VERSITA

Cent. Eur. J. Chem. • 10(4) • 2012 • 1042-1048
DOI: 10.2478/s11532-012-0031-4

Central European Journal of **Chemistry**

CORAL: the prediction of biodegradation of organic compounds with optimal SMILES-based descriptors

Short Communication

Andrey A. Toropov*[,1], Alla P. Toropova[1], Anna Lombardo[1], Alessandra Roncaglioni[1], Nicoletta De Brita[2], Giovanni Stella[2], Emilio Benfenati[1]

[1]*Mario Negri Institute for Pharmacological Research, 20156 Milan, Italy*

[2]*Technico-Scientific Center, 20052 Monza, Italy*

Received 24 August 2011; Accepted 16 January 2012

Abstract: CORAL software (http:/www.insilico.eu/coral) has been used to build up quantitative structure–biodegradation relationships (QSPR). The normalized degradation percentage has been used as the measure of biodegradation (for diverse organic compounds, n=445). Six random splits into sub-training, calibration, and test sets were examined. For each split the QSPR one-variable linear regression model based on the SMILES-based optimal descriptors has been built up. The average values of numbers of compounds and the correlation coefficients (r^2) between experimental and calculated biodegradability values of these six models for the test sets are n=88.2±11.7 and r^2=0.728±0.05. These six models were further tested against a set of chemicals (n=285) for which only categorical values (biodegradable or not) were available. Thus we also evaluated the use of the model as a classifier. The average values of the sensitivity, specificity, and accuracy were 0.811±0.019, 0.795±0.024, and 0.803±0.008, respectively.

Keywords: *QSPR • SMILES • Biodegradability • CORAL software*

1. Introduction

Persistent organic pollutants are toxic organic compounds that are persistent in the environment and thus may have a greater risk to accumulate in biological organisms [1-7]. Chemicals having the potential to persist in environmental media, to undergo long-range transport via water and the atmosphere, to accumulate in the tissues of living organisms, and (in some cases) to cause adverse biological effects after long-term exposure, are the focus of national and international risk management measures, due to the special concerns they raise for human health and the environment [8-10]. These compounds are classified as Persistent, Bioaccumulative and Toxic (PBT) or very Persistent and very Bioaccumulative (vPvB) according to the European REACH regulation for chemicals.

A chemical is defined as persistent if it resists degradation processes and is present in the environment for a long time [12]. Persistent (P) and very persistent (vP) refer to chemicals that have degradation half-lives above certain trigger values in surface water, sediment or soil [12]. The triggers are reported in the Annex XIII of REACH. This Annex also explicitly indicates the use of biodegradation QSAR models as screening methods: if they (safely) predict the compound as non-persistent it can be classified as nP (non-persistent), otherwise experimental tests are necessary to evaluate the biodegradability.

In the European Chemicals Agency (ECHA) guidelines, biodegradation is defined as the biologically-mediated degradation or transformation of chemicals carried out by microorganisms. Most of the models generate qualitative predictions (usually, ready vs. non-ready biodegradability) [8-10]. Indeed, the OECD tests for ready biodegradability according to OECD guideline 301 [11] represent the most prominent group of standardized experimental biodegradation screening tests [9,12]. The most common procedure is the OECD 301c test, based on the MITI1 test.

Quantitative structure–activity relationships (QSAR) are a tool for the modern natural sciences [14-16]. QSAR for the prediction of biodegradability have been built up [8,10,17].

* *E-mail: andrey.toropov@marionegri.it*

The aim of the present study is the estimation of CORAL software as a possible tool to model the biodegradation of organic compounds by means of QSARs calculated with the optimal descriptors based on the simplified molecular input line entry system (SMILES).

2. Experimental procedure

2.1. Data

Experimental data on the normalized degradation percentage (NDP) and qualitative data on the NDP were taken from different sources [18-21].

A dataset of 730 compounds with continuous and discrete values of ready biodegradability (*i.e.*, percentage of degradation at 28 days) was obtained extracting data from the OECD toolbox v2.0 and from BioWin v4.10 (data used to build and test the Linear and Non-Linear MITI Biodegradation Model (*i.e.*, Biowin 5 and Biowin 6). OECD toolbox v2.0 contains continuous values, whereas BioWin v4.10 contains discrete values: 0 means that the percentage of biodegradation is below the threshold of 60%, 1 that the threshold is reached (*i.e.*, the substance is ready biodegradable). All the data are obtained performing the MITI-1 test. The compounds in common between the two datasets were checked for agreement: the compounds with values in disagreement were eliminated.

The 445 compounds with continuous values were involved in order to build up the model. Six random splits were prepared on the basis of the 445 compounds. These splits are random, but the range of endpoint for sub-training, calibration, and test sets is approximately equivalent. The additional 285 compounds characterized by discrete values (0 means stable; 1 means biodegradable) were used to test the model. If the predicted biodegradability ≤0.5 then one should expect that a substance is stable; vice versa, if the predicted biodegradability >0.5 then one should expect a substance is biodegradable.

On the one hand, each model calculated with the Monte Carlo method is a random event. On the other hand, each model is a measurement of the statistical characteristics which are obtained by a given approach. Thus, average values of the statistical characteristics for a group split are more informative than the statistical characteristics of the model for solely one split. However, carrying out hundreds of such measurements results in extremely time-intensive calculations. We have estimated six random splits as a reasonable compromise between the reliability of the results and the time of the calculations.

2.2. Optimal SMILES-based descriptor

The structural descriptor used for one-variable models of the biodegradability is calculated as

$$DCW(Threshold, Nepoch) = CW(ATOMPAIR) + CW(BOND) + CW(NOSP) + CW(HALO) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (1)$$

where ATOMPAIR is defined in the following way. We consider nine SMILES elements: F, Cl, Br, N, O, S, P, double bond, triple bond. Then the software checks for the simulataneous presence of two of these SMILES elements. Similarly, the software searches for the occurrence of these bonds in the BOND index: double, triple, or stereochemical bonds, and if they are present at the same time in the molecule. The NOSP index looks specifically for the occurrence of these atoms: N, O, P, S, and if they are present together or not. Finally, the HALO index searches in the molecule the occurrence of halogens: F, Cl, Br, and if they are present simultaneously in the molecule. Table 1 contains an example of ATOMPAIR, BOND, NOSP, HALO, Sk, SSk, and SSSk which are extracted from SMILES. It should be noted that molecular features represented by ATOMPAIR and molecular features represented by BOND, NOSP, and HALO are different: *e.g.* ATOMPAIR = **....N...O...** is an indicator of the presence of nitrogen together with oxygen, whereas NOSP=NOSP11000000 is an indicator of the presence of nitrogen together with oxygen in the *absence* of sulphur and phosphorus.

SMILES is a sequence of symbols which are a representation of a molecular structure. Hence, one can speak about above-mentioned SMILES symbols as about molecular fragments (*e.g.* S_k, SS_k, SSS_k). However, ATOMPAIR, BOND, NOSP and HALO are not fragments: they are descriptors for combinations of different molecular features.

There are symbols which themselves are representations of a molecular feature, *e.g.* 'c', 'C', 'N', etc. There are undivided pairs of symbols which represent a molecular feature, *e.g.* 'Cl', 'Br', '@@', *etc*. We have denoted both these kinds of information as SMILES atoms (S_k). SS_k and SSS_k are combinations of two and three SMILES atoms. *E.g.* if the SMILES is ABCDE, the SS_k are AB, BC, CD, and DE; similarly SSS_k are ABC, BCD, and CDE. In order to avoid situations where the same molecular fragment is represented twice (*i.e.*, AB and BA), the SS_k and SSS_k are ordered according to ASCII codes of symbols. It is to be noted, that for a SMILES attribute that contains four SMILES-atoms $SSSS_k$, it is impossible to define the rule for selection of solely one "correct" possibility that is similar to the above-mentioned AB-BA or ABC-CBA. CW(x)

is the correlation weight for a SMILES attribute x (x = ATOMPAIR, BOND, NOSP, HALO, S_k, SS_k, and SSS_k). Each SMILES attribute for registration and for the Monte Carlo calculations is represented by a sequence of twelve symbols (Table 1). The first four symbols are the first zone; the second four symbols are the second zone; finally, the third four symbols are the third zone (Table 1). All three zones are necessary for the SMILES attributes involving three SMILES atoms (*i.e.*, SSS_k). The SS_k are represented in the first and second zones. The S_k are located in the first zone. Vacant positions in this twelve-symbols representation are indicated by 'x'. CW(x) are calculated with the Monte Carlo method. The classic scheme is to build up a model that is satisfactory for the training set and evaluate that the model is also appropriate for external optimization sets. However, the balance of correlation seems a more realistic approach. This approach, based on the split of the training set into sub-training and calibration sets, is aimed at avoiding overtraining by means of the control of the statistical quality of the model for the calibration set. The balance of correlation is the optimization with target function BC= R+R'-ABS(R-R'), where R and R' are correlation coefficients for the sub-training and calibration sets, respectively. Thus, the calibration set plays the role of a 'preliminary test set'.

The correlation weights of rare molecular features (which are represented by SMILES attributes) are improving the statistical quality only for compounds which are involved in the sub-training or calibration sets (but not for the test set). Thus, the reliable model must be based on molecular fragments which are not rare. For this reason we introduced a threshold which is a tool to select SMILES attributes which are 'not rare'. If the threshold is defined as five, then all SMILES attributes (including ATOMPAIR, BOND, NOSP, HALO, S_k, SS_k, and SSS_k) which take place only in four (or less) SMILES of the training set will be classified as rare. The correlation weights for "rare" attributes will be defined as zero. Table 1 contains an example of the calculation of the optimal descriptor (Split 1).

The *Supplementary materials* section contains SMILES, numerical (n=445), qualitative (n=285) data for studied substances, and correlation weights used for calculating DCW(2,11) in Eq. 2.

3. Results and discussion

Table 2 contains the statistical quality of the model obtained for cases with the threshold from 0-3 and the number of epochs of the Monte Carlo optimization for the correlation weights N_{epoch}=30.

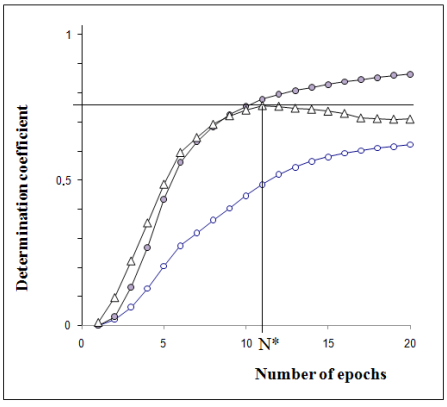

Figure 1. The co-evolution of correlations during 20 epochs of the Monte Carlo method optimization. N* is the value of the epochs which gives the maximumcorrelation coefficient for the external test set.

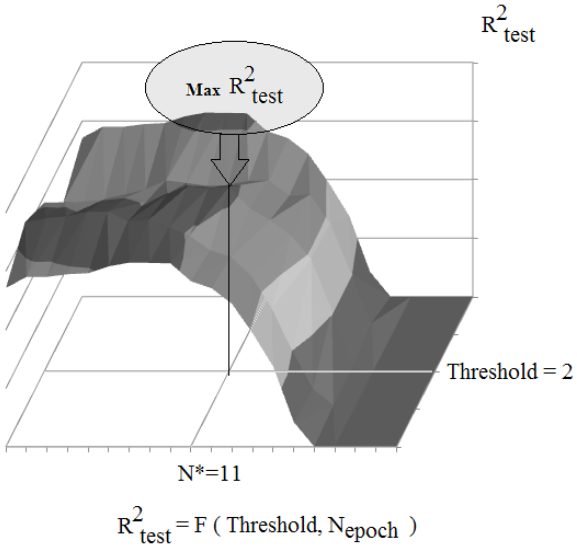

Figure 2. The correlation coefficient between experimental and calculated values of an endpoint for the external test set is a mathematical function of the threshold and the N_{epoch}. The N˚ is the number of epochs which gives the maximum correlation coefficient for the external test set.

Fig. 1 shows the representation of a co-evolution of correlations for the sub-training, calibration, and test sets for split 1. The preferable threshold for split 1 is 2 and the preferable number of epochs of the Monte Carlo optimization (N*) is 11 (Fig. 2). The preferable N* and T* give the maximum correlation coefficient between the experimental and calculated value of an endpoint for the test set. In fact, the correlation coefficient between experimental and predicted biodegradability is a mathematical function of the threshold and the number of epochs of the Monte Carlo optimization. Fig. 2 shows the scheme for definition of N* and T*. From Table 2, one can see that the N* values are 11,

Table 1. Example of the registration of SMILES attributes and the calculation of optimal descriptors for SMILES="N#CC(C)(C)O"; DCW= 23.4562619; NDP= 0.7623872 (Eq. 2).

SMILES attribute (SA)	Correlation Weight
S_k	
Nxxxxxxxxxx	0.1278310
#xxxxxxxxxx	-3.8165450
Cxxxxxxxxxx	-0.8716681
Cxxxxxxxxxx	-0.8716681
(xxxxxxxxxx	-1.2502889
Cxxxxxxxxxx	-0.8716681
(xxxxxxxxxx	-1.2502889
(xxxxxxxxxx	-1.2502889
Cxxxxxxxxxx	-0.8716681
(xxxxxxxxxx	-1.2502889
Oxxxxxxxxxx	0.7492272
SS_k	
Nxxx#xxxxxxx	-3.8706420
Cxxx#xxxxxxx	-4.5024889
CxxxCxxxxxxx	-0.6221895
Cxxx(xxxxxxx	0.2491393
Cxxx(xxxxxxx	0.2491393
Cxxx(xxxxxxx	0.2491393
(xxx(xxxxxxx	-1.6281275
Cxxx(xxxxxxx	0.2491393
Cxxx(xxxxxxx	0.2491393
Oxxx(xxxxxxx	-2.2514967
Nxxx#xxxCxxx	-4.4966284
SSS_k	
CxxxCxxx#xxx	1.5037750
CxxxCxxx(xxx	-0.2453117
Cxxx(xxxCxxx	1.2474111
(xxxCxxx(xxx	-0.0013274
Cxxx(xxx(xxx	1.4998647
Cxxx(xxx(xxx	1.4998647
(xxxCxxx(xxx	-0.0013274
Oxxx(xxxCxxx	2.0037458
NOSP	
NOSP11000000	-1.4985284
HALO	
HALO00000000	9.8771395
BOND	
BOND01000000	12.0024523
ATOMPAIR	
....N...O...	2.9950425
....O...B3..	10.4998562
....N...B3..	9.6267963

14, 19, 20, 20, and 15 for splits 1-6, respectively. The preferable threshold (T*) is 2 for all splits.

The statistical characteristics of the models for splits 1-6 calculated with threshold 2 (i.e. T*=2) and above-mentioned N* are the following:

Split 1

$$NDP = 0.1924(\pm 0.0011) + 0.0243(\pm 0.0001) * DCW(2,11) \quad (2)$$

n=236, r^2=0.5199, q^2=0.5133, s=0.276, F=253 (sub-training set)

n=124, r^2=0.7909, r^2_{pred}=0.7859, s=0.191, F=462 (calibration set)

n=85, r^2=0.7548, r^2_{pred}=0.7426, R^2_m=0.6573, s=0.211, F=256 (test set)

Split 2

$$NDP = 0.0519(\pm 0.0011) + 0.0337(\pm 0.0001) * DCW(2,14) \quad (3)$$

n=237, r^2=0.5920, q^2=0.5864, s=0.258, F=341 (sub-training set)

n=132, r^2=0.8336, r^2_{pred}=0.8295, s=0.183, F=651 (calibration set)

n=76, r^2=0.6338, r^2_{pred}=0.6167, R^2_m=0.5663, s=0.247, F=128 (test set)

Split 3

$$NDP = 0.0146(\pm 0.0010) + 0.0342(\pm 0.0001) * DCW(2,19) \quad (4)$$

n=265, r^2=0.6030, q^2=0.5984, s=0.249, F=399 (sub-training set)

n=104, r^2=0.8822, r^2_{pred}=0.8783, s=0.149, F=764 (calibration set)

n=76, r^2=0.7730, r^2_{pred}=0.7609, R^2_m=0.7595, s=0.203, F=252 (test set)

Split 4

$$NDP = 0.0002(\pm 0.0012) + 0.0353(\pm 0.0001) * DCW(2,20) \quad (5)$$

n=214, r^2=0.5683, q^2=0.5624, s=0.262, F=279 (sub-training set)

n=127, r^2=0.8248, r^2_{pred}=0.8200, s=0.177, F=588 (calibration set)

n=104, r^2=0.7120, r^2_{pred}=0.7032, R^2_m=0.7015, s=0.221, F=252 (test set)

Split 5

$$NDP = 0.0383(\pm 0.0014) + 0.0286(\pm 0.0001) * DCW(2,20) \quad (6)$$

n=208, r^2=0.5538, q^2=0.5472, s=0.273, F=256 (sub-training set)

n=133, r^2=0.8122, r^2_{pred}=0.8073, s=0.204, F=567 (calibration set)

Table 2. Search for most informative threshold T* and the number of epochs of the Monte Carlo optimization N* for splits 1-6: T* and N* are values which produce the maximum correlation coefficient between experimental and calculated endpoint values for the test set.

Split	Threshold	Probe 1	Probe 2	Probe 3	Average	Dispersion
1	$R^2_{(test)}$					
	0	0.7709	0.7610	0.7542	0.7620	0.0068
	1	0.7564	0.7484	0.7497	0.7515	0.0035
	2	0.7707	0.7613	0.7557	**0.7626**	0.0062
	3	0.7567	0.7343	0.7446	0.7452	0.0091
	N*					
	0	10	12	11	11.00	0.82
	1	11	9	11	10.33	0.94
	2	12	10	12	**11.33**	0.94
	3	10	11	12	11.00	0.82
2	$R^2_{(test)}$					
	0	0.6222	0.6325	0.6095	0.6214	0.0094
	1	0.6149	0.6282	0.6218	0.6216	0.0055
	2	0.6187	0.6414	0.6151	**0.6250**	0.0116
	3	0.6178	0.6246	0.6216	0.6214	0.0028
	N*					
	0	11	11	12	11.33	0.47
	1	10	12	11	11.00	0.82
	2	12	14	15	**13.67**	1.25
	3	10	12	12	11.33	0.94
3	$R^2_{(test)}$					
	0	0.7791	0.7481	0.7605	0.7626	0.0127
	1	0.7486	0.7709	0.7503	0.7566	0.0101
	2	0.7748	0.7751	0.7448	**0.7649**	0.0142
	3	0.6987	0.7080	0.7192	0.7086	0.0084
	N*					
	0	20	19	19	19.33	0.47
	1	19	20	17	18.67	1.25
	2	18	19	19	**18.67**	0.47
	3	16	17	20	17.67	1.70
4	$R^2_{(test)}$					
	0	0.6295	0.6341	0.6475	0.6370	0.0076
	1	0.6204	0.6565	0.6463	0.6410	0.0152
	2	0.7404	0.7354	0.7062	**0.7273**	0.0151
	3	0.7330	0.7148	0.7303	0.7260	0.0080
	N*					
	0	17	9	16	14.00	3.56
	1	7	9	8	8.00	0.82
	2	15	16	12	**14.33**	1.70
	3	13	14	14	13.67	0.47
5	$R^2_{(test)}$					
	0	0.7216	0.7048	0.7034	0.7099	0.0083
	1	0.7140	0.7361	0.7232	0.7244	0.0091
	2	0.7315	0.7287	0.7242	**0.7281**	0.0030
	3	0.7391	0.7285	0.7107	0.7261	0.0117
	N*					
	0	15	17	19	17.00	1.63
	1	14	18	16	16.00	1.63
	2	20	18	20	**19.33**	0.94
	3	20	19	18	19.00	0.82
6	$R^2_{(test)}$					
	0	0.7037	0.7120	0.7284	0.7147	0.0103
	1	0.7156	0.7119	0.7202	0.7159	0.0034
	2	0.7664	0.7979	0.8022	**0.7888**	0.0160
	3	0.7527	0.7201	0.7168	0.7299	0.0162
	N*					
	0	9	9	10	9.33	0.47
	1	12	9	10	10.33	1.25
	2	15	15	13	**14.33**	0.94
	3	15	13	11	13.00	1.63

n=104, r^2=0.7208, r^2_{pred}=0.7104, R^2_m=0.6496, s=0.213, F=263 (test set)

Split 6

$$NDP = 0.0228(\pm 0.0013) + 0.0329(\pm 0.0001) * DCW(2,15) \quad (7)$$

n=225, r^2=0.5689, q^2=0.5635, s=0.270, F=294 (sub-training set)

n=136, r^2=0.8330, r^2_{pred}=0.8290, s=0.184, F=669 (calibration set)

n=84, r^2=0.7978, r^2_{pred}=0.7897, R^2_m=0.7696, s=0.191, F=324 (test set)

As a further check on the model performance, we took the data for which we had only values reported as classes: degradable or not. These data could not be used to build up the model, because the CORAL model is a regression model, and thus continuous values are needed. Thus, this exercise can be considered as both an example of the use of the model as a classifier and as an external validation study. The predictive potential of the models calculated with Eqs. 2-7 has been checked with qualitative data on the biodegradability (n=286). Table 3 contains sensitivity, specificity and accuracy values for the validation set.

One can see (Eqs. 2-7 and Table 3) that the statistical quality of the described models varies for different splits. For random splits #1, #3, and #6 (Eqs. 2, 4, and 7) R^2_{test} ≥0.75, whereas for split 2 (Eq. 3) R^2_{test} =0.63. We submit that one should expect that correlation coefficients for different splits will be different. The range of correlation coefficients for a group of random splits is important information: the average value of the correlation coefficient for the test sets is 0.728±0.05 (Eqs. 2-7). Thus, the proposed approach can be used to model the biodegradability.

Probably, at present, a QSAR study on NDP is absent, but the prediction of rate constants for radical degradation of aromatic pollutants is a task similar to the prediction of the biodegradability [22]. The four-variable linear regression model based on DRAGON descriptors together with quantum mechanics parameters is statistically characterized by the following n=60, r^2=0.735, s=0.174 (training set) and n=18, r^2=0.760, s=0.200 (test set). Thus, the predictability of one-variable regression models calculated with Eqs. 2-7 can be estimated as equivalent to the predictability of the above-mentioned model. In other words, predictions which have been obtained by the approach described above are reasonably good.

Table 3. The testing of models (splits 1-6) against external qualitative data on biodegradability (n=285). True positive means a prediction where the biodegradability is modeled correctly and classified a chemical as biodegradable, analogical definitions for "true negative", "false positive", *etc.*

Split 1
The Number of True_Positive = 121 42.46% The Number of True_Negative = 110 38.60% The Number of False_Positive = 20 7.02% The Number of False_Negative = 34 11.93% Sensitivity* = 0.781 Specificity = 0.846 Accuracy = 0.811
Split 2
The Number of True_Positive = 129 45.26% The Number of True_Negative = 103 36.14% The Number of False_Positive = 27 9.47% The Number of False_Negative = 26 9.12% Sensitivity = 0.832 Specificity = 0.792 Accuracy = 0.814
Split 3
The Number of True_Positive = 129 45.26% The Number of True_Negative = 101 35.44% The Number of False_Positive = 29 10.18% The Number of False_Negative = 26 9.12% Sensitivity = 0.832 Specificity = 0.777 Accuracy = 0.807
Split 4
The Number of True_Positive = 125 43.86% The Number of True_Negative = 101 35.44% The Number of False_Positive = 29 10.18% The Number of False_Negative = 30 10.53% Sensitivity = 0.806 Specificity = 0.777 Accuracy = 0.793
Split 5
The Number of True_Positive = 123 43.16% The Number of True_Negative = 104 36.49% The Number of False_Positive = 26 9.12% The Number of False_Negative = 32 11.23% Sensitivity = 0.794 Specificity = 0.800 Accuracy = 0.796
Split 6
The Number of True_Positive = 127 44.56% The Number of True_Negative = 101 35.44% The Number of False_Positive = 29 10.18% The Number of False_Negative = 28 9.82% Sensitivity = 0.819 Specificity = 0.777 Accuracy = 0.800

*) $Sensitivity = \frac{True\_Positive}{True\_Positive + False\_Negative}$

$Specificity = \frac{True\_Negative}{True\_Negative + False\_Positive}$

$Accuracy = \frac{True\_Positive + True\_Negative}{True\_Positive + False\_Positive + True\_Negative + False\_Negative}$

Predicting the biodegradability of chemicals can be done with special rules [10] based on the presence of various molecular features, *e.g.* presence of C-O, C=C,

Cl, aromatic groups, various cycles, etc. The accuracy of the model for biodegradability described in the literature [10] is about 77-89% (350-400 compounds). In fact the accuracy of the models calculated with Eqs. 2-7 (Table 3) is similar: one can see (Table 3), that for all six models this criterion is about 0.80, consequently, the number of false positive and false negative predictions for all splits is about 53-55 compounds, *i.e.*, 19-20%. Hence, the numerical values of these criteria for models (classification into two classes: biodegradable or not biodegradable) which are calculated with Eqs. 2-7 are quite good.

There are several software packages available for the generation of SMILES notations (http://depth-first.com/articles/2007/04/03/creating-canonical-smiles-with-ruby-open-babel/). We have used the canonical version of SMILES notations (ACD/ChemSketch Freeware, v. 11.00, Inc., Toronto, Canada, www.acdlabs.com, 2007). However, we deem that selected global SMILES attributes (i.e. ATOMPAIR, BOND, NOSP, and HALO) most probably will be the same for the majority (maybe for all) versions of SMILES. Unfortunately, local SMILES attributes considerably depended upon the difference in various SMILES formats. Consequently, for a robust model, one should use the same SMILES for all compounds. The best way is to use canonic SMILES.

4. Conclusions

1. The CORAL software gives reasonable models for the biodegradability of organic compounds; 2. The model can be used both to predict continuous values, and, from these values, chemicals can be categorized as persistent or not; 3. The suggested modeling process for biodegradability is based on the representation of the molecular structure by SMILES and on the experimental data.

Acknowledgement

The authors express their gratitude to ANTARES (project number LIFE08-ENV/IT/00435) for financial support, and to Dr. L. Cappellini, Dr. G. Bianchi and Dr. R. Bagnati for valuable consultations on the computer use.

References

[1] A. Sabljic, Chemosphere 43, 363 (2001)
[2] P. Gramatica, F. Consolaro, S. Pozzi, Chemosphere 43, 655 (2001)
[3] V. Uddameri, M. Kuchanur, Chemosphere 54,771 (2004)
[4] S. Dimitrov, D. Nedelcheva, N. Dimitrova, O. Mekenyan, Sci. Tot. Environ. 408, 3811 (2010)
[5] E. Papa, P. Gramatica, J. Mol. Graph. Model. 27,59 (2008)
[6] A. Mostrąg, T. Puzyn, M. Haranczyk, Environ. Sci. Pollut. Res. 17,470 (2010)
[7] T. Puzyn, M. Haranczyk, N. Suzuki, T. Sakurai, Mol Divers 15, 173 (2011)
[8] M. Pavan, A.P. Worth, QSAR Comb. Sci. 27,32 (2007)
[9] A. Sabljic, W. Peijnenburg, Pure Appl. Chem. 73, 1331 (2001)
[10] J.R. Baker, D. Gamberger, J.R. Mihelcic, A. Sabljic, Molecules 9, 989 (2004)
[11] OECD 301c: Ready Biodegradability - MODIFIED MITI TEST (I)
[12] T. Junker, C. Paatzsch, T. Knacker, Sci. Tot. Environ. 408, 3803 (2010)
[13] J. Hermens, S. Balaz, J. Damborsky, W. Karcher, M. Müller, W. Peijnenburg, A. Sabljic, M. Sjöström, SAR QSAR Environ. Res. 3, 223 (1995)
[14] M. Karelson; V.S. Lobanov, A.R. Katritzky, Chem. Rev. 96, 1027 (1996)
[15] P.K. Ojha, I. Mitra, R.N. Das, K. Roy, Chemometr. Intel. Lab. 107, 194 (2011)
[16] L.M.A. Mullen, P.R. Duchowicz, E.A. Castro, Chemometr. Intel. Lab. 107, 269 (2011)
[17] R.S. Boethling, A. Sabljić, Environ. Sci. Technol. 23, 672 (1989)
[18] OECD toolbox v2.0: http://www.oecd.org/document/54/0,3746,en_2649_34379_42923638_1_1_1_1,00.html
[19] EPISuite v4.1: http://www.epa.gov/opptintr/exposure/pubs/episuite.htm
[20] J. Tunkel, P.H. Howard, R.S. Boethling, W. Stiteler, H. Loonen, Environ. Toxicol. Chem. 19, 2478 (2000)
[21] L. Han, Y. Wang, S.R. Bryant, Bioinformatics 9, 401 (2008)
[22] H. Kusuc, B. Rasulev, D. Lesczcynska, J. Leszczynski, N. Koprivanac, Chemosphere 75, 1128, (2009)