	QMRF identifier (JRC Inventory): **Q13-203-0011**
	QMRF Title: **QSAR for acute toxicity to Pimephales promelas (fathead minnow)**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

QSAR for acute toxicity to Pimephales promelas (fathead minnow)

1.2.Other related models:

1.3.Software coding the model:

MOBY DIGS

Software for multilinear regression analysis and variable subset selection by Genetic Algorithm, ver. 1.0 beta for Windows, 2004

Todeschini Roberto, Talete srl, Milan (Italy)

2.General information

2.1.Date of QMRF:

29/05/2009

2.2.QMRF author(s) and contact details:

[1]Ester Papa University of Insubria ester.papa@uninsubria.it

[2]Fulvio Villa University of Insubria fulvio.villa@uninsubria.it

[3]Paola Gramatica University of Insubria paola.gramatica@uninsubria.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Ester Papa University of Insubria ester.papa@uninsubria.it

[2]Fulvio Villa University of Insubria fulvio.villa@uninsubria.it

[3]Paola Gramatica University of Insubria paola.gramatica@uninsubria.it

2.6.Date of model development and/or publication:

2005

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

Papa E, Villa F & Gramatica P (2005). Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow). Journal of Chemical Information and Modeling 45, 1256-1266.

2.8.Availability of information about the model:

This model is not proprietary, training and test sets are available

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

no information available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Pimephales promelas (Fathead Minnow)

3.2.Endpoint:

3.3.Acute toxicity to fish (lethality) 203Fish, Acute Toxicity Test

3.3.Comment on endpoint:

Flow-through bioassays, conducted with juvenile fathead minnows.

3.4.Endpoint units:

The median lethal concentrations are reported as the logarithm of the inverse molar concentration: log(1/LC50).

3.5.Dependent variable:

log1/EC50

3.6.Experimental protocol:

Experimentally determined 96h LC50 values for 468 industrial organic chemicals were collected from Russom et al. (1997) (original source: U.S.-E.P.A. Duluth Fathead Minnow Database). The data relate to flow-through bioassays, conducted with juvenile fathead minnows, on chemicals selected from a cross section of the Toxic Substances Control Act Inventory of industrial organic chemicals.

3.7.Endpoint data quality and variability:

A detailed analysis of the quality of the data reported in Duluth Fathead minnow database was made by Russom et al. (1997).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

multilinear regression QSAR

Log (1/LC50)96h = -2.54 (±0.4) + 0.91 (±0.06) WA + 6.2 (±0.6) Mv + 0.08 (±0.01) H-046 + 0.22 (±0.03) nCb - 0.19 (±0.04) MAXDP - 0.33 (±0.06) nN

4.3.Descriptors in the model:

[1]WA topological descriptor representing the mean Weiner indexv

[2]Mv constitutional descriptor, mean atomic van der Waals volume

[3]H-046 H attached to C-O sp3

[4]nCb- number of C sp2 in substituted benzenes

[5]MAXDP topological descriptor, maximal electrotopological positive variation

[6]nN number of nitrogen atoms

4.4.Descriptor selection:

A total of 1200 molecular descriptors of different kinds (0D, 1D, 2D, 3D) were calculated by the DRAGON software to describe the chemical diversity of the compounds. Constant values and descriptors found to be correlated pairwise were excluded in a prereduction step (when there was more than 98% pairwise correlation, one variable was deleted). Quantum-chemical descriptors such as HOMO (highest occupied molecular orbital), LUMO (lowest unoccupied molecular orbital), HOMO-LUMO gap (DHL), ionisation potential (P ion), and heat of formation (H), were added to Dragon descriptors. The Genetic Algorithm (GA) was applied to a final set of 400 descriptors for variable selection.

4.5.Algorithm and descriptor generation:

Multiple linear regression (Ordinary Least Square method) was applied to generate the model.

Molecular descriptors were generated by the DRAGON software. The input

files for descriptor calculation contain information on atom and bond types, connectivity, partial charges and atomic spatial coordinates, relative to the minimum energy conformation of the molecule, and were obtained by the Molecular Mechanics method of Allinger (MM+) using the package HYPERCHEM.

Quantum-chemical descriptors were calculated by the semiempirical PM3 Hamiltonian for the geometry optimization method available in the HYPERCHEM package.

4.6.Software name and version for descriptor generation:

DRAGON - 2005, version 5.2 for Windows

Software for the calculation of molecular descriptors

R. Todeschini, Talete s.r.l. Milano

HYPERCHEM - ver. 7.03

Software for molecular drawing and conformational energy optimization

4.7.Chemicals/Descriptors ratio:

249 chemicals / 6 descriptors = 41.5

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Structural Applicability Domain - high leverage chemicals (training set):

nicotine sulfate, 2,2'-methylenebis(3,4,6- trichlorophenol),

hexachloro-1,3-butadiene, pentachloropyridine, rotenone, and 2,6-di-*tert*-butyl-4-methylphenol.

Response domain - response outliers (training set):

chloroacetonitrile.

Structural Applicability Domain - high leverage chemicals (validation set):

tetrachloroethylene, hexachloroethane, 2,4,5-tribromoimidazole,

3-amino-5,6-dimethyl-1,2,4- triazine, caffeine, pentabromophenol.

Most of the chemicals falling outside the AD of the model belong to the specific acting compound class

5.2.Method used to assess the applicability domain:

The structural AD of the model was checked by the Leverage approach. The presence of outliers (i.e. compounds with cross-validated standardized residuals greater than 2.5 standard deviation units) and chemicals very structurally influential in determining model parameters (i.e. compounds with a high*leverage*value (*h*) greater than 3*p'/n* (*h**), where*p'*is the number of model variables plus one,

and*n*is the number of the objects used to calculate the model)

was verified by the Williams plot.

5.3.Software name and version for applicability domain assessment:

MOBY DIGS - MOdels BY Descriptors In Genetic Selection - ver.1 beta for Windows, Talete S.r.l., Milan, Italy, 2004.

Calculation of hat values, calculated - predicted values

Excel

Calculation of standardised residuals (compounds with crossvalidated standardized residuals greater than 2.5 standard deviation units

5.4.Limits of applicability:

High leverage compounds: hat value > 0.084

Outliers for the response: standardised residuals > 2.5 standard deviation units

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:

249 compounds

6.6.Pre-processing of data before modelling:

Transformation of LC50 into Log1/LC50

6.7.Statistics for goodness-of-fit:

n = 249; R^2 = 0.79; SDEP = 0.613; SDEC = 0.595; RMSE (training. set) = 0.38;

K_X= 34.81; K_{XY}= 39.94

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

Q^2_{LOO}= 0.78

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

$Q^2_{LMO(50\%)}$= 0.77

6.10.Robustness - Statistics obtained by Y-scrambling:

$R^2_{Y\text{-}SC}$= 0.024

6.11.Robustness - Statistics obtained by bootstrap:

Q^2_{BOOT}= 0.78

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:

200 compounds

7.6.Experimental design of test set:

The splitting of the original dataset into a training set of 249 chemicals representative of the entire data set and a validation set of 200 chemicals (splitting 50%) was realized by Kohonen artificial neural network (K-ANN) using the software KOALA: the three most significant principal components, calculated from each group of DRAGON molecular descriptors, were used to synthesize the structural information of the chemicals. This structural information and the response were used as variables to organize the structure of a Kohonen map. At the end of the net training, similar chemicals fell within the same neuron, i.e., they carried the same information. To select the training set of chemicals it was assumed that the compound closest to each neuron centroid was the most representative of all the chemicals within the same neuron.Thus, the training set chemicals were selected according to the minimal distance from the centroid of each cell in the top map. The remaining objects, close to the training set chemicals, were used for the validation set.

7.7.Predictivity - Statistics obtained by external validation:

Q^2_{EXT} = 0.71; RMSE (validation set) = 0.64

7.8.Predictivity - Assessment of the external validation set:

The application of Kohonen maps Artificial Neural Networks allowed for the selection of a large structurally representative validation set

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model was developed by a statistical approach. No mechanistic basis was defined a priori.

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori mechanistic interpretation. The theoretical descriptors selected in this model are a combination of global structural features, able to represent the high structural heterogeneity of the training and test sets: WA (topological descriptor representing the mean Weiner index), Mv (mean atomic van der Waals volume), nCb- (number of C sp2 in

substituted benzenes), H-046 (H attached to C-O sp3), MAXDP (maximal electrotopological positive variation), and nN (number of nitrogen atoms). The information related to dimensional features is condensed in WA and Mv, and the electronic distribution in MAXDP, while some counters (nN, nCb, and H-046) are mainly needed to model some particular chemicals in the data set.

8.3.Other information about the mechanistic interpretation:

No significant difference was found in the performances obtained with or without the inclusion of Log P.

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Russom CL, Bradbury SP, Broderius SJ, Drummond RA, & Hammermeister DE (1997). Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). Environmental Toxicology and Chemistry 16, 948-967.

[2]US EPA. Duluth Fathead Minnow Database http://www.epa.gov/med/databases/fathead_minnow.htm

9.3.Supporting information:

PimephalesTraining_249.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-203-0011/attachment/A644
Pimephales Test_200.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q13-203-0011/attachment/A645

Test set(s)

10.Summary (JRC QSAR Model Database)

Supporting information

10.1.QMRF number:

Q13-203-0011

10.2.Publication date:

2013-06-21

10.3.Keywords:

fathead minnow;Pimephales promelas;acute toxicity;

10.4.Comments:

formerQ2-17-11-126