **22**, 1006 (2003)
17. L. Ralaivola, S.J. Swamidass, S. Hiroto, P. Baldi, Neural Netw. **18**, 1093 (2005)
18. S.A. Rahman, M. Bashton, G.L. Holliday, R. Schrader, J.M. Thornton, J. Cheminform. **1**, 12 (2009)
19. M.S. Armstrong, J. Mol. Graph. Model. **28**, 368 ((2009)
20. P.M. Petrone, B. Simms, F. Nigsch, E. Lounkine, P. Kutchukian, A. Cornett, Z. Deng, J.W. Davies, J.L. Jenkins, M. Glick, ACS Chem. Biol. **17**, 1399 (2012)
21. S. Nallusamy, S. Selvaraj, Bioinformation **8**, 498 (2012)
22. C. Li, L.M. Colosi, SAR QSAR Environ. Res. **24**, 679 (2013)
23. T.A. Roy, A.J. Krueger, C.R. Makerer, W. Neil, A.M. Arroyo, J.J. Yang, Dermal penetration capacity of some PAHs. SAR QSAR Environ. Res. **9**, 171 (1998)
24. C. Rücker, M. Meringer, A. Kerber, Boiling point of some fluoroalkanes. J. Chem. Inf. Model. **45**, 74 (2005)
25. O. Ivanciuc, T. Ivanciuc, P.A. Filip, D. Cabrol-Bass, Viscosity of quite various compounds. J. Chem. Inf. Sci. **39**, 515 (1999)
26. J.S. Barker, C.K. Hattotuwagama, M.G.B. Drew, Sweetness power of some guanidine derivatives. Pure Appl. Chem. **74**, 1207 (2002)
27. L.H. Hall, T.A. Vaughn, Med. Chem. Res. **7**, 407 (1997)
28. K. Roy, G. Ghosh, Int. Elect. J. Mol. Des. **2**, 599 (2003)
29. PCModel Program is Available from Gajewski, J. J.; Gilbert, K. E., Serena Software, Box 3076, Bloomington, IN, USA

30. MOPAC Program is Available from Stewart, J.J.P., 15210 Paddington Circle, Colorado Springs, CO 80921 E-mail: MrMOPAC@OpenMOPAC.net; http://www.openmopac.net/
31. PRECLAV Program is Available from Center of Organic Chemistry—Bucharest—Romanian Academy; ltarko@cco.ro; tarko_laszlo@yahoo.com
32. L. Tarko, Rev. Chim. (Bucuresti) **55**, 539 (2004)
33. C.E. Shannon, Bell Syst. Tech. J. **27**, 379 (1948)
34. L. Tarko, J. Math. Chem. **49**, 2330 (2011)
35. L. Tarko, Rev. Chim. (Bucuresti) **55**, 169 (2004)
36. See the documentation of PRECLAV, last version