	QMRF identifier (JRC Inventory): *To be entered by JRC*
	QMRF Title: *Insubria QSAR PaDEL-Descriptor model for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow)*
	Printing Date: *Feb 11, 2014*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for Modeling Aquatic Toxicity of Organic Chemicals in *Pimephales promelas* (Fathead Minnow)

1.2.Other related models:

Papa E., Villa F., Gramatica P., Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), J. Chem. Inf. Model., 2005,. 45, 1256-1266.[7]

1.3.Software coding the model:

[1]PaDEL-Descriptor 2.18 A software to calculate molecular descriptors and fingerprints http://padel.nus.edu.sg/software/padeldescriptor/index.html
[2]QSARINS 1.2 Software for the development, analysis and validation of QSAR MLR models paola.gramatica@uninsubria.it www.qsar.it

2.General information

2.1.Date of QMRF:

14-11-2013

2.2.QMRF author(s) and contact details:

Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Paola Gramatica DiSTA, University of Insubria (Varese - Italy) paola.gramatica@uninsubria.it www.qsar.it
[2]Stefano Cassani DiSTA, University of Insubria (Varese - Italy) +390332421439 stefano.cassani@uninsubria.it www.qsar.it

2.6.Date of model development and/or publication:

July 2013

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

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132 [1]
[2]Gramatica P., et al. QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J. Comput. Chem. (Software News and Updates), 2013.

2.8.Availability of information about the model:

The model is non-proprietary and published in a scientific peerreviewed journal. All information in full details are available (e.g.training and prediction set, algorithm, ecc...).

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.Ecotoxic effects 3.3.Acute toxicity to fish (lethality)

3.3.Comment on endpoint:

A selected set of experimental 96h LC50 data (from the original U.S.-E.P.A. Duluth Fathead Minnow Database) was taken from Russom et al. (1997) [2]. It consists of 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.

3.4.Endpoint units:

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

3.5.Dependent variable:

log(1/LC50)

3.6.Experimental protocol:

Experimentally determined 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)

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 - Multiple linear regression model (OLS - Ordinary Least Square)

4.2.Explicit algorithm:

Log 1/LC50 split model

MLR-OLS method. Model developed on a training set of 249 compounds.

Log 1/LC50 full model

MLR-OLS method. Model developed on a training set of 449 compounds.

Split model equation: log(1/LC50) 96h = 2.25 + 0.57 VP-1 - 1.09 MFLER_BH + 0.13 nAtomLAC - 0.88 HybRatio + 0.18 naasC - 0.25 nN

Full model equation: log(1/LC50) 96h = 2.31 + 0.56 VP-1 - 1.09

MFLER_BH + 0.12 nAtomLAC - 0.86 HybRatio + 0.17 naasC - 0.23 nN

4.3.Descriptors in the model:

[1]VP-1 Valence path, order 1
[2]MFLER_BH Overall or summation solute hydrogen bond basicity
[3]HybRatio Fraction of sp3 carbons to sp2 carbons
[4]nAtomLAC Number of atoms in the longest aliphatic chain
[5]naasC Count of atom-type E-State: :C:-
[6]nN number of nitrogens

4.4.Descriptor selection:

A total of 681 molecular descriptors of differing types (0D, 1D, 2D) were calculated in PaDEL-Descriptor 2.18. Constant and semi-constant values and descriptors found to be correlated pairwise were excluded in a pre-reduction step (one of any two descriptors with a correlation greater than 0.98 was removed to reduce redundant information), and a final set of 166 molecular descriptors were used as input variables for variable subset selection by genetic algorithm (GA-VSS).The models were initially developed by the all-subset-procedure until two variable. Then the GA was applied in order to explore new combinations of variables, selecting the variables by a mechanism of reproduction/mutation. The optimized parameter used was Q2LOO (leave-one-out).

4.5.Algorithm and descriptor generation:

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

Molecular descriptors were generated by PaDEL-Descriptor 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 firstly obtained by the semi empirical AM1 method using the package HYPERCHEM. Then, these files were converted by OpenBabel into MDL-MOL format and used as input for the calculation of descriptors in PaDEL-Descriptor.

4.6.Software name and version for descriptor generation:

PaDEL-Descriptor 2.18
A software to calculate molecular descriptors and fingerprints
Yap Chun Wei, Department of Pharmacy, National University of Singapore.
http://padel.nus.edu.sg/software/padeldescriptor/index.html

HYPERCHEM - ver. 7.03
Software for molecular drawing and conformational energy optimization

OpenBabel ver.2.3.2
Open Babel: The Open Source Chemistry Toolbox. Used for conversion between HYPERCHEM files (hin) and MDL-MOL files.

http://openbabel.org

4.7.Chemicals/Descriptors ratio:

Split Model: 249 chemicals / 6 descriptors = 41.5

Full Model: 449 chemicals / 6 descriptors = 74.8

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model was verified by the leverage approach and fixed thresholds has been used to define both structural and response outliers (see section 5.4). The plot of leverages (hat diagonals) versus standardised residuals, i.e. the Williams plot, verified the presence of response 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 parameters (i.e. compounds with a leverage value (h) greater than 3p'/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). For new compounds without experimental data, leverage can be used as a quantitative measure for evaluating the degree of extrapolation: for compounds with a high leverage value (h > h*), that are structural outliers, predictions should be considered less reliable.

Response and descriptor space:

Range of experimental log(1/LC50) values: 0.04 / 8.45

Range of descriptor values: nN (0 / 4), VP-1 (0.45 / 11.11), naasC (0 / 10), HybRatio (0 / 1), nAtomLAC (0 / 13), MLFER_BH (-0.41 / 3.38)

5.2.Method used to assess the applicability domain:

As it has been stated in section 5.1, the structural applicability domain of the model was assessed by the leverage approach, providing a cut-off hat value (h*=0.047). HAT values are calculated as the diagonal elements of the HAT matrix:

$H = X(X^TX)^{-1}X^T$

The response applicability domain can be verified by the standardized residuals, calculated as: $r'i = ri / s\sqrt{(1-hii)}$, where $ri = Yi-\hat{Y}i$.

5.3.Software name and version for applicability domain assessment:

QSARINS 1.2

Software for the development, analysis and validation of QSAR MLR models

paola.gramatica@uninsubria.it

www.qsar.it

5.4.Limits of applicability:

Split model domain: outliers for structure, hat>0.084 (h*): Nicotine Sulfate (65-30-5), Hexachlorophene (70-30-4), 1,2,4-Triazin-3-amine (17584-12-2), Rotenone (83-79-4), Caffeine (58-08-2), Fensulfothion (115-90-2), Diazinon (333-41-5), Malathion (121-75-5), carbophenothion (786-19-6), 2,4-dinitroaniline (97-02-9). Outliers for response, standardised residuals > 2.5 standard deviation units:

chloroacetonitrile (107-14-2), hexylene glycol (107-41-5), propylene glycol monoacrylate (999-61-1), 2-hydroxyethyl acrylate (818-61-1), hexachloroethane (67-72-1), 1-Octyne-3-ol (818-72-4), But-2-yn-1-ol (764-01-2), isovalerylaldehyde (590-86-3), 3-butyn-2-ol (65337-13-5). **FULL model domain**: outliers for structure, hat>0.047 (h*): Nicotine Sulfate (65-30-5), Hexachlorophene (70-30-4), 1,2,4-Triazin-3-amine (17584-12-2), Rotenone (83-79-4), Caffeine (58-08-2), Fensulfothion (115-90-2), Diazinon (333-41-5), Malathion (121-75-5), carbophenothion (786-19-6), Chlorpyrifos (2921-88-2). Outliers for response, standardised residuals > 2.5 standard deviation units: chloroacetonitrile (107-14-2), hexylene glycol (107-41-5), propylene glycol monoacrylate (999-61-1), 2-hydroxyethyl acrylate (818-61-1), hexachloroethane (67-72-1), 1-Octyne-3-ol (818-72-4), 2-butyn-1-ol(764-01-2).

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:Yes
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:

The training set of the Split Model consists of 249 heterogeneous organic compounds (including almost all the principal functional groups present mainly in pesticides) with a range of log(1/LC50) values from 0.04 to 8.45. Training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis (SOM).

6.6.Pre-processing of data before modelling:

Transformation of LC50 into Log1/LC50

6.7.Statistics for goodness-of-fit:

R^2= 0.77; CCCtr [3]=0.87; RMSE= 0.62

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

Q^2LOO= 0.76; CCCcv=0.86; RMSEcv= 0.64

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

Q^2LMO= 0.76

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2y-sc= 0.02

6.11.Robustness - Statistics obtained by bootstrap:

No information available (since we have calculated Q^2LMO)

6.12.Robustness - Statistics obtained by other methods:

No information available

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:Yes
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:

The external prediction set consists of 200 heterogeneous organic compounds with a range of log(1/LC50) values from 0.84 to 6.72. Training and prediction set are structurally balanced, being the splitting based on the structural similarity analysis.

7.6.Experimental design of test set:

The splitting of the original data set (449 compounds) into a training set of 249 compounds and a prediction set of 200 compounds was realized by Kohonen artificial neural network.

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [4]= 0.71; Q^2extF2 [5]= 0.71; Q^2extF3 [6]= 0.77; CCCex=0.84; RMSE= 0.63

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on similarity analysis (performed by the application of the Kohonen maps Artificial Neural Networks - KANN) allowed for the selection of a meaningful training set and a representative prediction set.

Training and prediction set are balanced according to both structure and response. In particular, for response the range of log(1/LC50) values are [0.04 / 8.45] and [0.84 - 6.72] respectively for training and prediction set. As much as concern structural representativity, the range of descriptors values are:

nN: training set (0 / 4), prediction set (0 / 4)
VP-1: training set (0.45 / 11.11), prediction set (0.81 / 8.71)
naaasC: training set (0 / 10), prediction set (0 / 6)
HybRatio: training set (0 / 1), prediction set (0 / 1)
nAtomLAC: training set (0 / 13), prediction set (0 / 12)

MLFER_BH: training set (-0.41 / 3.38), prediction set (-0.16 / 2.29)

7.9.Comments on the external validation of the model:

No information available

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model was developed by statistical approach. No mechanistic basis for this physico-chemical property was set a priori, but a mechanistic interpretation of molecular descriptors was provided a posteriori (see 8.2).

8.2.A priori or a posteriori mechanistic interpretation:

The DRAGON model published in Papa et al.[7] is: $\text{Log}(1/LC50)_{96h} = -2.54 + 0.91WA + 6.2Mv + 0.08H\text{-}046 + 0.22nCb - 0.19MAXDP - 0.33nN$ WA= mean Wiener index Mv= mean atomic van der Waals volume (scaled on Carbon atom) H-046= H attached to CO(sp3) no X attached to next C nCb= number of C sp2 in substituted benzenes MAXDP= maximal electrotopological positive variation nN= number of Nitrogen atoms. The equation of the new PaDEL-descriptor model included in QSARINS is : $\log(1/LC50)_{96h} = 2.31 + 0.56\ VP\text{-}1 - 1.09\ MFLER\_BH + 0.12\ nAtomLAC - 0.86\ HybRatio + 0.17\ naasC - 0.23\ nN$ VP-1: Valence path, order 1 MFLER_BH: Overall or summation solute hydrogen bond basicity HybRatio: Fraction of sp3 carbons to sp2 carbons nAtomLAC: Number of atoms in the longest aliphatic chain naasC: Count of atom-type E-State: :C:- nN: number of nitrogens The theoretical descriptors selected in this model (see Section 4.3 for a short explanation) are a combination of global structural features, able to represent the high structural heterogeneity of the training and test sets: VP-1, MFLER_BH, HybRatio, nAtomLAC, naasC, nN. The information related to dimensional features is condensed in VP-1, while some counters (nN, nAtomLAC, naaSC) are mainly needed to model some particular chemicals in the data set.

8.3.Other information about the mechanistic interpretation:

No other information available

9.Miscellaneous information

9.1.Comments:

To predict log(1/LC50) in *Pimephales promelas* for new chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=449), thus ensuring a wider applicability domain.

The equation (reported also in section 4.2) and the statistical parameters of the full model are:

$\log(1/LC50)\ 96h = 2.31 + 0.56\ VP\text{-}1 - 1.09\ MFLER\_BH + 0.12\ nAtomLAC - 0.86\ HybRatio + 0.17\ naasC - 0.23\ nN$

$N = 449;\ R^2 = 0.75;\ Q2 = 0.74;\ Q2LMO = 0.75;\ CCC = 0.86;$

CCCcv = 0.85 ;RMSE= 0.626; RMSEcv = 0.637

9.2.Bibliography:

[1]Gramatica P., et al. QSARINS: A new software for the development, analysis and validation of QSAR MLR models, J. Comput. Chem. (Software News and Updates), 2013, 34 (24), 2121-2132.
[2]Russom, C. L.; Bradbury, S. P.; Broderius, S. J.; Drummond, R. A.; Hammermeister, D. E. Predicting modes of toxic action from chemical structure: Acute toxicity in the fathead minnow (Pimephales promelas). EnViron. Toxicol. Chem. 1997, 16, 948-967.
[3]Chirico N. and Gramatica P., Real External Predictivity of QSAR Models. Part 2. New Intercomparable Thresholds for Different Validation Criteria and the Need for Scatter Plot Inspection, J. Chem. Inf. Model. 2012, 52, pp 2044– 2058
[4]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.
[5]Schuurman G. et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficient - Test Set Activity Mean vs Training Set Activity Mean, J. Chem. Inf. Model. 48 (2008) 2140-2145.
[6]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678
[7]Papa E., Villa F., Gramatica P., Statistically Validated QSARs, Based on Theoretical Descriptors, for Modeling Aquatic Toxicity of Organic Chemicals in Pimephales promelas (Fathead Minnow), J. Chem. Inf. Model., 2005,. 45, 1256-1266.

9.3.Supporting information:

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

10.Summary (JRC Inventory)

10.1.QMRF number:

To be entered by JRC

10.2.Publication date:

To be entered by JRC

10.3.Keywords:

To be entered by JRC

10.4.Comments:

To be entered by JRC