QMRF identifier (JRC Inventory):	
QMRF Title: *QSAR model for Bioaccumulation, BAF smelt*	
Printing Date: *16.12.2010*	

1.QSAR identifier

1.1.QSAR identifier (title):

QSAR model for Bioaccumulation, BAF smelt

1.2.Other related models:

1.3.Software coding the model:

QSARModel 4.0.4 Molcode Ltd., Turu 2, Tartu, 51014, Estonia http://www.molcode.com

2.General information

2.1.Date of QMRF:

10.12.2010

2.2.QMRF author(s) and contact details:

[1]Indrek Tulp Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[2]Tarmo Tamm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[3]Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[4]Dimitar Dobchev Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[5]Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[6]Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[7]Eneli Härk Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[8]Mati Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.6.Date of model development and/or publication:

10.06.2010

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

[1]Karelson M, Dobchev D, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Karelson G (2008). Correlation of blood-brain penetration

and human serum albumin binding with theoretical descriptors. ARKIVOC 16, 38-60.
[2]Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Dobchev D (2009). QSAR study of pharmacological permeabilities. ARKIVOC 2, 218–238.

2.8.Availability of information about the model:

All information in full detail is available

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

None to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Smelt

3.2.Endpoint:

2.Environmental fate parameters 5.Bioaccumulation 2.5.a.BAF fish

3.3.Comment on endpoint:

bioaccumulation factor on smelt

3.4.Endpoint units:

unitless

3.5.Dependent variable:

log(BAF)

3.6.Experimental protocol:

Bioaccumulation: flow-through fish test was determined using the OECD Test Guideline 305 (EU C.13). Bioaccumulation is the process of accumulation of chemicals by fish through non-dietary routes under flow-through conditions and is one of the key steps through which chemicals are able to enter into the biosphere from physical surroundings. Bioaccumulation factor (BAF) is defined as the ratio of the concentration of a chemical accumulated inside an organism (from food and direct exposure) to the concentration in the surrounding environment: BAF=Corg/Cenv

The release of polychlorinated biphenyls (PCBs) from industrial sources and their persistence in the environment have resulted in widespread contamination of water and soil, with subsequent potential exposure of the general population. PCBs have a low metabolism rate in the aquatic species, which makes them very persistent. Even low levels of PCBs in water can result in significant bioaccumulation in the food chain, due to their very slow degradation and high lipophilicity [1,2].

Experimentally determined BAF values for 46 PCBs were extracted from the Green Bay Mass Balance Study [3] and Hudson River Reassessment Remedial Investigation/Feasibility Study (RI/FS) [4]. For this multilaboratory study, a demanding quality assurance plan was used, and the data generated in this study are generally considered to be of the highest quality. The sampling design for the GBMB Study was based

upon fish habitat zones, and fish were collected in spring (April-June 1989), summer (July-August, 1989), and fall (October-November 1989) from each of the habitat zones. Species-specific composite samples, based upon the fish age were analyzed. Lipids in the fish were determined gravimetrically using 50:50 hexane-methylene chloride extraction solvent. For the Hudson River Reassessment RI/FS, the quality of PCB data has been extensively reviewed and documented [5], and the analytical methods used were highly comparable to those used in the GBMB Study.

3.7.Endpoint data quality and variability:

Experimental data from different sources (EPA database) has been used

Statistics:
max value: 8.68
min value: 5.90
standard deviation: 0.796
skewness: -0.0252

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

2D and 3D regression-based QSAR

4.2.Explicit algorithm:

multilinear regression QSAR
multilinear regression QSAR derived with BMLR (Best Multiple Linear Regression) method

log(BAF) = -296.262
-0.743Highest e-n attraction (AM1) for C - Cl bonds
+7.734E-002*Schultz principal eigenvalue of D matrix
+3.306E-002*WPSA2 Weighted PPSA (PPSA2*TMSA/1000) (Zefirov)

4.3.Descriptors in the model:

[1]Highest e-n attraction (AM1) for C - Cl bonds [eV] Highest electron - nuclei attraction energy of C-Cl bonds in the molecule
[2]Schultz principal eigenvalue of D matrix [unitless] Principal eigenvalue of Distance matrix
[3]WPSA2 Weighted PPSA (PPSA2*TMSA/1000) (Zefirov) [mÅ3] Surface weighted charged partial positive charged surface area WPSA2

4.4.Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules (one-parameter equations: Fisher criterion and R2 over threshold, variance and t-test value over threshold, intercorrelation with another descriptor not over threshold),

(two-parameter equations: intercorrelation coefficient below threshold, significant correlation with endpoint, in terms of correlation coefficient and t-test)

Stepwise trial of additional descriptors not significantly correlated to

any already in the model.

4.5.Algorithm and descriptor generation:

1D, 2D, and 3D theoretical calculations. Quantum chemical descriptors derived from AM1 calculation. Model developed by using multilinear regression.

4.6.Software name and version for descriptor generation:

QSARModel 4.0.4
QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling
Molcode Ltd, Turu 2, Tartu, 51014, Estonia
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

10.3 (31 chemicals / 3 descriptors)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:

a) by chemical identity: Polychlorinated biphenyls (PCBs)

b) by descriptor value range: The model is suitable for compounds that have the descriptors

in the following minimal-maximal range:

Highest e-n attraction (AM1) for C - Cl bonds: -399.4 - -397.9

Schultz principal eigenvalue of D matrix: 64.2 - 111

WPSA2 Weighted PPSA (PPSA2*TMSA/1000) (Zefirov): 20.6 - 47.8

5.2.Method used to assess the applicability domain:

Range of descriptor values in training set with ±30% confidence. Descriptor values must fall between maximal and minimal descriptor values of training set ±30%.

5.3.Software name and version for applicability domain assessment:

QSARModel 4.0.4
QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling
Molcode Ltd, Turu 2, Tartu, 51014, Estonia
http://www.molcode.com

5.4.Limits of applicability:

See 5.1

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:Yes
INChI:No

MOL file:Yes

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:

data points,
0 negative values,
31 positive values

6.6.Pre-processing of data before modelling:

n/a

6.7.Statistics for goodness-of-fit:

R2 = 0.960 (Correlation coefficient)
s2 = 0.170(Standard error of the estimate)
F = 116 (Fisher function)

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

R2CV = 0.950

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

R2CVMO = 0.948

6.10.Robustness - Statistics obtained by Y-scrambling:

n/a

6.11.Robustness - Statistics obtained by bootstrap:

n/a

6.12.Robustness - Statistics obtained by other methods:

ABC analysis (2:1 training : prediction) on sorted (in increased order of endpoint value) data divided into 3 subsets (A;B;C). Training set formed with 2/3 of the compounds (set A+B, A+C, B+C) and validation set consisted of 1/3 of the compounds (C, B, A).
average R2 (fitting) = 0.962
average R2 (prediction) = 0.946

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:Yes
INChI:No
MOL file:Yes

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:

15 data points,
0 negative values,
15 positive values

7.6.Experimental design of test set:

From sorted data source each 3rd was subjected to the test set.

7.7.Predictivity - Statistics obtained by external validation:

R2 = 0.915 (Coefficient of determination)

7.8.Predictivity - Assessment of the external validation set:

All compounds are in range of applicability domain:
Highest e-n attraction (AM1) for C - Cl bonds: -399.3 - -397.8
Schultz principal eigenvalue of D matrix: 66.3 - 106
WPSA2 Weighted PPSA (PPSA2*TMSA/1000) (Zefirov): 27.3 - 44.0

7.9.Comments on the external validation of the model:

The validation coefficient of determination (R2) is good and close to those coefficients of internal validation (R2CV and R2CVMO).

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Model possesses one size related topological descriptor "Schultz principal eigenvalue of D matrix" which considers chlorine substitutions. Charge distribution related descriptors "WPSA2 Weighted PPSA (PPSA2*TMSA/1000) (Zefirov)" and quantum-chemically calculated descriptor "Highest e-n attraction (AM1) for C - Cl bonds" are reflecting the charge distribution through electron densities counting resonance and induction effects in aromatic ring.

8.2.A priori or a posteriori mechanistic interpretation:

a posteriori mechanistic interpretation, consistent with published scientific interpretations of experiments

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Modeling the bioconcentration factors and bioaccumulation factors of polychlorinated biphenyls with posetic quantitative super-structure/activity relationships (QSSAR) DOI: 10.1007/s11030-005-9003-3
[2]Evaluation of two methods for prediction of bioaccumulation factors DOI: 10.1021/es0303889
[3]U.S. Environmental Protection agency, Great Lakes initiative http://www.epa.gov/greatlakes/
[4]U.S. Environmental Protection Agency. Database for the Hudson River

PCBs Reassessment RI/FS-Phase 2 Report; Release 4.1. U.S. EPA, Region II: New York, 1998.
[5]U.S. Environmental Protection Agency. Further Site Characterization and Analysis. Volume 2C, Data Evaluation and Interpretation Report. Hudson River PCBs Reassessment RI/FS; U.S. EPA, Region II: New York, 1997.

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

Karelson Arkivoc 2008	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf83_Karelson Arkivoc 2008.pdf
Karelson Arkivoc 2009	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf83_Karelson Arkivoc 2009.pdf

10.Summary (ECB Inventory)

10.1.QMRF number:

10.2.Publication date:

10.3.Keywords:

10.4.Comments: