1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for Modeling PFC inhalation toxicity in mouse

1.2.Other related models:

Bhhatarai B., Gramatica P., Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling, Chem. Res. Toxicol., 2010, 23, 528–539 [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:

02/12/2013

2.2.QMRF author(s) and contact details:

Alessandro Sangion DiSTA, University of Insubria (Varese - Italy) a.sangion@hotmail.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) 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]QSARINS-Chem: Insubria Datasets and New QSAR/QSPR Models for Environmental Pollutants in QSARINS, submitted to J. Comput. Chem. (Software News and Updates)

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Mouse (Mus)

3.2.Endpoint:

4.Human health effects 4.1.Acute inhalation toxicity

3.3.Comment on endpoint:

lethal concentration 50 (LC50)

Standard measure of the toxicity of the surrounding medium that will kill half of the sample population of a specific test-animal in a specified period through exposure via inhalation (respiration). LC50 is measured in micrograms (or milligrams) of the material per liter, or parts per million (ppm), of air or water

3.4.Endpoint units:

The median lethal concentrations are reported as the inverse log of the molar concentration: pLC50 mouse (mmol/m^3)

3.5.Dependent variable:

pLC50

3.6.Experimental protocol:

The experimental data on mouse LC50 inhalation toxicities were collected from ChemID plus [2]

3.7.Endpoint data quality and variability:

The ChemID plus data was verified as much as possible and filtered by performing principle component analysis (PCA) and by omitting the spurious compounds which could badly influence the regression models.

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:

pLC50 PaDEL-Descriptor full model for PFC Mouse inhalation Toxicity
OLS - Multiple linear Regression Model developed on a training set of 56 chemicals

pLC50 PaDEL-Descriptor split model (SOM) for PFC Mouse inhalation Toxicity
OLS - Multiple linear Regression Model developed on a training set of 40 chemicals

pLC50 PaDEL-Descriptor split model (Ordered Response) for PFC Mouse inhalationToxicity

OLS - Multiple linear Regression Model developed on a training set of 44 chemicals

Full model equation: pLC50= 2.95 + 1.36 VP-3 + 0.05 TopoPSA - 1.03 nsssCH - 0.42 XlogP

Split by SOM model equation: pLC50= 3.28 + 1.27 VP-3 -1.06 nsssCH -0.44 XLogP + 0.04 TopoPSA

Split by Ordered Response model equation: pLC50= 2.98 + 1.35 VP-3 + 0.04 TopoPSA -1.10 nsssCH -0.42 XLogP

4.3.Descriptors in the model:

[1]VP-3 Valence path, order 3

[2]nsssCH Count of atom-type E-State: >CH-

[3]XlogP A logP calculated in PaDEL-Descriptor

[4]TopoPSA Topological polar surface area

4.4.Descriptor selection:

A total of 1609 molecular descriptors of different kinds (0D, 1D, 2D, fingerprints) were calculated by PaDEL-Descriptor software to describe the chemical diversity of the compounds. Constant and semi-constant (at least 20% compounds must have values different from zero or from the values of other chemicals) values and descriptors found to be pair-wise correlated more than 0.98 were excluded in a prereduction step. The Genetic Algorithm (GA) was applied to a final set of 144 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 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

An open source software to calculate molecular descriptors and fingerprints, ver. 2.13, 2012.

Yap C.W, 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.2.3.0, 2010
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:

Full model: 56 chemicals / 4 descriptros = 14
Split by SOM: 40 chemicals / 4 descriptors = 10
Split by Ordered response: 44 chemicals / 4 descriptors = 11

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 pLC50 values: 0.269 / 6.542.
Range of descriptor values:VP-3 (0 / 2.88) XLogP (0.619 / 7.81) TopoPSA (0 / 47.58) nsssCH (0 / 3)

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*=0.268). 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:

Full model domain:outliers for structure, hat>0.268 (h*): Propane, 2-chloro-1,1,3,3-tetrafluoro- (19041-02-2); Perfluorodibutyl ether (308-48-5); Outliers for response, standardised residuals > 2.5 standard deviation units: no

Split by SOM model domain: outliers for structure, hat>0.375 (h*): Propane, 2-chloro-1,1,3,3-tetrafluoro- (19041-02-2). Outliers for response, standardised residuals > 2.5 standard deviation units: no

Split by Ordered Response model domain: outliers for structure, hat>0.341 (h*): Propane, 2-chloro-1,1,3,3-tetrafluoro- (19041-02-2); Perfluorodibutyl ether (308-48-5); Pentadecafluorotriethylamine (359-70-6).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:

The training set of the Split by SOM Model consists of 40 perfluorinated compounds with a range of pLC50 values from 0.269 to 6.542.

The training set of the Split by Ordered Response Model consists of 44 perfluorinated compounds with a range of pLC50 values from 0.315 to 6.255.

6.6.Pre-processing of data before modelling:

The original g/m^3 data were converted into the mmol/m^3 and expressed in inverse log unit for modeling which are represented as pLC50

6.7.Statistics for goodness-of-fit:

Split by SOM Model:

R^2: 0.79; CCCtr[3]: 0.88; RMSEtr: 0.68

Split by Ordered Response Model:

R^2: 0.72; CCCtr: 0.84; RMSEtr: 0.77

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

Split by SOM Model:

Q^2loo: 0.74; CCCcv: 0.85; RMSEcv: 0.76

Split by Ordered Response Model:

Q^2loo: 0.66; CCCcv: 0.81; RMSEcv: 0.85

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

Split by SOM Model: Q^2LMO: 0.73

Split by Ordered Response Model: Q^2LMO: 0.67

6.10.Robustness - Statistics obtained by Y-scrambling:

Split by SOM Model: R^2Yscr: 0.10

Split by Ordered Response Model: R^2Yscr: 0.09

6.11.Robustness - Statistics obtained by bootstrap:

No information available (since we have calculated Q2LMO)

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=56) was split, before model development, into a training set used for model development and a prediction set used later for external validation. Two different splitting techniques were applied: by **Ordered Response** (n external validation set =12) and by **structural similarity (SOM)** (n external validation set =16).

7.6.Experimental design of test set:

In the case of split by **Ordered Response model**, chemicals were ordered according to their increasing activity, and one out of every five chemicals was put in the prediction set (always including the most and the least active compounds in the training set). The splitting by **SOM model** takes advantages of the clustering capabilities of Kohonen Artifical Neural Network (K-ANN), allowing the selection of a structurally meaningful training set and an equally representative prediction set.

7.7.Predictivity - Statistics obtained by external validation:

Split by SOM model: n prediction= 16; R^2ext = 0.78; Q^2ext F1[4] = 0.77; Q^2 ext F2 [5]= 0.71; Q^2 ext F3 [6]= 0.70; CCCex = 0.81; RMSEex = 0.80; MAEex = 0.60.

Split by Oredered Response model: n prediction= 12; R^2ext = 0.95; Q^2ext F1= 0.95; Q^2 ext F2 = 0.95; Q^2 ext F3 = 0.93; CCCex = 0.97;

RMSEex = 0.40; MAEex = 0.35 .

7.8.Predictivity - Assessment of the external validation set:

Range of response for prediction set (**SOM split**, n=16) compounds:

log(1/LC50) $mmol/m^3$: 0.315 / 5.368 (range of corrispondig training set: 0.269 / 6.542)

Range of modeling descriptors for prediction set (**SOM split**, n=16) compounds:

VP-3: 0.15 / 1.43 (range of corrispondig training set: 0 / 2.88)
XLogP: 0.859 / 4.616 (range of corrispondig training set: 0.619 / 7.81)

TopoPSA: 0 / 26.3 (range of corrispondig training set: 0 / 47.58)
nsssCH: 0 / 2 (range of corrispondig training set:0 / 3)

Range of response for prediction set (**Ordered Response split**, n=12) compounds: log(1/LC50) $mmol/m^3$: 0.269 / 6.542 (range of corrispondig training set: 0.315 / 6.255) Range of modeling descriptors for prediction set (**Ordered Response split**, n=12) compounds: VP-3: 0.247 / 2.488 (range of corrispondig training set: 0 / 2.876)
XLogP: 1.043 / 7.81 (range of corrispondig training set: 0.619 / 4.543)
TopoPSA: 0 / 47.58 (range of corrispondig training set: 0 / 47.58)

nsssCH: 0 / 3 (range of corrispondig training set:0 / 2)

The distribution of response values of the chemicals in the two different training sets is comparable to the distribution of the response values of the two prediction set.

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 was defined a priori.

8.2.A priori or a posteriori mechanistic interpretation:

The DRAGON model published in Bhhatarai B. and Gramatica P. [7] is:

pLC50= 4.21 - 1.27 MLOGP + 1.43 X3v + 0.38 F01[C-C] - 1.14 H-048

where MLOGP: Moriguchi octanol-water partition coeff.
X3v: valence connectivity index chi-3
F01[C-C]: frequancy of C-C at topological distance 01
H-048: H attached to C2(sp3)/C1(sp2)/C0(sp)

The most influential descriptor was MlogP with the negative coefficient, which shows that for fluorinated chemicals, the contribution of hydrophobicity, within this combination of descriptors, demonstrates a decreasing trend for mouse inhalation toxicity.

The other two successive descriptors X3v and F01[C-C] had a positive influence on mouse toxicity. The X3V (valence connectivity index) accounts for the presence of the heteroatom and double and triple bonds present in the compound. Increase in one or the other features increases the value of X3v in total. The F01[C-C] represents the total number of C-C bonds. As the alkyl chain length increases, C-C increases, giving high values to the longer chain. Thus, increasing chain length and increase in bond order, as well as the presence of the heteroatom contributes to increase in mouse inhalation toxicity.

The least significant H-048 and its negative coefficient shows the decrease in toxicity value for compounds which have higher values of this descriptor.

The equation of the new PaDEL-descriptor model included in QSARINS is :

pLC50= 2.95 + 1.36 VP-3 + 0.05 TopoPSA - 1.03 nsssCH - 0.42 XlogP

where VP-3= Valence path, order 3

nsssCH= Count of atom-type E-State: >CH- XlogP= a calculated logP value

TopoPSA= Topological polar surface area

Two variables in DRAGON and PaDEL model give exactly the same structural information: X3v and VP-3 (correlation:1), similarly regarding H-048 and nsssCH, which are highly correlated (0.92). Both models are based on logP, even if differently calculated (MLOGP in DRAGON model, XlogP in PaDEL-Descriptor model), with an intercorrelation of 0.83.

Therefore, out of 4 modeling descriptors, 3 shows high correlation and the two models carry the similar structural information.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict inhalation toxicity in mouse for new PFC chemicals without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=56), thus ensuring a wider applicability domain.

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

Full model equation: pLC50= 2.95 + 1.36 VP-3 + 0.05 TopoPSA - 1.03 nsssCH - 0.42 XlogP

N = 56; R^2 = 0.79; Q^2 = 0.75; Q^2LMO = 0.75; CCC = 0.88; CCCcv = 0.86; RMSE= 0.70; RMSEcv = 0.76

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]ChemID Plus http://chem.sis.nlm.nih.gov/chemidplus/

[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]Bhhatarai B., Gramatica P., Per- and Polyfluoro Toxicity (LC50 Inhalation) Study in Rat and Mouse Using QSAR Modeling, Chem. Res. Toxicol., 2010, 23, 528–539

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