	QMRF identifier (JRC Inventory): *To be entered by JRC*
	QMRF Title: *Insubria QSAR PaDEL-Descriptor model for prediction of Esters toxicity in Daphnia magna*
	Printing Date: *Jan 20, 2014*

1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of Esters toxicity in *Daphnia magna*

1.2.Other related models:

E.Papa, F. Battaini, P.Gramatica. Ranking of aquatic toxicity of esters modelled by QSAR, Chemosphere (58), 2005, 559-570.[9]

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:

05/12/2013

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

2.6.Date of model development and/or publication:

September 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 other information available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia magna

3.2.Endpoint:

3.Ecotoxic effects 3.1.Short-term toxicity to Daphnia (immobilisation)

3.3.Comment on endpoint:

Experimental toxicity data (EC50) were taken from the literature (Cash and Clements, 1996; Staples et al., 1997; IUCLID, 2000)[2-4]; all data are reported in mmol/l and transformed in logarithmic units.

3.4.Endpoint units:

The median effect concentrations are reported as the logarithm of the inverse molar concentration: log(1/EC50) mmol/L

3.5.Dependent variable:

log (1/EC50) or pEC50

3.6.Experimental protocol:

The data selected in IUCLID are related to tests performed according to OECD and GPL norms.

3.7.Endpoint data quality and variability:

No information available

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:

pEC50 D.magna PC1 Split model
OLS-MLR method. Model developed on a training set of 24 compounds

pEC50 D.magna FULL model
OLS-MLR method. Model developed on a training set of 29 compounds

PC1 Split model equation: pEC50 = 0.03 + 0.51 VP-2 - 0.62 nsCH3 + 1.61 minHdCH2

Full model equation: pEC50 = 0.37 + 0.53 VP-2 - 0.78 nsCH3 + 1.70 minHdCH2

4.3.Descriptors in the model:

[1]VP-2 Valence path, order 2
[2]nsCH3 Count of atom-type E-State: -CH3
[3]minHdCH2 Minimum atom-type H E-State: =CH2

4.4.Descriptor selection:

A total of 720 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 109 molecular descriptors were used as input variables for variable subset selection. The models were initially developed by the all-subset-procedure, and then GA was applied to obtain the final population of models (three variables). 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 by PC1 model: 24 chemicals / 3 descriptors = 8
Full model: 29 chemicals / 3 descriptors = 9.67

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 pEC50 *D.magna* values: -1.14 / 2.51

Range of descriptor values: VP-2: 0.70 / 7.91 ; nsCH3: 0 / 4; minHdCH2: 0 / 0.56.

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*=0.414). HAT values are calculated as the diagonal elements of the HAT matrix:

H = X(XTX)-1XT

The response applicability domain can be verified by the standardized residuals, calculated as: r'i = ri / s√(1-hii), where ri = Yi-Ŷi.

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:

PC1 Split model domain: outliers for structure, hat>0.500 (h*): 2-hydroxyethyl acrylate (818-61-1). Outliers for response, standardised residuals > 2.5 standard deviation units: methyl acrylate (96-33-3). **FULL model domain**: outliers for structure, hat>0.414 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: no.

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:

To verify the predictive capability of the proposed models, the dataset (n=29) was split, before model development, into a training set used for model development and a prediction set used later for external validation. The splitting is based on PCA analysis (Ordered PC1 score) and the training set is composed of 24 chemicals.

6.6.Pre-processing of data before modelling:

Transformation of EC50 into Log1/EC50 (or pEC50) mmol/L

6.7.Statistics for goodness-of-fit:

R^2= 0.86; CCCtr[5]=0.92; RMSE= 0.37

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

Q^2LOO= 0.80; CCCcv=0.89; RMSEcv= 0.44

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

Q^2LMO= 0.78.

6.10.Robustness - Statistics obtained by Y-scrambling:

R^2y-sc= 0.13

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:

To verify the predictive capability of the proposed models, the dataset (n=29) was split, before model development, into a training set used for model development and a prediction set used later for external validation. The splitting is based on PCA analysis (Ordered PC1 score) and the prediction set is composed of 5 chemicals, with a range of pEC50 : -0.24 / 2.51.

7.6.Experimental design of test set:

Chemicals were ordered according to their increasing PC1 score (after a PCA analysis of the modeling descriptors), and one out of every five chemicals was put in the prediction set.

7.7.Predictivity - Statistics obtained by external validation:

Q^2extF1 [6]= 0.82; Q^2extF2 [7]= 0.75; Q^2extF3 [8]= 0.66; CCCex=0.87; RMSE= 0.57

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based PC1 score allowed for the selection of meaningful training sets and representative prediction sets.

Training and prediction sets are balanced according to both structure and response. In particular, for response the range of pEC50 values are [-1.14 / 2.34][-0.24 / 2.51] respectively for training and prediction set.

As much as concern structural representativity, the range of descriptors values is:

VP-2: training set (0.70 / 7.91), prediction set (0.73 / 5.1);

nsCH3: training set (1 / 4), prediction set (0 / 3);

minHdCH2: training set (0/ 0.56), prediction set (0/ 0.56).

7.9.Comments on the external validation of the model:

no other 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.[9] is:

pEC50 = -0.193 + 0.05 TIC0 -0.82 nCp + 0.94 n=CH2

where

TIC0: total information content index (neighborhood symmetry 0-order)

nCp: number of terminal primary C(sp3)(could be an indicator of both molecule dimension and shape)

n=CH2: is a counter of double bonds and could show the relevance of reactivitysites in the structure.

The equation of the new PaDEL-descriptor model included in QSARINS is : pEC50 = 0.37 + 0.53 VP-2 -0.78 nsCH3 + 1.70 minHdCH2

where

VP-2= Valence path, order 2

nsCH3= Count of atom-type E-State: -CH3 maxHdsCH2= Maximum atom-type H E-State: =CH2

The correlation between VP-2 and TIC0 is very high (99%), therefore

the two descriptors have similar structural relevance in modeling the response, between nCp and nsCH3 is acceptable (72%). Also if n=CH2 is no longer available in DRAGON, its meaning ("a counter of double bonds and could show the relevance of reactivity sites in the structure") is comparable with PaDEL-Descriptor variable maxHdsCH2.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict pEC50 for new Esters without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=29), thus ensuring a wider applicability domain. The full model equation (reported also in section 4.2) and the statistical parameters are the following:

pEC50 = 0.37 + 0.53 VP-2 - 0.78 nsCH3 + 1.70 minHdCH2

N = 29; R^2 = 0.86; Q^2 = 0.82; Q^2LMO = 0.81; CCC = 0.93; CCCcv = 0.90; RMSE= 0.39; RMSEcv = 0.44.

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]Cash, G.G., Clements, R.G., Comparison of structure– activity relationships derived from two methods for estimating octanol–water partition coefficients. SAR QSAR Environ. Res. 5, 1996, 113–124.
[3]Staples, C.A., Adams, W.J., Parkerton, T.F., Gorsuch, J.W., Biddinger, G.R., Reinert, K.H., Aquatic toxicity of eighteen phthalate esters. Environ. Toxicol. Chem. 16, 1997, 875– 891
[4]IUCLID CD-ROM, 2000. European Commission Joint Research Centre.
[5]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
[6]Shi L.M. et al. QSAR Models Using a Large Diverse Set of Estrogens, J. Chem. Inf. Comput. Sci. 41 (2001) 186–195.
[7]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.
[8]Consonni V. et al. Comments on the Definition of the Q2 Parameter for QSAR Validation, J. Chem. Inf. Model. 49 (2009) 1669-1678
[9]E.Papa, F. Battaini, P.Gramatica. Ranking of aquatic toxicity of esters modelled by QSAR, Chemosphere (58), 2005, 559-570.

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