
	<b>QMRF identifier (ECB Inventory):</b> To be entered by ECB	
	<b>QMRF Title:</b> Model to predict bioconcentration factors (BCF).	
	<b>Printing Date:</b> 28-Jul-2009	

## 1. QSAR identifier

### 1.1. QSAR identifier (title):

Model to predict bioconcentration factors (BCF).

### 1.2. Other related models:

Two models, model A and model B, have been used to build hybrid model, model C. In the proposed approach, the outputs of the individual models (model A and B) were used as inputs of the hybrid model.

Model A was developed by Radial Basis Function Neural Networks (RBFNN) using an heuristic method to select the optimal descriptors; model B was developed by Radial Basis Function Neural Networks (RBFNN) using genetic algorithm for the descriptors selection.

Details of model A and B are provided in Table1.pdf in 9.3, Supporting information.

The combined model C (hybrid) was built by an in house software made as PC-Windows Excel macro (available on request); codessa software, version 2.21 for HM (Heuristic Method); Moby Digs, version 1.0 (<http://www.taletе.mi.it>), genetic algorithm for GA-VSS (variable selection strategy); RBFNN (Wan and Harrington, 1999), matlab function, available on request; 2D descriptors: DRAGON version 5.4, MDL descriptors, ACD labs (version 9.08), Kowin (version 1.67).

### 1.3. Software coding the model:

In house software made as PC-Windows Excel macro (available on request).

## 2. General information

### 2.1. Date of QMRF:

21/07/2008

### 2.2. QMRF author(s) and contact details:

Elena Boriani Istituto di Rcerche Farmacologiche Mario Negri boriani@marionegri.it

### 2.3. Date of QMRF update(s):

27/02/2009

### 2.4. QMRF update(s):

Manuela Pavan

mpavan@miantd.com

Modified fields: 1.2; 1.3; 2.3; 2.4; 2.5; 2.6; 2.8; 4.2; 4.3; 4.4; 4.5; 4.6; 4.7; 5.1; 6.2; 6.5; 7.2; 8.2; 9.3

### 2.5. Model developer(s) and contact details:

[1] Chuyan Zhao Department of Chemistry, Lanzhou University, Lanzhou 730000, China

[2] Elena Boriani Istituto di Ricerche Farmacologiche Mario Negri boriani@marionegri.it <http://www.marionegri.it/mn/it/dipLab.html?id=94&ti=4>

[3] Antonio Chana Istituto di Ricerche Farmacologiche Mario Negri

[4] Alessandra Roncaglioni Istituto di Ricerche Farmacologiche Mario Negri aroncaglioni@marionegri.it <http://www.marionegri.it/mn/it/dipLab.html?ti=4&id=549>

[5]Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri benfenati@marionegri.it  
<http://www.marionegri.it/mn/it/dipLab.html?lab=168>

#### **2.6.Date of model development and/or publication:**

The model was published in 2008.

#### **2.7.Reference(s) to main scientific papers and/or software package:**

Zhao, C., Boriani, E., Chana, A., Roncaglioni, A., Benfenati, E. A new hybrid system of QSAR models for predicting bioconcentration factors (BCF). *Chemosphere* (2008), 73, 1701-1707.

#### **2.8.Availability of information about the model:**

The hybrid model is available (see attached supporting information MATLAB\_function.rar). To use the model matlab software, dragon descriptors and MDL descriptors are required. Soon free descriptors will be made available to use the model. A client server application will allow a simplified access to the model, at <http://www.caesar-project.eu>

#### **2.9.Availability of another QMRF for exactly the same model:**

### **3.Defining the endpoint - OECD Principle 1**

#### **3.1.Species:**

Fish (two databases combined; experimental data obtained according to OECD 305 protocol; fish species: Cyprinos Carpio and salmonids).

#### **3.2.Endpoint:**

2.Environmental fate parameters 4.Bioconcentration 2.4.a.BCF fish

#### **3.3.Comment on endpoint:**

BCF This endpoint is particularly required under REACH regulation. A good prediction for BCF endpoint may reduce the number of animal (fish) in experimental tests. REACH regulation states that a substance is identified as bioaccumulative (B) when  $BCF > 2000$  ( $\log BCF > 3.301$ ) and verybioaccumulative (vB) when  $BCF > 5000$  ( $\log BCF > 3.699$ ). Thus the endpoint could also be treated in classification and distinguish not B from v+vB compounds(it is not possible to distinguish V from vB because of the high variability in experimental data). This hybrid model performed well as classifier for B and vB chemicals.

#### **3.4.Endpoint units:**

adimensional

#### **3.5.Dependent variable:**

Log BCF

#### **3.6.Experimental protocol:**

OECD 305 (also required for REACH testing protocol).

#### **3.7.Endpoint data quality and variability:**

Variability of the test data: 0.75 log units (Dimitrov at al., 2005), reference in Bibliography , 9.2.

Quality check of the structures has been done (details on the procedure and the list of compounds discarded are described in qualityProcedure.pdf provided in the supplementary information )

### **4.Defining the algorithm - OECD Principle 2**

#### **4.1.Type of model:**

QSAR Hybrid model derived from model A developed by Radial Basis Function Neural Networks (RBFNN) using an heuristic method to select the optimal descriptors and model B developed by Radial Basis Function Neural Networks (RBFNN) using genetic algorithm for the descriptors selection.

Hybrid algorithm reference is provided in supporting information: Table1.pdf lists details of the models which gave the best results on the test set. Models 1-6 are single models.

QSAR Hybrid model derived from model A developed by Radial Basis Function Neural Networks (RBFNN) using an heuristic method to select the optimal descriptors and model B developed by Radial Basis Function Neural Networks (RBFNN) using genetic algorithm for the descriptors selection.

Hybrid algorithm reference is provided in supporting information: Table1.pdf lists details of the models which gave the best results on the test set. Models 1-6 are single models.

The hybrid model is based on these single models #2 and #5. The basic idea of a hybrid model is that each model brings a different content of the complex system which is modelled. (Amaury et al., 2007).

#### **4.2.Explicit algorithm:**

Hybrid model

Model available (see attached supporting information MATLAB\_function.rar), hybrid model and explanation to run it in README.txt in MATLAB\_function.rar in supporting information.

If mean (value given by models to combine) > 2.410

$$\log \text{BCF} = 1.052 * [\text{mean (value given by models to combine)}] - 0.065$$

If  $1.355 < \text{mean (value given by models to combine)} \leq 2.410$

$$\log \text{BCF} = 0.996 * [\text{min (value given by models to combine)}] + 0.042$$

Otherwise

$$\log \text{BCF} = 0.936 * [\text{mean (value given by models to combine)}] - 0.123$$

#### **4.3.Descriptors in the model:**

[1]Moriguchi octanol-water partition coefficient (MlogP) Moriguchi et al., 1994

[2]Moran autocorrelation (MATS5V) Molecular descriptor calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all paths of the considered path length (the lag)

[3]Number of chlorine atoms (Cl-089) Cl attached to C1(sp2)

[4]Absolute sums of eigenvalues (BEHp2) Molecular descriptor obtained from the positive and negative eigenvalues of the adjacency matrix, weighting the diagonal elements with atom weights.

[5]Geary autocorrelation (GATS5V) Molecular descriptor calculated from the molecular graph by summing the products of atom weights of the terminal atoms of all paths of the considered path length (the lag).

[6]X0Solv Solvation connectivity index (XOSolv) Molecular descriptor designed for modelling solvation entropy and describing dispersion interactions in solution.

[7]SsCl Sum of all (-Cl) E-state values in molecule

[8]Aeige Absolute eigenvalues sum from electronegativity weighted distance matrix

#### **4.4.Descriptor selection:**

The set of descriptors initially screened is made of 2D molecular descriptors, calculated by DRAGON version 5.4 (759 descriptors), MDL descriptors (249 descriptors), ACD labs version 9.08, (13 descriptors) and KOWIN (1 descriptor). Thus, 1022 descriptors were obtained including different logP and logD values calculated with these programs.

Heuristic and genetic algorithm methods were then used to select the optimal descriptors.

The hybrid model was derived from model A(#2) (HM +RBFNN) and model B (#5) (GA +RBFNN). A heuristic (HM) (Zhao et al., 2005) and genetic algorithm (GA) methods were used to select optimal descriptors. The software CODESSA version 2.21 was used for the HM, to give a complete search for the best multilinear correlations in the ordinary least squares regression (OLS) method. MobyDigs version 1.0 (<http://www.talete.mi.it>) was used for Genetic Algorithm-Variable Subset Selection strategy (GA-VSS).

#### **4.5. Algorithm and descriptor generation:**

Two models, model A and model B, have been used to build hybrid model, model C. In the proposed approach, the outputs of the individual models (model A and B) were used as inputs of the hybrid model. Model A was developed by Radial Basis Function Neural Networks (RBFNN) using an heuristic method to select the optimal descriptors; model B was developed by Radial Basis Function Neural Networks (RBFNN) using genetic algorithm for the descriptors selection.

Multiple Linear Regression (MLR) was used to develop the linear model of the property of interest, with CODESSA software. Radial Basis Function Neural Network (RBFNN) (Wan and Harrington, 1999) was used with a Matlab function for building the models. This function and result files containing the models are available on request. The hybrid model approach is based on the idea of using more representations of the problem, more paradigms of knowledge representation, and more algorithms to find a solution. As in other cases (Lo Piparo, 2006; Amaury et al., 2007), the outputs of the individual models were used as inputs of the hybrid model. An in-house software made as a PC-Windows Excel macro was used to build combined models.

The descriptors were calculated with DRAGON software.

#### **4.6. Software name and version for descriptor generation:**

DRAGON, version 5.4

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Prof. R. Todeschini - distributed by Talete srl, via Pisani 13, 20124 Milano, Italy

<http://www.talete.mi.it>

#### **4.7. Descriptors/Chemicals ratio:**

378 chemicals training / 8 descriptors = 47.25

### **5. Defining the applicability domain - OECD Principle 3**

#### **5.1. Description of the applicability domain of the model:**

Outliers were these class of compounds: perfluorinated sulfonic acid derivatives, phosphonothioate, phosphorothioate, compounds including thioester functions, compound highly reactive and highly degradable by humidity and light, i.e double peroxide, and Michael acceptor with high probability of reacting with a carbanion, compound with high degree of topological symmetry. Details on the outliers are provided in supporting information (Table2.pdf).

Within CAESAR a special tool was developed for all models. This tool, available at the website, shows the six most similar compounds present in our data set, and the related experimental and predicted values. In this way the user can have a direct, transparent, and clear assessment of the errors for similar compounds, and thus have a good basis for the evaluation of the applicability domain specific for a certain compound. Indeed, this information is related to the compound of interest.

Details on the tool are provided in caesar\_applet.pdf in supporting information.

### **5.2.Method used to assess the applicability domain:**

Expert judgement, from outliers. The applicability domain was evaluated with different methods. Some studies have been done characterising the composition of the training set, considering composition (atom type), complexity and polarity. For this target, different tools have been used, such as the Atom Centred Fragment developed within ChemProp, which characterise the compounds on the basis of typical features; or tools (also using ChemProp) for the characterisation of the about 80 functional groups, or molecular complexity (composition with C and H, or also N, O, P, halogens, etc), and of the polarizability of the compounds, in terms of non-polar or progressively more polar compounds, also keeping into account H-bond donors. The BCF data set is characterised by chemicals with a good presence of hydrocarbons and halogenated compounds, containing many chemicals with single functional groups in a high percentage. Many compounds are non polar or weakly polar.

In addition to this analysis, based on the composition of the training set, we made an analysis of the outliers, in order to identify chemical features, which can be related to lower model performance. This analysis is done ex post, and thus it keeps into account also the endpoint values, while the typical analysis of the applicability domain, as above done, is done only on the chemiometric information related to the chemical space (input), without considering the output. The analysis of the outliers allowed identifying a series of chemical features, based on the manual examination of some possible residues/fragments which could be related with the outliers occurrence. These hypotheses have been checked statistically exploring with chemiometric tools (ChemOffice) the presence of the candidate chemical features in compounds with correct and wrong predictions.

### **5.3.Software name and version for applicability domain assessment:**

### **5.4.Limits of applicability:**

See outliers description in 7.9

## **6.Internal validation - OECD Principle 4**

### **6.1.Availability of the training set:**

Yes

### **6.2.Available information for the training set:**

CAS RN:Yes

Chemical Name:Yes

Smiles:No

Formula:No

INChI:No

MOL file:No

### **6.3.Data for each descriptor variable for the training set:**

All

**6.4.Data for the dependent variable for the training set:**

All

**6.5.Other information about the training set:**

Classification for REACH for compounds v+vB and notB.

The whole training set is provided in supporting information (Training set.xls).

The training set structures are provided in supporting information (BCF.sdf)

**6.6.Pre-processing of data before modelling:**

**6.7.Statistics for goodness-of-fit:**

In supporting information, Table3.pdf are reported all the statistics done for goodness-of-fit for the models.

**6.8.Robustness - Statistics obtained by leave-one-out cross-validation:**

See 6.7

**6.9.Robustness - Statistics obtained by leave-many-out cross-validation:**

Leave many out (20%) cross validation models (20% of the compounds on the training set were randomly selected (sub-test set) and a model developed with the remaining ones (sub-training set). This procedure was repeated 10 times. Results is: Rcv2= 0.79 , SDEP = 0.66

**6.10.Robustness - Statistics obtained by Y-scrambling:**

**6.11.Robustness - Statistics obtained by bootstrap:**

**6.12.Robustness - Statistics obtained by other methods:**

**7.External validation - OECD Principle 4**

**7.1.Availability of the external validation set:**

Yes

**7.2.Available information for the external validation set:**

CAS RN:Yes

Chemical Name:Yes

Smiles:No

Formula:No

INChI:No

MOL file:No

**7.3.Data for each descriptor variable for the external validation set:**

All

**7.4.Data for the dependent variable for the external validation set:**

All

**7.5.Other information about the external validation set:**

Classification for REACH for compounds v+vB and notB.

The test set is provided in supporting information (Test set.xls).

The test set structures are provided in supporting information (BCF.sdf).

**7.6.Experimental design of test set:**

See 6.9

**7.7.Predictivity - Statistics obtained by external validation:**

Only five of the outliers (55, 57, 90, 291, 476) are false negatives (see 5.1 applicability domain).

## 7.8. Predictivity - Assessment of the external validation set:

In view of the large number of chemicals, the distribution of chemicals in the training and test set was not expected to affect the model evaluation, even using random separation.

## 7.9. Comments on the external validation of the model:

After generating the models and the hybrid system, the BCF value for outliers was further evaluated. For this study, the outliers can be defined as compounds with absolute predicted residual greater than 2 SDEP (Standard Deviation Error in Prediction). Outliers in the hybrid model are listed in Table 2.pdf and Supporting information.pdf. REPORT on outliers: Six outliers, out of 16, are wrongly predicted due to a significant difference between the MlogP and experimental BCF values. The high correlation between this descriptor and the experimental BCF values is fundamental in the models where the other descriptors affect the final value for a better adjustment between MLogP and experimental BCF. Among the outliers, the perfluorinated sulfonic acid derivatives (compounds 55, 57) were a small group, with no chemicals correctly predicted. The user of the models should be aware of this limitation, relative to this chemical class. Other chemical groups with more outliers than correctly predicted chemicals are phosphonothioate (484) and phosphorothioate (488) where the sulphur atom seems to give them an odd behavior as regards the phosphates (16 in total) while phosphonates are not included at all. Compounds including thioester functions should therefore be avoided for prediction. Again, this is the case of the thiocyanate function (511) which is found only once in the whole set of compounds used for model construction.

Another reason for outliers is the compound reactivity, so the BCF might actually refer to a chemical different from that introduced into the model. This seems to be the case of outliers 135, 356, 476 and 487 where compound 135 can easily be converted to the anhydride form due to the steric disposition.

Compound 356 is a highly reactive compound and highly degradable by humidity and light. Compound 476 is double peroxide, thus extremely reactive, and compound 487 is a Michael acceptor with high probability of reacting with a carbanion. Among the remaining outliers 63, 90, 288, 291 and 400 have an extremely high degree of topological symmetry. So the autocorrelation descriptors might be overestimated as regards the rest of the set. The two remaining outliers 111 and 486 show rather odd behavior since they can easily undergo esterification and hydrolysis, but other chemicals carrying carboxylic acids or phosphate esters are correctly predicted.

## 8. Providing a mechanistic interpretation - OECD Principle 5

### 8.1. Mechanistic basis of the model:

The model largely rely on logP, which typically is the main descriptor used for BCF. Corrections are applied to balance the use of the specific logP calculator, MLogP. Indeed, this particular descriptor gives good results when chemicals contain C,N,O, but it may be less accurate in case of compounds with other atoms, like Cl and P.

### 8.2. A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation of the model is provided a posteriori, i.e. by interpretation of the final set of the selected descriptors.

### 8.3. Other information about the mechanistic interpretation:

## 9. Miscellaneous information

### 9.1. Comments:

Prediction accuracy was 98%, sensitivity 96%, specificity 100% and geometric mean 0.98. Thus, the hybrid model also performed well as a classifier for “B” and “vB” chemicals. Another important feature of models for regulatory purposes is reproducibility, as we mentioned. To obtain that, the parameters of the model have to be fixed. The model has been codified (the software is available on request), and the user does not have to optimise the model parameters, which are now fixed. Thus, any user will get exactly the same results when introducing the descriptors for a given chemical, using the software described before.

This shows that the model is reproducible.

A further fundamental point for models for regulatory purposes is the quality check. The used experimental BCF values were obtained according to official protocols. Furthermore, as explained in the qualityProcedure.pdf, all structures were checked one-by-one within the EC funded project CAESAR, by at least two scientists.

### 9.2. Bibliography:

- [1] Dimitrov, S. et al., 2005. SAR QSAR Environ. Res., 16, 531-554
- [2] Amaury, N. et al., 2007. In Benfenati, E. (Ed.) Quantitative Structure-Activity Relationship (QSAR) for Pesticide Regulatory Purposes, Elsevier, Amsterdam, The Netherlands, pp. 149-183 and 187-189
- [3] Zhao et al., 2005. SAR QSAR Env. Res., 16, 349-367.
- [4] Lo Piparo et al., 2006. J. Med. Chem., 49, 5702-5709.
- [5] Wan, C. 1999. J. Chem. Inf. Comput. Sci., 39, 1049-1056.
- [6] Katrizky, A.R., et al. (2005). Comprehensive Descriptors for structural and Statistical Analysis. University of Florida. <http://www.semichem.com/codessa/default.php>
- [7] Hur, J., and Kim, J.W., (2008). A hybrid classification method using error pattern modelling. Expert Syst. Appl. 34, 231-241
- [8] Moriguchi, L. et al., 1994. Chem Pharm. Bull., 42, 976-978.

### 9.3. Supporting information:

#### Training set(s)

Training set.xls	file:///D:/IRFMN/BCF/Supporting Information/Training set.xls
BCF.sdf	file:///D:/IRFMN/BCF/Supporting Information/BCF.sdf

#### Test set(s)

Test set.xls	file:///D:/IRFMN/BCF/Supporting Information/Test set.xls
BCF.sdf	file:///D:/IRFMN/BCF/Supporting Information/BCF.sdf

#### Supporting information

Table 1.pdf	file:///D:/IRFMN/BCF/Supporting Information/Table 1.pdf
Supporting Information.pdf	file:///D:/IRFMN/BCF/Supporting Information/Supporting Information.pdf



qualityProcedure.pdf	file:///D:\IRFMN\BCF\Supporting Information\qualityProcedure.pdf
MATLAB_function.rar	file:///D:\IRFMN\BCF\Supporting Information\MATLAB_function\MATLAB_function.rar
Table 2.pdf	file:///D:\IRFMN\BCF\Supporting Information\Table 2.pdf
Table 3.pdf	file:///D:\IRFMN\BCF\Supporting Information\Table 3.pdf
caesar_applet.pdf	file:///D:\IRFMN\caesar_applet.pdf

## 10. Summary (ECB Inventory)

### 10.1. QMRF number:

To be entered by ECB

### 10.2. Publication date:

To be entered by ECB

### 10.3. Keywords:

To be entered by ECB

### 10.4. Comments:

To be entered by ECB