J Math Chem (2010) 47:174–190
DOI 10.1007/s10910-009-9540-6

ORIGINAL PAPER

Monte Carlo method for identification of outlier molecules in QSAR studies

Tarko Laszlo

Received: 2 December 2008 / Accepted: 24 February 2009 / Published online: 14 March 2009

Abstract The paper presents some difficulties that appear in the application of the classical formula in the identification of "outliers" in a given objects set. The paper proposes a new Monte Carlo-like method for the identification of "outliers" in the calibration set used in QSPR/QSAR computations. Sub-sets of molecules are randomly extracted thousands of times from the given calibration set. The method relies on the idea that the presence of "outlier" molecules in a certain sub-set decreases the prediction power of the QSAR equation that used this particular sub-set of molecules. The presence of "outlier" molecules often leads to poor quality QSAR equations and rarely to high quality QSAR equations. The paper proposes a specific formula for "outlier index". The molecule with the highest value of the outlier index is eliminated out of the calibration set. The identification/elimination process is repeated until the maximum value of the outlier index stops decreasing. The paper presents five examples of outliers' identification using various kinds of calibration sets. We compare the results with the results obtained by a classical outlier index formula, using the same calibration set, the same set of descriptors and the same outlier identification/elimination procedure.

Keywords Monte Carlo · Outliers · Qsar

1 Introduction

Let us consider a certain set of objects having some common features, e.g., weight, number of atoms, height, smell, blood pressure, age, political opinions, electrical charge, etc. In the widest sense of the word, the "outlier" object in certain group is

T. Laszlo (✉)
Center of Organic Chemistry "C. D. Nenitzescu"–Romanian Academy, 6th Sector,
202B Spl. Independentei, PO Box 35-108, Bucharest, MC 060023, Romania
e-mail: ltarko@cco.ro

the one that "resembles" only a small number of the other objects in the group. The resemblance between objects, from the point of view of one given property P, can be measured quantitatively when P has a specified numeric value for each object. Here, the properties having computable numeric values are called "descriptors".

The distribution of descriptor values, used for comparison, can be Gaussian. In this case, most descriptor values are close to the average value and a few of them are further apart. The graph of the distribution function f(x) is also symmetrical.

$$f(x) = a \cdot e^{y(x)} \tag{1}$$

where

$y(x) = -b \cdot (x - c)^2/d^2$
$a, b, c, d > 0$

The values of height for all the people of Earth are distributed Gaussian. There are a few people with very large or very small height, but a very large number of people with height close to average and the graph of number of people versus height is symmetrical. The values of velocity of a pure gas molecules are distributed (almost) Gaussian (actually according to Maxwell–Boltzmann function). There are few molecules very fast or very slow, but a very large number of molecules have a velocity close to the average and the graph of molecule number vs. velocity is symmetrical.

For such Gaussian distribution, the "outlier" term means "far from the average" [1,2]. Here, this "outlier" objects are named "type A outliers" and obey the classical condition (2).

$$|V - V_m| / \sigma > k \tag{2}$$

where

V is the descriptor value for the analyzed object
V_m is the average of descriptor values in the group of analyzed objects
σ is the standard deviation of descriptor values in the group of analyzed objects
k is a factor, real number

In practice, in the absence of any theoretical reasons, the value of factor k for outlier objects is empirically assigned within range [1.5, 3.0].

If the distribution of the descriptor's values, used for comparison of objects, is not Gaussian, there might be many objects whose descriptor values are far from the average value and only a few near the average. For instance, the positive and negative electric charges of ions in salt crystals are far apart from the average; there are no ions having almost zero electrical charge, which would be considered close to the average value of charge. Similarly, the weights of the individual members of a species with high sexual dimorphism are far from the average weight. There are only few individuals with a weight close to the average weight. The atomic masses of elements, at Universe scale, present a non-Gaussian distribution. In such non-Gaussian cases, the identification of type A outliers requires different formulae and procedures [3–8]. For instance, the objects are grouped in classes ("clusters"), based on the descriptor

value; the "outliers" are the objects included in the clusters having small number of objects [9,10].

The objects from the analyzed group can be molecules with a certain structure. A QSPR (Quantitative Structure Property Relationship) is the mathematical formula of some property, common to all the analyzed molecules. In this formula, obtained by statistical methods, the studied property is called "dependent property" and expressed as a function of various molecular descriptors. To obtain a QSPR we use the observed values of the dependent property and of the molecular descriptors for the group of analyzed molecules, called here "calibration set". If the dependent property is biochemical activity then Property = Activity and QSPR = QSAR. In practice, in the QSPR/QSAR field, the specialty literature makes a distinction between "type I outlier" (an object with an erroneously measured value of the descriptor or the dependent property) and "type II outlier" (an object with a correctly measured but far from the average value of the descriptor or the dependent property) [11–14].

For the molecules in the calibration set, the estimated values of biochemical activity (using QSAR equation) are not always very close to the observed values. The molecules in the calibration set for which the QSAR equation yields much weaker predictions than for rest of the molecules are called here "type B outliers". Type B outliers can be either type I or type II. To identify type B outliers a formula like (2) can be used. In this case, the descriptor is "the difference between the observed and the computed value of the dependent property" [15,16]. As a rule, the values of this descriptor are within $(-\infty, +\infty)$ range. It is assumed that the distribution of values of this descriptor is Gaussian. In formula (2) $V = V_{obs} - V_{calc}$, V_m is the average of these differences (usually very close to zero), and σ the standard deviation of these differences. Criterion (2) becomes:

$$\left|V_{obs} - V_{calc}\right| / \sigma > k \tag{3}$$

The ratio $|V_{obs} - V_{calc}|/\sigma$ can be used as the "outlier index".

If σ in formula (3) is very large compared to the average value of V_{obs}, criterion (3) is difficult to fulfill, and only very few type B outliers are usually identified. Indeed, if the estimation quality for the whole calibration set is very low then it is improbable that the estimation quality for a few certain molecules will be even lower. A large number of "outlier" molecules indicates a low quality estimation, but a small number of "outliers" says nothing about the quality of the estimation overall.

As a rule, the elimination of certain "outliers" induces the homogenization of the calibration set. According to statistical logic, the elimination of "outlier" molecules from the calibration set will lead to another QSAR, with better predictive power. The new QSAR equation, obtained using the reduced calibration set (without the identified "outlier" molecules), might include other descriptors and the value of the weighting factors will be different. The new QSAR may identify other type B "outlier" molecules because, in formula (3), lowering of numerator value is accompanied by the lowering of the denominator value. Consequently, the value of ratio in formula (3), may remain the same or may even increase. Therefore, the elimination of "outlier" molecules using criterion (3) is no warranty for lowering of outlier number in the next QSAR computations. It is, probably, the reason why the specialty literature

recommends the elimination of "outlier" molecules out of the calibration set only once [1,17–20]. Another theory [21] maintains that the elimination of "outlier" molecules must be repeated as long as the prediction is ... wrong enough, in other words as long as the denominator value in formula (3) remains large enough. An iterative discarding method for the selection of accurate data points is presented in [22].

Within a non-homogeneous calibration set, each sub-class of molecules may include a large number of molecules. Consequently, there are not "outliers" according to criterion (3), although each sub-class is "outlier" for the other classes.

Within non-homogeneous calibration set, each sub-class of molecules is best described by another QSPR equation. In this case, computation of equations with acceptable predictive power for the whole calibration set becomes doubtful, irrespective of presence/absence of "outlier" molecules. The incorrect choice of descriptors used in computations has a similar effect.

Some molecules are identified as "outliers" because the calibration set includes other "outlier" molecules. Consequently, it is not very clear if the elimination of "outliers" must be applied on all molecules that fulfill criterion (3) or only on the molecule with the highest value of "outlier index".

The observed values V_{obs} of biochemical activity may be within a narrow range (small value of variance coefficient) or a large range (large value of variance coefficient). The difference $|V_{obs} - V_{calc}|$ does not offer information about the rank of the molecules in sets ordered by the values of V_{obs} and V_{calc}. A good agreement of observed/computed values may be associated with a poor agreement of ranks of observed/computed values or the other way round. There is not any theoretical explanation for the usage, in formula (2) and (3), of values, not of the ranks of values.

Different statistical algorithms identify, within the same calibration set, different sets of "outliers", because they use different QSPR/QSAR equations.

Within a given calibration set, a molecule might be identified as "type B outlier" for various reasons:

- a different biochemical mechanism from other molecules, because, maybe, there are several active sites on the receptor macromolecule or unusual binding mode [23,24]
- same biochemical mechanism and very different chemical structure
- wrong computed value of descriptors due to faulty chemical structure [25]
- erroneous observed value V_{obs} used in computations
- presence of other "outlier" molecules in the calibration set, etc.

Some "type B outliers" need special attention if:

- the chemical structure, used in computations, is correct
- the observed value V_{obs}, used in computations, is correct
- the observed value V_{obs} is "large"
- very pronounced "outlier index" value

These "type B outlier" molecules are possibly good starting points for the development of new classes of biochemical active compounds because they are both very active and very different from the other molecules in the calibration set. This type of molecules is quoted as "outliers for lead hopping" [26,27]. As a rule, this term may

be used also when the dependent property is different from "biochemical activity". We think the identification of "outliers for lead hopping" is the main goal of using of formula (3). Consequently, the use of "robust" statistical methods, having a low sensibility to the presence of outliers, must be avoided.

Is it necessary and obligatory to eliminate "outlier" molecules? Theoretically, the aim of the computations in the absence of "outlier" molecules is to find a QSAR equation with high predictive power. However, simply finding a QSAR with better predictive power is not a goal by itself.

In practice, we frequently find out that by elimination of "outlier" molecules the structure of QSAR obtained with the reduced set of molecules is only slightly modified (number and types of descriptors, the algebraic sign of the coefficients). In these situations, the data considered useful from a drug design point of view are not influenced by the presence/absence of "outlier" molecules. However, the elimination of "outlier" molecules gives a higher degree of confidence in the results obtained from the new QSPR equation, whatever those conclusions might be.

Many QSAR studies are done in the presence of an "external validation set" (containing molecules with known values of the dependent property) or in the presence of a "prediction set" (containing molecules that have not yet been synthesized, with unknown values of the dependent property), sets containing molecules that have NOT been used for calculating the QSAR equation. The increase in the predictive value of the calibration set is not always followed by an increase in the predictive value for the prediction set or for the external validation set. When there is a prediction set or an external validation set the elimination of "outlier" molecules from the calibration set must be made very carefully, in order to not diminish the "representative sample" character of the calibration set in the group calibration set + prediction/external validation set [28,29].

In molecular analysis, the Monte Carlo type methods are usually applied in the "Molecular Dynamics" studies and "Conformational Analysis". In QSAR practice, the Monte Carlo type methods are used in the selection of "significant" descriptors [30] and in the identification of the optimal QSAR equation [31]. Along the same line, the "leave-n-out cross-validation" ($n > 1$) procedures select randomly various sets of molecules in order to calculate the quality of a certain set of descriptors [32].

This paper proposes a new Monte Carlo-like method for the identification of type B outlier molecules in a given calibration set.

2 Methods and formulas

Given a calibration set with N molecules and N observed values of the dependent property.

The calculation of descriptors has been done after geometry optimization using the MMX method (PCModel) [33] and PM6 method (MOPAC) [34,35]. For every molecule, we have calculated values for around 400 "whole molecule descriptors", specific to the PRECLAV software [36–39].

Eqs. 4, 5, and 6 define the constants L, n and p.

$$\mathrm{L} = \mathrm{x} \cdot N^{1/2} \tag{4}$$

$$n = N / \mathrm{y} \tag{5}$$

$$\mathrm{p} = N / (\mathrm{y} \cdot \mathrm{z}), \text{ but larger than 1 and smaller than 11} \tag{6}$$

where

x, y and z are factors, real numbers

The used values for L, $n$and p are integers obtained by rounding up the calculated values.

The proposed method for the identification of type B outlier molecules in the calibration set with N molecules implies the following steps:

(a) random extraction from the set of N molecules of a group of n molecules
(b) identification, by a heuristic procedure, in the set of molecular descriptors, of the "optimal" set of descriptors with the highest "quality", considered the best for describing the group of n molecules
(c) the quality of the "optimal" set identified in step (b) becomes the quality of the group of n molecules extracted in step (a)
(d) steps (a), (b), and (c) are repeated, retaining the extracted groups and their quality until the number of extracted groups and quality values equals $2 \cdot \mathrm{L}$
(e) the $2 \cdot \mathrm{L}$ groups are arranged by quality; thus, we obtain L "worst groups" of molecules and L "best groups" of molecules
(f) steps (a), (b), (c) are redone; depending on the calculated value of the quality, the group extracted at step (a) will replace the "bad group" with the highest quality, will replace the "good group" with the lowest quality or will not be used
(g) the L "worst groups" and the L "best groups" are arranged according to their quality
(h) for every molecule present in the $2 \cdot \mathrm{L}$ groups the function WB (worst – best) is calculated

$$\mathrm{WB} = (\mathrm{w} - \mathrm{b}) / \mathrm{L} \tag{7}$$

where

w is presence number within "worst" groups
b is presence number within "best" groups
L is defined in formula (4)

steps (f), (g), and h) are repeated until a criterion for stopping the calculations applies

The values of function WB are within the range $[-1, +1]$. The average of WB is always very close to zero. The standard deviation of WB will be represented as σ_{WB} from now on.

During computation, the value of WB increases gradually and the value of the ratio $\mathrm{WB}/\sigma_{\mathrm{WB}}$ remains approximately constant. As the computations proceed, the quality of the best groups becomes better and the quality of the worst groups becomes smaller.

Some molecules have a higher contribution to the decreasing of the group quality. These molecules are present many times w inside the worst groups and rarely b inside the best groups ($w >> b$).

The WB algorithm may use different values for the constants x, y, and z, different random procedures for step (a) and different heuristic procedures for step (b). A variety of quality functions can be used as well as a number of criteria for stopping the calculation. One can define various criteria for the "outlier" character, depending on the absolute value of WB or the relative value of WB by comparing the values of WB for all molecules.

The results presented here have been obtained in particular conditions, which will be presented below.

For the constants in formulas (4), (5), and (6) we have used $x = 20$, $y = 4$, $z = 3$. The value for x has been chosen empirically. When $y = 4$ the subsets randomly extracted represent a quarter of the calibration set, percentage that seems intuitively correct. For z, we have chosen the value three because the number of descriptors in the calculated equation has to be small enough in comparison with the number of molecules in the calibration (sub)set.

For step (a), we used a pseudo-random procedure. To be precise, we chose, one by one, each molecule (first, second, third, etc.) included in the set of N molecules. Every time, the set of n molecules has been filled with other $n - 1$ molecules, chosen at random. The use of the last molecule in set of N molecules marks the end of a "computation loop".

The heuristic procedure from step (b) calculates multilinear Eq. 8.

$$\mathrm{A} = \mathrm{c}_0 + \sum_{i=1}^{p} \mathrm{c}_i \cdot \mathrm{d}_i \tag{8}$$

where

- A is biochemical activity or other dependent property
- c_0 is intercept, real number
- c_i are weighting factors, real numbers
- d_i are descriptor values
- p is descriptor number in formula (6)

This procedure adds the descriptors successively. For every equation, we calculate the square of the linear Pearson correlation r^2 between the observed values and the calculated values (using the equation in question) for the activity A, for the set of n molecules. Descriptor $\mathrm{d}_1$is the one that determines the highest value for r^2 in mono-linear equations. Descriptor d_2 is the descriptor that, together with d_1, gives the highest value for r^2 in bilinear equations. Descriptor d_3 is the descriptor that, together with d_1 and d_2 gives the highest value of r^2 in tri-linear equations, etc. The set considered here as "optimal" is the one that contains p descriptors. For the computation of coefficients c_i in formula (8), we have used Ordinary Least Square Method. The heuristic procedure used does not eliminate the "non-significant" descriptors and does not check the intercorrelation of the descriptors in the same set. The calculated value

of r^2 for a set with p descriptors is the "quality" of the "optimal" set of descriptors and has been used in step (c). This heuristic procedure has been used because the computation time required is rather short.

The ratio WB/σ_{WB}, calculated for each molecule, has been considered here "outlier index". The computations have been stopped when, for 80 consecutive "computation loops", the same set of molecules with $WB/\sigma_{WB} > 2.5$ has been identified or, rarely, when the number of consecutive "computation loops" has reached 400. These criteria for stopping the computations insure a good reproducibility of results.

The molecule for which we calculate the highest value of the ratio WB/σ_{WB} was considered "possible outlier". This "possible outlier" is eliminated from the calibration set and all the computations are repeated using the reduced calibration set. If we observe a real decrease in the value of the ratio WB/σ_{WB} then the "possible outlier" becomes a "true outlier" and a new "possible outlier" is identified. The procedure of elimination/identification of "possible outlier"/"true outlier" is repeated as long as the maximum value of the ratio WB/σ_{WB} keeps decreasing.

The application of the WB algorithm has been implemented with a computation module made with that end in view.

The proposed Monte Carlo type method extracts at random, thousands of times, sets of molecules included in the given calibration set and, every time, calculates a different QSAR equation, using other descriptors. We hope that the identification of an "outlier" will be, in these conditions, not so much influenced by the presence in the calibration set of a certain number of other "outlier" molecules and by the variation coefficient of the values of the dependent property.

3 Results and comments

We present here, as an example, the results of the application of the WB algorithm for the identification of "outlier" molecules included in various calibration sets.

Comparing the results with the results of the application of the classic criteria (3) is difficult because the QSAR programs using criteria (3) use, for step b), more complex heuristic procedures, use only "significant" descriptors selected according to specific procedures and the intercorrelation of the descriptors in the same set is low. Moreover, the programs for QSAR computations use a single equation for the identification of "outlier" molecules, more precisely the equation that best describes the calibration set as a whole. However, we will present here the results obtained with the PRECLAV software. The set of descriptors used for selection was the same and the "PRECLAV outlier index" was considered the ratio $|V_{obs} - V_{calc}|/\sigma$. The same procedure was used for the elimination/identification of "possible outlier"/"true outlier".

Analysis 1

Calibration set: 50 substituted phenols in Table 1
Dependent property: toxicity T against *Tetrahymena pyriformis*, where $T = 0.431 + \log(1/C)$. The observed values of C (nM) are quoted in literature [40,41].

Table 1 Structure and toxicity of substituted phenols

Index	Substituents	T_{obs}	Index	Substituents	T_{obs}
1	H	0.000	26	3,5-dichloro	1.993
2	2-fluoro	0.679	27	2,4-dibromo	1.834
3	3-fluoro	0.904	28	2-chloro-5-methyl	1.071
4	4-fluoro	0.448	29	2-methyl-4-chloro	1.131
5	2-chloro	0.708	30	3-methyl-4-chloro	1.226
6	4-chloro	0.976	31	2,3-dimethyl	0.553
7	2-bromo	0.935	32	2,4-dimethyl	0.559
8	4-bromo	1.112	33	2,5-dimethyl	0.440
9	3-iodo	1.549	34	3,4-dimethyl	0.553
10	4-iodo	1.285	35	3,5-dimethyl	0.544
11	3-methyl	0.369	36	2-*tert*-butyl-4-methyl	1.728
12	4-methyl	0.239	37	2,6-diphenyl	2.544
13	2-ethyl	0.607	38	2,4,5-trichloro	2.531
14	3-ethyl	0.660	39	2,6-dichloro-4-bromo	2.210
15	2-*iso*-propyl	1.234	40	2,4,6-tribromo	2.481
16	3-*iso*-propyl	1.040	41	2-methyl-4-bromo-6-chloro	1.708
17	4-*iso*-propyl	0.904	42	2,4-dibromo-6-phenyl	2.638
18	3-*tert*-butyl	1.161	43	3,5-dimethyl-4-chloro	1.634
19	4-*tert*-butyl	1.344	44	2-*iso*-propyl-4-chloro-5-methyl	2.293
20	2-phenyl	1.525	45	2,6-dimethyl-4-bromo	1.709
21	2,6-difluoro	0.827	46	2,3,6-trimethyl	0.849
22	3-chloro-4-fluoro	1.273	47	3,4,5-trimethyl	1.361
23	2,3-dichloro	1.702	48	2,4,6-trimethyl	2.126
24	2,4-dichloro	1.467	49	2,4-dimethyl-6-*tert*-butyl	1.676
25	2,5-dichloro	1.559	50	2,6-di-*tert*-butyl-4-methyl	2.219

"Possible" WB outliers/WB outlier index: molecule **48**/5.012 → molecule **23**/2.676 → molecule **47**/2.871
"True" WB outliers: molecule **48**
"Possible" PRECLAV outliers/PRECLAV outlier index: molecule **48**/3.818 → molecule **23**/2.320 → molecule **47**/2.223 → molecule **41**/2.015 → molecule **9**/2.248
"True" PRECLAV outliers: molecules **48, 23**, and **47**
Quality of PRECLAV QSAR in presence of all molecules: $r^2 = 0.8537 : F = 91.5$
Quality of PRECLAV QSAR in absence of the outlier **48** : $r^2 = 0.9032 : F = 143.1$
Quality of PRECLAV QSAR in absence of all identified outliers: $r^2 = 0.9269 : F = 186.0$

There is a high difference in toxicity between molecule **48** and molecules **46** and **47**, although the chemical structures are very similar. It is possible that the observed

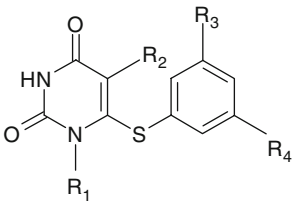


Fig. 1 Anti-HIV agents that present cytotoxic activity

value for the toxicity of molecule **48** is wrong. The elimination of "outlier" molecules greatly increases the quality of the calculated QSAR equations.

Analysis 2

Calibration set: 49 HEPT (1-[(2-hydroxyethoxy)-methyl]-6-(phenylthio)-timine) analogues in Fig. 1; Table 2
Dependent property: cytotoxic activity A = $\log(743/CC_{50})$. The observed values of $CC_{50}(\mu M)$ are quoted in literature [42].
"Possible" WB outliers/WB outlier index: molecule **20**/4.501 → molecule **46**/2.554 → molecule **45**/3.049
"True" WB outliers: molecule **20**
"Possible" PRECLAV outliers/PRECLAV outlier index: molecule **20**/2.725 → molecule **21**/3.345
"True" PRECLAV outliers: none
Quality of PRECLAV QSAR in presence of all molecules: $r^2 = 0.6717 : F = 23.0$
Quality of PRECLAV QSAR in absence of the molecule **20** : $r^2 = 0.3217 : F = 10.9$

Molecule **20** presents the highest observed (correct?) value for activity, much higher than the average. The mediocre quality of the PRECLAV equation in the presence of molecule **20**, suggests a lack of homogeneity of the calibration set in Table 2. By eliminating molecule **20,** we seem to accentuate the lack of homogeneity, at least from the point of view of PRECLAV.

Analysis 3

Calibration set: 68 heterocycles in Fig. 2; Table 3
Dependent property: antifungal activity against *Candida albicans*. The observed activity values are quoted in literature [43].
"Possible" WB outliers/WB outlier index: molecule **56**/2.965
→ molecule **46**/3.073
"True WB outliers": none
"Possible" PRECLAV outliers/PRECLAV outlier index: molecule **46**/2.723 → molecule **41**/2.365 → molecule **43**/2.595
"True" PRECLAV outliers: molecule **46**
Quality of PRECLAV QSAR in presence of all molecules: $r^2 = 0.7400 : F = 18.7$
Quality of PRECLAV QSAR in absence of molecule **46** : $r^2 = 0.7511 : F = 22.3$

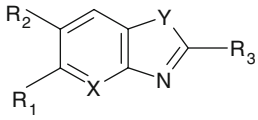
Table 2 Structure and cytotoxic activity for molecules in Fig. 1

Index	R_1	R_2	R_3	R_4	A_{obs}
1	CH_2–O–$(CH_2)_2$–OH	CH_3	CH_3	H	0.248
2	CH_2–O–$(CH_2)_2$–OH	CH_3	C_2H_5	H	0.613
3	CH_2–O–$(CH_2)_2$–OH	CH_3	*tert*-C_4H_9	H	0.996
4	CH_2–O–$(CH_2)_2$–OH	CH_3	CH_2–OH	H	0.406
5	CH_2–O–$(CH_2)_2$–OH	CH_3	CF_3	H	0.579
6	CH_2–O–$(CH_2)_2$–OH	CH_3	F	H	0.421
7	CH_2–O–$(CH_2)_2$–OH	CH_3	Cl	H	0.549
8	CH_2–O–$(CH_2)_2$–OH	CH_3	Br	H	0.722
9	CH_2–O–$(CH_2)_2$–OH	CH_3	I	H	0.846
10	CH_2–O–$(CH_2)_2$–OH	CH_3	NO_2	H	0.641
11	CH_2–O–$(CH_2)_2$–OH	CH_3	OH	H	0.222
12	CH_2–O–$(CH_2)_2$–OH	CH_3	CH_3	CH_3	0.485
13	CH_2–O–$(CH_2)_2$–OH	CH_3	Cl	Cl	0.757
14	CH_2–O–$(CH_2)_2$–OH	CH_3	$COOCH_3$	H	0.527
15	CH_2–O–$(CH_2)_2$–OH	CH_3	$COCH_3$	H	0.513
16	CH_2–O–$(CH_2)_2$–OH	CH_3	COOH	H	0.324
17	CH_2–O–$(CH_2)_2$–OH	CH_3	$COONH_2$	H	0.385
18	CH_2–O–$(CH_2)_2$–OH	CH_3	CN	H	0.502
19	CH_2–O–$(CH_2)_2$–OH	CH_2–CH=CH_2	H	H	0.609
20	CH_2–O–$(CH_2)_2$–OH	$COOCH_3$	H	H	2.051
21	CH_2–O–$(CH_2)_2$–OH	$COONHC_6H_5$	H	H	1.616
22	CH_2–O–$(CH_2)_2$–OH	C_2H_5	H	H	0.269
23	CH_2–O–$(CH_2)_2$–OH	*n*–C_3H_7	H	H	0.484
24	CH_2–O–$(CH_2)_2$–OH	*iso*-C_3H_7	H	H	0.507
25	CH_2–O–$(CH_2)_2$–OH	C_2H_5	CH_3	CH_3	0.698
26	CH_2–O–$(CH_2)_2$–OH	*iso*-C_3H_7	CH_3	CH_3	0.764
27	CH_2–O–$(CH_2)_2$–OH	C_2H_5	Cl	Cl	1.163
28	CH_2–O–$(CH_2)_2$–OH	H	H	H	0.000
29	CH_2–O–$(CH_2)_2$–OCH_3	CH_3	H	H	0.395
30	CH_2–O–$(CH_2)_2$–O–*n*–C_5H_{11}	CH_3	H	H	1.131
31	CH_2–O–$(CH_2)_2$–O–$CH_2C_6H_5$	CH_3	H	H	1.218
32	CH_2–O–CH_3	CH_3	H	H	0.484
33	CH_2–O–C_2H_5	CH_3	H	H	0.507
34	CH_2–O–*n*–C_3H_7	CH_3	H	H	0.704
35	CH_2–O–*n*–C_4H_9	CH_3	H	H	0.952
36	CH_2–O–$(CH_2)_2$–$Si(CH_3)_3$	CH_3	H	H	1.366
37	CH_2–O–CH_2–C_6H_5	CH_3	H	H	0.893
38	CH_2–O–C_2H_5	C_2H_5	H	H	0.664
39	CH_2–O–C_2H_5	C_2H_5	Cl	Cl	1.218
40	CH_2–O–*iso*-C_3H_7	C_2H_5	H	H	0.716
41	CH_2–O–*ciclo*-C_6H_{11}	C_2H_5	H	H	1.641

Table 2 continued

Index	R_1	R_2	R_3	R_4	A_{obs}
42	CH_2–O–CH_2–C_6H_5	C_2H_5	H	H	1.340
43	CH_2–O–$(CH_2)_2$–C_6H_5	C_2H_5	H	H	1.291
44	CH_2–O–C_2H_5	*iso*-C_3H_7	H	H	0.846
45	CH_2–O–C_2H_5	*ciclo*-C_3H_5	H	H	0.521
46	H	CH_3	H	H	0.473
47	CH_3	CH_3	H	H	0.695
48	C_2H_5	CH_3	H	H	0.898
49	*n*–C_4H_9	CH_3	H	H	0.922

Fig. 2 Antifungal heterocyclic molecules



The calibration set in Table 3 lacks "outlier" molecules, at least according to WB algorithm. The mediocre quality of the equations (low enough value of F) calculated with PRECLAV in the presence/absence of molecule **46** suggests the lack of homogeneity of the calibration set in Table 3. The elimination of molecule **46** has a very weak effect over the quality of QSAR equation.

Analysis 4

Calibration set: 33 molecules in Table 4
Dependent property: melting point (°K). The observed values of dependent property are quoted in Internet databases [44,45].
"Possible" WB outliers/WB outlier index: molecule **9**/2.456 → molecule **21**/2.160 → molecule **33**/1.663 → molecule **32**/2.001
"True WB outliers": molecules **9** and **21**
"Possible" PRECLAV outliers/PRECLAV outlier index: molecule **11**/2.123 → molecule **33**/2.245
"True" PRECLAV outliers: none
Quality of PRECLAV QSPR in presence of all molecules: $r^2 = 0.7836 : F = 36.2$
Quality of PRECLAV QSPR in absence of the molecules **9** and **21**: $r^2 = 0.8206 : F = 42.7$

Intuitively, for chemists, the last five molecules in Table 4 should be "outliers", because they present a combination of chemical groups very different from the structure of the hydrocarbons. The results obtained suggest though that the chemical groups have a much-reduced influence over the melting point. Probably, the melting point is influenced more by the size of the molecule, its symmetry and its ability to form intermolecular hydrogen bonds. From the point of view of the PRECLAV program, the elimination of WB outliers has a weak effect over the quality of the QSAR equation.

Table 3 Structure and antifungal activity for compounds in Fig. 2

Index	X	Y	R_1	R_2	R_3	A_{obs}
1	CH	O	H	H	C_6H_5	3.892
2	CH	O	H	H	C_6H_4-*para*-$C(CH_3)_3$	4.001
3	CH	O	H	H	C_6H_4-*para*-NH_2	3.924
4	CH	O	H	H	C_6H_4-*para*-$NHCH_3$	3.952
5	CH	O	Cl	H	C_6H_4-*para*-C_2H_5	4.013
6	CH	O	Cl	H	C_6H_4-*para*-$NHCOCH_3$	4.059
7	CH	O	Cl	H	C_6H_4-*para*-$NHCH_3$	4.015
8	CH	O	Cl	H	C_6H_4-*para*-Cl	4.024
9	CH	O	Cl	H	C_6H_4-*para*-NO_2	4.040
10	CH	O	NO_2	H	C_6H_5	4.282
11	CH	O	NO_2	H	C_6H_4-*para*-CH_3	4.308
12	CH	O	NO_2	H	C_6H_4-*para*-$C(CH_3)_3$	4.375
13	CH	O	NO_2	H	C_6H_4-*para*-NH_2	4.310
14	CH	O	NO_2	H	C_6H_4-*para*-Cl	4.342
15	CH	O	NO_2	H	C_6H_4-*para*-Br	4.406
16	CH	O	NH_2	H	C_6H_4-*para*-C_2H_5	3.979
17	CH	O	NH_2	H	C_6H_4-*para*-F	3.960
18	CH	O	CH_3	H	C_6H_4-*para*-$N(CH_3)_2$	4.005
19	CH	O	CH_3	H	C_6H_4-*para*-CH_3	3.950
20	CH	O	CH_3	H	C_6H_4-*para*- C_2H_5	3.977
21	CH	O	CH_3	H	C_6H_4-*para*-OCH_3	3.980
22	CH	O	CH_3	H	C_6H_4-*para*-F	3.958
23	CH	O	CH_3	H	C_6H_4-*para*-$NHCOCH_3$	4.027
24	CH	O	CH_3	H	C_6H_4-*para*-$NHCH_3$	3.979
25	CH	O	H	H	C_6H_4-*para*-$N(CH_3)_2$	4.004
26	N	O	H	H	C_6H_4-*para*-CH_3	4.225
27	N	O	H	H	C_6H_4-*para*- C_2H_5	4.253
28	N	O	H	H	C_6H_4-*para*-OCH_3	4.257
29	N	O	H	H	C_6H_4-*para*-OC_2H_5	4.283
30	N	O	H	H	C_6H_4-*para*-NH_2	4.227
31	N	O	H	H	C_6H_4-*para*-NO_2	4.285
32	CH	O	H	H	$CH_2C_6H_5$	4.223
33	CH	O	H	H	$CH_2C_6H_4$-*para*-OCH_3	4.282
34	CH	O	H	H	$CH_2C_6H_4$-*para*-Cl	4.290
35	CH	O	H	H	$CH_2C_6H_4$-*para*-NO_2	4.308
36	CH	O	Cl	H	$CH_2C_6H_5$	4.290
37	CH	O	Cl	H	$CH_2C_6H_4$-*para*-OCH_3	4.340
38	CH	O	Cl	H	$CH_2C_6H_4$-*para*-Br	4.410
39	CH	O	Cl	H	$CH_2C_6H_4$-*para*-NO_2	4.363
40	CH	O	NO_2	H	$CH_2C_6H_5$	4.609
41	CH	O	NO_2	H	$CH_2C_6H_4$-*para*-OCH_3	4.657

Table 3 continued

Index	X	Y	R_1	R_2	R_3	A_{obs}
42	CH	O	NO_2	H	$CH_2C_6H_4$-*para*-Br	4.725
43	CH	O	NO_2	H	$CH_2C_6H_4$-*para*-Cl	4.664
44	CH	O	NO_2	H	$CH_2C_6H_4$-*para*-NO_2	4.680
45	CH	O	CH_3	H	$CH_2OC_6H_5$	3.980
46	CH	O	H	NO_2	$CH_2OC_6H_5$	3.732
47	CH	O	Cl	NO_2	$CH_2OC_6H_5$	3.785
48	CH	O	Cl	NO_2	$CH_2OC_6H_4$-*para*-Cl	3.831
49	CH	O	NO_2	H	$CH_2SC_6H_5$	4.359
50	CH	O	CH_3	H	$CH_2SC_6H_5$	4.009
51	N	O	H	H	$CH_2OC_6H_5$	4.260
52	N	O	H	H	$CH_2OC_6H_4$-*para*-Cl	4.319
53	CH	NH	CH_3	H	$CH_2OC_6H_4$-*para*-Cl	4.037
54	CH	NH	NO_2	H	$CH_2SC_6H_5$	4.358
55	CH	NH	CH_3	H	$CH_2SC_6H_5$	4.009
56	CH	O	$COOCH_3$	H	$CH_2OC_6H_5$	4.054
57	CH	O	$COOCH_3$	H	$CH_2OC_6H_4$-*para*-Cl	4.104
58	CH	NH	$COOCH_3$	H	$CH_2OC_6H_4$-*para*-Cl	4.102
59	CH	NH	$COOCH_3$	H	$CH_2SC_6H_5$	4.076
60	CH	O	NO_2	H	$(CH_2)_2C_6H_5$	4.331
61	N	O	H	H	$(CH_2)_2C_6H_5$	4.253
62	CH	O	NH_2	H	C_6H_4-*para*-Br	4.110
63	CH	O	H	H	$CH_2C_6H_4$-*para*-Br	4.360
64	CH	O	H	H	$CH_2OC_6H_4$-*para*-Cl	4.016
65	CH	NH	NO_2	H	$CH_2OC_6H_5$	4.283
66	CH	NH	H	H	$CH_2OC_6H_4$-*para*-Cl	4.015
67	CH	NH	Cl	H	$CH_2SC_6H_5$	4.041
68	CH	NH	H	H	$(CH_2)_2C_6H_5$	4.078

Analysis 5

Calibration set: 33 molecules in the same Table 4
Dependent property: Log P. The values of dependent property were computed using KowWin algorithm [46] and EPI software [47].
"Possible" WB outliers/WB outlier index: molecule **32**/4.001 → molecule **23**/2.877 → molecule **29**/3.137
"True WB outliers": molecule **32**
"Possible" PRECLAV outliers/PRECLAV outlier index: molecule **32**/3.330 → molecule **26**/2.240 → molecule **30**/2.403
"True" PRECLAV outliers: molecule **32**
Quality of PRECLAV QSPR in presence of all molecules: $r^2 = 0.9757 : F = 621.8$

Table 4 Name, melting point (°K) and Log P values

No.	Molecule name	Melting point	Log P_{KowWin}
1	Propane	85	1.81
2	Hexane	178	3.29
3	Dodecane	264	6.23
4	2-methyl-butane	115	2.72
5	2,4-dimethyl-pentane	154	3.63
6	2,2,5,5-tetra-methyl-hexane	260	5.03
7	4-ethyl-4-propyl-heptane	203	6.12
8	Cyclobutane	182	2.19
9	Cyclohexane	280	3.18
10	Cyclononane	284	4.65
11	Penta-methyl-benzene	327	4.73
12	Ethyl-benzene	178	3.03
13	Hexyl-benzene	212	5.00
14	*para*-diethyl-benzene	230	4.07
15	*iso*-propyl-benzene	178	3.45
16	1,4-di-*iso*-propyl-benzene	256	4.90
17	4-*tert*-butyl-toluene	222	4.45
18	Tetraline	238	3.96
19	Cyclopropyl-phenyl-methane	278	3.83
20	Cyclohexyl-benzene	280	4.81
21	(1-cyclohexyl-ethyl)-benzene	219	5.72
22	1,3-*trans*-butadiene	164	2.03
23	1,5-hexadiene	132	3.02
24	2,7-nonadiyne	277	3.33
25	1,3,5,7-tetramethyl-anthracene	553	6.53
26	2,4,6,8-tetramethyl-azulene	374	5.57
27	5-methyl-crisene	390	6.07
28	Triphenylene	470	5.52
29	Glycerin	291	−1.65
30	Nitro-benzene	279	1.81
31	Oxalic acid	462	−1.74
32	Dimethyl-sulfoxide	292	−1.22
33	Acetamide	354	−1.16

Quality of PRECLAV QSPR in absence of the molecule **32** : $r^2 = 0.9853 : F = 1002.9$

It is difficult to understand why only molecule **32** is identified as outlier, when the dependent property is Log P. Maybe the cause is low value of Log P despite of the absence of OH/NH/COOH chemical groups. The elimination of molecule **32** has a positive effect on the quality of the QSAR equation.

4 Conclusions

The analysis of a large number of calibration sets for the identification of type B "outlier" molecules, using WB algorithm, leads to the following conclusions:

- regardless of the formula used for the identification of "outliers" and regardless of the value for the "outlier index", it is recommended to eliminate only one molecule, more precisely the molecule with the highest calculated "outlier index"; the computations and the elimination procedure should be repeated as long as the maximum value of the "outlier index" is decreasing
- if the elimination of "possible outliers" has a weak effect over the quality of the QSAR equation and the overall quality of the QSAR equation remains low, this means that the calibration set includes more sub-sets (classes) of molecules and each sub-set includes many molecules; in this situation, there are no "outliers" in the classic sense and the classes of molecules must be identified by specific procedures (cluster analysis)
- the set of "outliers" identified depends on the statistic method used and on the set of descriptors used
- the WB procedure is an alternative worthy of consideration, but the computation time is sensibly increased
- if the same set of descriptors is used, it is recommended to compare the results of the classic formula with the WB procedure
- if one wishes to use the ranks of the values instead of the values, then this replacement must be included in all the computation steps, for instance when using Least Square Method; more precisely all the "parametric" formulas and procedures must be replaced with "non-parametric" formulas and procedures

References

1. see Internet site http://www.statsoft.com/textbook/stathome.html?stbasic.html&1
2. V. Barnett and D. Roberts, Communications in statistics. Theory Methods **22**, 2703 (1993)
3. K. Carling, Comput. Stat. & Data Anal. **33**, 249 (2000)
4. M.B. Kremer, R.D. Martin, Comput. Intell. Finan. Eng. (CIFEr) **29**, 212 (1998)
5. Q. Zhou, S. Li, X. Li, W. Wang, Z. Wang, Clin. Chim. Acta **372**, 94 (2006)
6. M.M. Breunig, H. Kriegel, R.T. Ng, J. Sander, Proceedings of the ACM SIGMOD conference, (Dallas, 2000), p. 93
7. M. Ester, H. Kriegel, J. Sander, X. Xu, Proceedings of the 2nd international conference on knowledge discovery and data mining, (1996), p. 226
8. A.G. Steele, B.M. Wood, R.J. Douglas, Metrologia **42**, 32 (2005)
9. E.M. Knorr, R.T. Ng, Proceedings of the 24th international conference on very large data bases, (New York, 1998), p. 392
10. V. Šaltenis, Informatica **15**, 399 (2004)
11. C.R. Moorhead, J. Royal Stat. Soc. (B) **48**, 39 (1986)
12. A.J. Fox, J. Royal Stat. Soc. (B) **34**, 350 (1972)
13. P. Verboon, I.A. van der Lans, Psychometrika **59**, 485 (1994)
14. B.C. Sutradhar, Ind. J. Stat. **57**, 299 (1995)
15. I.L. Ruiz, M.U. Cuadrado, M.A. Gomez-Nieto, Proc. World Acad. Sci. Eng. Technol. **22**, 302 (2007)
16. K. Kodithala, A.J. Hopfinger, E.D. Thompson, M.K. Robinson, Toxicol. Sci. **66**, 336 (2002)
17. H.J. Motulsky, R.E. Brown, BMC Bioinformatics **7**, 123 (2006)
18. H.E. Solberg, A. Lahti, Clin. Chem. **51**, 2326 (2005)

19. F. Hristea, Math. Rep. **54**, 177 (2002)
20. R.P. Verma, C. Hansch, Bioorg. & Med. Chem. **13**, 4597 (2005)
21. L. Tarko, Rev. Chim. (Bucuresti) **59**, 185 (2008)
22. G.H. Schmid, V.M. Csizmadia, P.G. Mezey, I.G. Csizmadia, Can. J. Chem. **54**, 3330 (1976)
23. G. Maggiora, J. Chem. Inf. Model. **46**, 1535 (2006)
24. K.H. Kim, J. Comp-Aid. Mol. Des. **21**, 63 (2007)
25. R.C.A. Martins, G.A. Magaly, R.B. Alencastro, J. Braz. Chem. Soc. **13**, 816 (2002)
26. R.D. Cramer, R.J. Jilek, S. Guessregen, S.J. Clark, B. Wendt, R.D. Clark, J. Med. Chem. **47**, 6777 (2004)
27. J.C. Saeh, P.D. Lyne, B.K. Takasaki, D.A. Cosgrove, J. Chem. Inf. Comput. Sci. **45**, 1122 (2005)
28. E. Furusjo, A. Svenson, M. Rahmberg, M. Andersson, Chemosphere **63**, 99 (2006)
29. M.T.D. Cronin, W. Schultz, J. Mol. Struct. Theochem **622**, 39 (2003)
30. D.A. Konovalov, N. Sim, E. Deconinck, Y.V. Heyden, D. Coomans, J. Chem. Inf. Model. **48**, 370 (2008)
31. Y. Dalin, L. Yizeng, X. Qingsong, Comp. Appl. Chem. **23**, 569 (2006)
32. D.A. Konovalov, L.E. Llewellyn, Y.V. Heyden, D. Coomans, J. Chem. Inf. Model. **48**, 2081 (2008)
33. PCModel v. 9.1 is available from Serena Software, Box 3076, Bloomington, IN, 47402-3076, USA, see Internet site http://www.serenasoft.com/
34. J.J.P. Stewart, J. Mol. Model. **13**, 1173 (2007)
35. Last version of MOPAC is available from Internet site http://www.openmopac.net/
36. L. Tarko, Rev. Chim. (Bucuresti) **56**, 639 (2005)
37. L. Tarko, I. Lupescu, D. Groposila-Constantinescu, Arkivoc **X**, 254 (2005)
38. L. Tarko, C.T. Supuran, Bioorg. & Med. Chem. **15**, 5666 (2007)
39. PRECLAV software is available from Center of Organic Chemistry (CCO), Bucharest – Romanian Academy; managing director pfilip@cco.ro; author ltarko@cco.ro
40. L.H. Hall, T.A. Vaughn, Med. Chem. Res. **7**, 407 (1997)
41. K. Roy, G. Gosh, Int. Electr. J. Mol. Design **2**, 599 (2003)
42. J.M.J. Tronchet, M. Grigorov, N. Dolatshahi, F. Moriaud, J. Weber, Eur. J. Med. Chem. **32**, 279 (1997)
43. O. Ursu, A. Costescu, M.V. Diudea, B. Parv, Croat. Chim. Acta. **79**, 483 (2006)
44. see Internet site http://chembiofinder.cambridgesoft.com/
45. see Internet site http://webbook.nist.gov/chemistry/
46. W.M. Meylan, P.H. Howard, J. Pharm. Sci. **84**, 83 (1995)
47. Estimation Programs Interface (EPI) software is available from Internet site http://www.epa.gov/oppt/exposure/pubs/episuitedl.htm