European Journal of Medicinal Chemistry 46 (2011) 1400–1403

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Short communication

CORAL: Building up the model for bioconcentration factor and defining it's applicability domain

A.A. Toropov [a,*], A.P. Toropova [a], A. Lombardo [a], A. Roncaglioni [a], E. Benfenati [a], G. Gini [b]

[a] Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
[b] Department of Electronics and Information, Politecnico di Milano, Via Ponzio 34/5, 20133 Milano, Italy

ARTICLE INFO

Article history:
Received 25 October 2010
Received in revised form
7 January 2011
Accepted 12 January 2011
Available online 21 January 2011

Keywords:
QSPR
SMILES
Bioconcentration factor
Optimal descriptor
Co-evolution of correlations

ABSTRACT

CORAL (**COR**relation **A**nd **L**ogic) software can be used to build up the quantitative structure – property/activity relationships (QSPR/QSAR) with optimal descriptors calculated with the simplified molecular input line entry system (SMILES). We used CORAL to evaluate the applicability domain of the QSAR models, taking a model of bioconcentration factor (log BCF) as example. This model's based on a large training set of more than 1000 chemicals. To improve the model is predictivity and reliability on new compounds, we introduced a new function, which uses the Delta(obs) = log BCF(expr) – log BCF(calc) of the predictions on the chemicals in the training set. With this approach, outliers are eliminated from the phase of training. This proved useful and increased the model's predictivity.

1. Introduction

The bioconcentration factor (BCF) is useful to characterize the environmental behavior of a chemical, particularly to see whether it has an accumulative effect. BCF defines the ratio between the concentration in the organism and the medium. This is an important characteristic from a regulatory point of view, since it is used in the GHS and REACH [1].

Besides the experimental model, which uses more than one hundred fish, takes at least one month and costs several tens of thousands of euros for each substance, quantitative structure – property relationships (QSPR) have been used to model this endpoint [2–12]. Thus, like in many other cases, developing the computer models for predicting the BCF of chemicals is motivated by the fact that the experimental measurements are time-consuming, expensive, and not feasible for the many thousands of chemicals of potential regulatory interest [13].

The aim of the present study is to build up a QSPR model for log BCF and to define its applicability domain. The definition is based on the QSPR model of the endpoint which is calculated as Delta = log BCF(obs) – log BCF(calc). This can be useful to classify as potential outliers the substances of the external test set which have large Delta values.

2. Method

2.1. Data

The experimental data for the log BCF were taken from Ref. [14], but for six compounds the log BCF were recalculated taking into account additional experimental data on the log BCF (CAS 361377-29-9, 892-20-6, 25155-30-0, 535-77-3, 71751-41-2, and 119446-68-3) and one compound was removed because it was very large molecule (CAS 71751-41-2). Three random splits A, B, and C (with approximately 50% of substances in the sub-training set, 30% in the calibration set, and 20% in the test set) were examined. In total were examined 1035 substances.

2.2. Optimal descriptors

SMILES is a representation of the molecular structure. One can calculate with SMILES a molecular descriptor similarly to the well-known descriptors calculated with molecular graphs. SMILES-based optimal descriptors for QSPR modeling of log BCF and the property Delta = log BCF(obs) – log BCF(calc) were calculated respectively as

* Corresponding author.
E-mail address: andrey.toropov@marionegri.it (A.A. Toropov).

0223-5234/$ – see front matter © 2011 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2011.01.018

Table 1
Calculation of the NOSP index.

N	O	S	P	Comments
0	0	0	0	Nitrogen, oxygen, sulphur, and phosphorus are absent
0	0	0	1	The molecule only contains phosphorus
0	0	1	0	The molecule only contains sulphur
0	0	1	1	The molecule contains sulphur and phosphorus
0	1	0	0	The molecule only contains oxygen
0	1	0	1	The molecule contains oxygen and phosphorus
0	1	1	0	The molecule contains oxygen and sulphur
0	1	1	1	The molecule contains oxygen, sulphur, and phosphorus
1	0	0	0	The molecule only contains nitrogen
1	0	0	1	The molecule contains nitrogen and phosphorus
1	0	1	0	The molecule contains nitrogen and sulphur
1	0	1	1	The molecule contains nitrogen, sulphur, and phosphorus
1	1	0	0	The molecule contains nitrogen and oxygen
1	1	0	1	The molecule contains nitrogen, oxygen and phosphorus
1	1	1	0	The molecule contains nitrogen, oxygen, and sulphur
1	1	1	1	The molecule contains nitrogen, oxygen, sulphur, and phosphorus

$$DCW(T) = CW(NOSP) + CW(HALO) + CW(BOND) + \sum CW(S_k) + \sum CW(SS_k), \quad (1)$$

$$DCW(T) = CW(NOSP) + CW(HALO) + CW(BOND) + \sum CW(S_k) + \sum CW(SS_k) + \sum CW(SSS_k) \quad (2)$$

T is the threshold. If the number of SMILES in the sub-training set which contain an attribute, A is less than T, then CW(A) = 0. Thus the T influences the modeling process. CW is the correlation weight used for modeling. NOSP index and HALO index are descriptors which are mathematical functions of the presence in molecules of combination of chemical elements (Tables 1 and 2). BOND index is the mathematical function of the presence of different chemical bonds (Table 3). S_k, SS_k, SSS_k are one-, two-, and three-element SMILES attributes [14].

Eq. (1) has been used to build up BCF models and more details can be found in literature [14]. Briefly, the optimal descriptors (Eqs. (1) and (2)) are mathematical functions of the following parameters: (a) SMILES; (b) threshold; and (c) the number of epochs (N_{epoch}) of the optimization. The numerical data for these parameters were selected empirically (Table 4). The statistical characteristics of the log BCF model for the test set were used as the criteria for this selection.

Models of log BCF were built up with descriptors calculated with Eq. (1). Models of Delta(obs) were built up with descriptors calculated with Eq. (2).

For three random splits the following steps have been done.

1. Selection of the preferable threshold, T, and the number of iterations for the Monte Carlo optimization, N_{epoch}, (Table 4);
2. Building up the general log BCF model, i.e., experiment 1 (Fig. 1);
3. Calculation of the "observed" Delta(obs) = log BCF(expr) − log BCF(calc);
4. Building up the Delta(obs) model as a mathematical function of the molecular structure represented by SMILES, i.e., Delta (calc) = F(SMILES);
5. Calculation of the outliers as structures with Delta(calc) without range $(\overline{d} - d,\ \overline{d} + d)$ according to scheme represented in Fig. 2;
6. Experiment 2 (Fig. 1);
7. Experiment 3 (Fig. 1).

Table 2
Calculation of the HALO index.

F	Cl	Br	Comments
0	0	0	Fluorine, chlorine and bromine are absent
0	0	1	The molecule only contains bromine
0	1	0	The molecule only contains chlorine
0	1	1	Molecule contains chlorine and bromine
1	0	0	The molecule only contains fluorine
1	0	1	The molecule contains fluorine and bromine
1	1	0	The molecule contains fluorine and chlorine
1	1	1	The molecule contains fluorine, chlorine, and bromine

Table 3
Calculation of the BOND index.

=	#	@	Comments
0	0	0	There are no double, triple, or stereo chemical bonds
0	0	1	The molecule only contains stereo chemical bonds
0	1	0	The molecule only contains triple bonds
0	1	1	The molecule contains triple and stereo chemical bonds
1	0	0	The molecule only contains double bonds
1	0	1	The molecule contains double bonds and stereo chemical bonds
1	1	0	The molecule contains double and triple bonds
1	1	1	The molecule contains double, triple, and stereo chemical bonds

Table 4
Definition of preferable threshold and N_{epoch}.

Split	Threshold	N_{epoch}	$r^2_{(test)}$
A	1	20 ± 1	0.796 ± 0.001
B	1	8 ± 1	0.733 ± 0.001
C	3	10 ± 1	0.751 ± 0.001

3. Results and discussion

Fig. 3 shows co-evolutions of correlations between the DCW(T) and log BCF for the sub-training, calibration, and test sets, for splits A, B, and C. We used 35 epochs of the Monte Carlo optimization which involved two phases. In the first phase the correlation coefficient between DCW(T) and log BCF increases for the sub-training, calibration, and test sets. In the second phase the correlation coefficient increases for the sub-training and calibration sets, but decreases for the test set. Thus, the range of transition from the first to second phase is an indicator of the model with the maximum predictive potential.

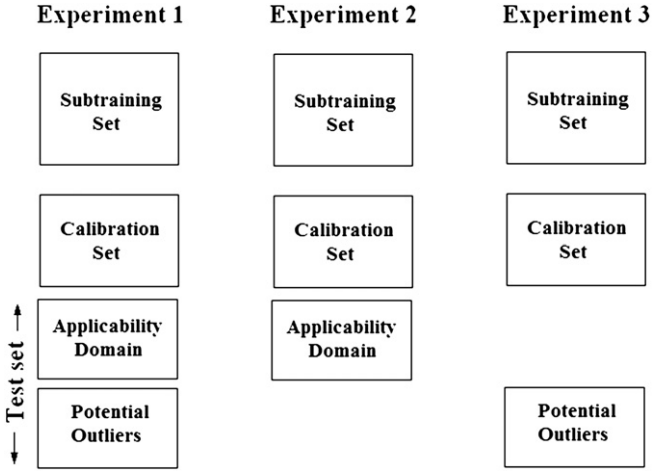

Fig. 1. The schemes of experiments 1, 2, and 3.

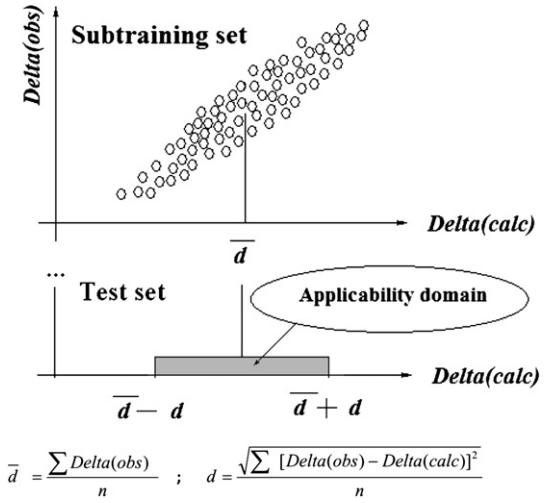

Fig. 2. The scheme of calculation of the domain of applicability (DA) for the log BCF model. Table 5 gives the numerical data on the $\overline{d}$ and d for splits A, B, and C.

The correlation coefficient between the experimental log BCF and calculated log BCF is a mathematical function of the threshold and N_{epoch}. Analysis of the surface for the mathematical function $r^2_{(test)} = F(Threshold, N_{epoch})$ shows that there is a maximum of the $r^2_{(test)}$ for splits A, B, and C. Thus, one can use the surface to define the preferable threshold and the number of epochs for the Monte Carlo optimization (Table 4).

The majority of substances have a typical ('average') behavior and are the basis for building up the log BCF model. However, there are substances with atypical behavior in both the sub-training and calibration sets (Fig. 4). During *the first phase* of the Monte Carlo optimization the main contribution for building up the model comes from information about the substances with 'average' behavior. When the real information contained in these runs out, overtraining starts. The essence of overtraining is a modification of the correlation weights of available attributes for improving only the model *for the sub-training set*. Unfortunately, that reduces the predictive potential of the model for the external test set. However, the preferable N_{epoch} can be selected by analyzing the co-evolutions of correlations (Fig. 3), and the function $r^2_{(test)} = F(Threshold, N_{epoch})$ serves to select both the preferable N_{epoch} and the preferable threshold.

Table 5 illustrates the statistical quality of the log BCF models using Eq. (2) for experiments 1, 2, and 3 with splits A, B, and C. The statistical quality of the model for the substances selected according to rule:

$$\text{Delta(calc)} \in (\overline{d} - d,\ \overline{d} + d) \tag{3}$$

is best for all three splits.

The model for log BCF (split A, experiment 2) is the following

$$\log \text{BCF} = 0.0037(\pm 0.0037) + 0.0922(\pm 0.0001)*\text{DCW}(1) \tag{4}$$

$n = 502$, $r^2 = 0.5537$, $Q^2_{LOO} = 0.5500$, RMSE = 0.897, F = 620 (sub-training set);
$n = 322$, $r^2 = 0.7780$, $R^2_{pred} = 0.7751$, RMSE = 0.627, F = 1122 (calibration set);
$n = 165$, $r^2 = 0.8277$, $R^2_{pred} = 0.8241$, RMSE = 0.545, F = 783;
$k = 1.0321$; $k' = 0.9206$; $R^2_m = 0.795$ (test set)

where

$$Q^2_{LOO} = 1 - \frac{\sum \left[Y_{pred} - Y\right]^2}{\sum [Y - \overline{Y}(\text{sub} - \text{training})]^2} \quad \left(Y \text{ and } Y_{pred} \text{ on sub} - \text{training set}\right) \tag{5}$$

$$R^2_{pred} = 1 - \frac{\sum \left[Y_{pred} - Y\right]^2}{\sum [Y - \overline{Y}(\text{sub} - \text{training})]^2} \quad \left(Y \text{ and } Y_{pred} \text{ on calibration or test set}\right) \tag{6}$$

Y and Y_{pred} are experimental and predicted values of the log BCF, respectively; $\overline{Y}(\text{sub} - \text{training})$ is an average of the experimental values of the log BCF over the sub-training set.

These values above indicate that the model is good, judging by criteria indicated in the literature: slopes k and k′ should be in the range 0.85–1.15 [15], and R^2_m should be larger than 0.5 [16]. Fig. 4 shows the model calculated with Eq. (4).

In all attempts to define the applicability domain with Eq. (3) (i.e., for splits A, B, and C; in experiments 1, 2, and 3) the statistical quality was better for substances classified within the applicability domain than those classified as outliers. But unfortunately some substances which are not outliers were classified as outliers. However, even under those circumstances, this approach can be useful for QSAR analysis. Software and data for described

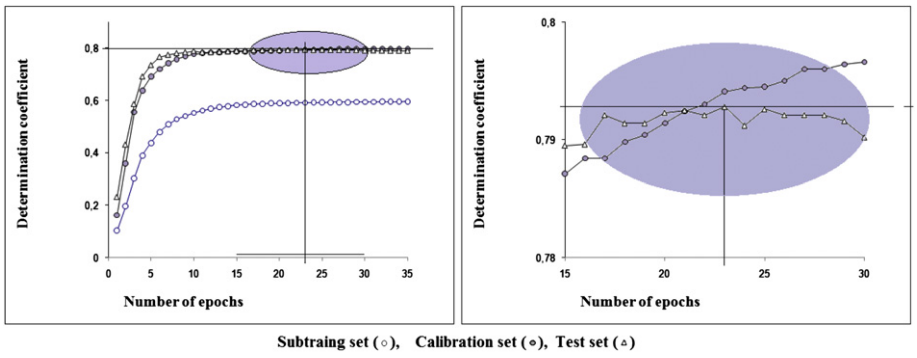

Fig. 3. Co-evolution of correlations between experimental log BCF and log BCF calculated for the sub-training, calibration, and test sets. Each version of the DCW(T) calculated with Eq. (1) or (2), has N_{epoch} that produces a maximum of the determination coefficient for test set. The experiment 1 for split A is shown here.

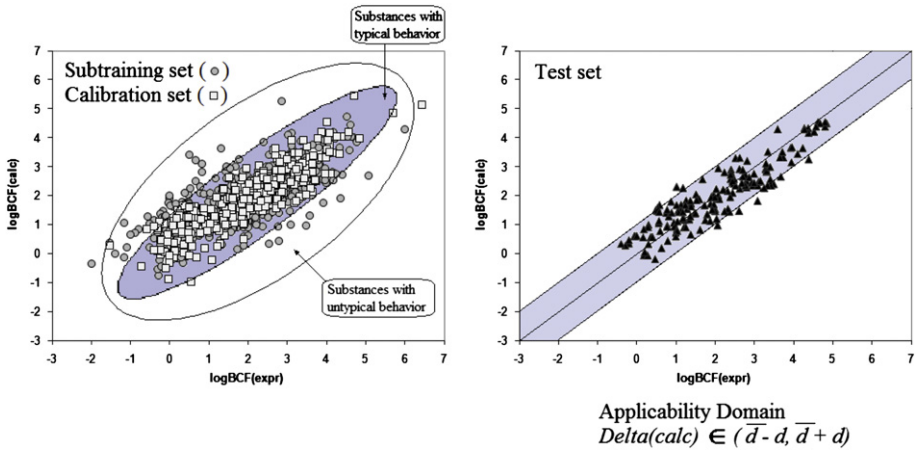

Fig. 4. Experimental log BCF and log BCF calculated with Eq. (4) (split A, experiment 2).

Table 5
Statistical characteristics of the models of the bioconcentration factor for three random splits. The best models are indicated by bold.

Set	n	r^2	Q^2_{LOO} or R^2_{pred}	RMSE	F
Split A					
Experiment 1, $\overline{d}$ = 0.001, d = 0.638 (Fig. 2)					
Training	502	0.5916	0.5885	0.858	724
Calibration	322	0.7950	0.7924	0.602	1241
Test	211	0.7912	0.7875	0.580	792
Experiment 2					
Training	502	0.5537	0.5500	0.897	620
Calibration	322	0.7780	0.7751	0.627	1122
Test	165	**0.8277**	**0.8241**	**0.545**	783
Experiment 3					
Training	502	0.5628	0.5593	0.888	644
Calibration	322	0.7757	0.7727	0.630	1106
Test	46	0.5943	0.5579	0.772	64
Split B					
Experiment 1, $\overline{d}$ = 0.001, d = 0.640 (Fig. 2)					
Training	484	0.5976	0.5943	0.915	716
Calibration	343	0.7384	0.7353	0.629	962
Test	208	0.7126	0.7077	0.652	511
Experiment 2					
Training	484	0.5892	0.5857	0.924	691
Calibration	343	0.7340	0.7308	0.632	941
Test	151	**0.7439**	**0.7380**	**0.628**	433
Experiment 3					
Training	484	0.5978	0.5945	0.915	716
Calibration	343	0.7468	0.7437	0.619	1006
Test	57	0.5999	0.5742	0.706	82
Split C					
Experiment 1, $\overline{d}$ = 0.000, d = 0.636 (Fig. 2)					
Training	451	0.5419	0.5374	0.956	531
Calibration	367	0.7325	0.7296	0.671	999
Test	217	0.7336	0.7291	0.643	592
Experiment 2					
Training	451	0.5433	0.5388	0.954	534
Calibration	367	0.7329	0.7301	0.670	1002
Test	150	**0.7617**	**0.7560**	**0.620**	473
Experiment 3					
Training	451	0.5360	0.5313	0.962	519
Calibration	367	0.7366	0.7338	0.668	1021
Test	67	0.6735	0.6546	0.686	134

computational experiments are available on the Internet at http://www.insilico.eu/coral.

4. Conclusions

We have introduced a new function to optimize QSPR models avoiding the use of chemicals characterized by poor predictions. This scheme, presented in Fig. 2, gave for all splits a robust applicability domain of the model (Table 5). Predictive ability of QSPR model for log BCF obtained in this study is better than BCF model that has been previously reported in literature [14].

Acknowledgements

The authors express their gratitude to OSIRIS for financial support and to Dr. L. Cappellini (*Istituto di Ricerche Farmacologiche Mario Negri, Milano*) for technical assistance.

References

[1] REACH.http://www.reachlegislation.com/index.php.
[2] Y. Zhang, X. Huang, X. Jiang, S. Huang, J. Theor. Comput. Chem. 8 (2009) 783–798.
[3] A.A. Toropov, A.P. Toropova, E. Benfenati, Eur. J. Med. Chem. 44 (2009) 2544–2551.
[4] C. Zhao, E. Boriani, A. Chana, A. Roncaglioni, E. Benfenati, Chemosphere 73 (2008) 1701–1707.
[5] P.P. Roy, J.T. Leonard, K. Roy, Chemometr. Intell. Lab. 90 (2008) 31–42.
[6] K. Roy, I. Sanyal, P.P. Roy, SAR QSAR Environ. Res. 17 (2006) 563–582.
[7] T. Ivanciuc, O. Ivanciuc, D.J. Klein, Mol. Divers. 10 (2006) 133–145.
[8] S. Dimitrov, N. Dimitrova, T. Parkerton, M. Comber, M. Bonnell, O. Mekenyan, SAR QSAR Environ. Res. 16 (2005) 531–554.
[9] S.D. Dimitrov, N.C. Dimitrova, J.D. Walker, G.D. Veith, O.G. Mekenyan, QSAR Comb. Sci. 22 (2003) 58–68.
[10] S.D. Dimitrov, N.C. Dimitrova, J.D. Walker, G.D. Veith, O.G. Mekenyan, Pure Appl. Chem. 74 (2002) 1823–1830.
[11] S.D. Dimitrov, O.G. Mekenyan, J.D. Walker, SAR QSAR Environ. Res. 13 (2002) 177–184.
[12] D. Wei, A. Zhang, C. Wu, S. Han, L. Wang, Chemosphere 44 (2001) 1421–1428.
[13] S. Kumar, M. Kumar, K. Thurow, R. Stoll, U. Kragl, Environ. Modell. Softw. 24 (2009) 44–53.
[14] A.P. Toropova, A.A. Toropov, A. Lombardo, A. Roncaglioni, E. Benfenati, G. Gini, Eur. J. Med. Chem. 45 (2010) 4399–4402.
[15] A. Golbraikh, A. Tropsha, J. Mol. Graph. Model. 20 (2002) 269–276.
[16] P.P. Roy, K. Roy, QSAR Comb. Sci. 27 (2008) 302–313.