1.QSAR identifier

1.1.QSAR identifier (title):

Insubria QSAR PaDEL-Descriptor model for prediction of NitroPAH mutagenicity.

1.2.Other related models:

P. Gramatica, P. Pilutti, E.Papa. Approaches for externally validated QSAR modelling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity, SAR and QSAR in Environmental Research (18) 1-2, 2007, 169-178. [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:

06/12/2013

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:

Salmonella typhimurium

3.2.Endpoint:

4.Human health effects 4.10.Mutagenicity

3.3.Comment on endpoint:

The mutagenicity potency in TA100 (without the S9 activation system) for the 48 modelled nitro-PAHs was obtained from the Benigni Report for OECD [2].

3.4.Endpoint units:

No information available

3.5.Dependent variable:

log TA100

3.6.Experimental protocol:

Mutagenicity assay: the bacterial Ames test in *Salmonella typhimurium* TA100 strain

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:

logTA100 PC1 Split model

OLS-MLR method. Model developed on a training set of 33 compounds

logTA100 Ordered Response Split model

OLS-MLR method. Model developed on a training set of 33 compounds

logTA100 Full model

OLS-MLR method. Model developed on a training set of 48 compounds

PC1 Split model equation: logTA100= -5.96 + 0.80 C3SP2 + 6.81 maxHaaCH

Ordered Response Split model equation: logTA100= -6.22 + 0.83 C3SP2 + 6.89 maxHaaCH

Full model equation: logTA100= -6.23 + 0.80 C3SP2 + 7.17 maxHaaCH

4.3.Descriptors in the model:

[1]C3SP2 Doubly bound carbon bound to three other carbons
[2]maxHaaCH Maximum atom-type H E-State: :CH:

4.4.Descriptor selection:

A total of 1605 molecular descriptors of differing types (0D, 1D, 2D, Fingerprints) 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 0.95 was removed to reduce redundant information), and a final set of 74 molecular descriptors were used as input variables for variable subset selection. The models were developed by the all-subset-procedure. 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: 33 chemicals / 2 descriptors = 16.5
Split by Ordered Response model: 33 chemicals / 2 descriptors= 16.5

Full model: 48 chemicals / 2 descriptors = 24

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 logTA100 values: -2.1 / 4.74
Range of descriptor values: C3SP2: 0 / 6 ; maxHaaCH 0.55 / 0.86

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*=0.188). 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>0.273 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: 1,3,6,8-TETRANITROPYRENE (28767-61-5). **Ordered Response Split model domain**: outliers for structure, hat>0.273 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: 1,3,6,8-TETRANITROPYRENE (28767-61-5). **FULL model domain**: outliers for structure, hat>0.188 (h*): no. Outliers for response, standardised residuals > 2.5 standard deviation units: 2-nitroanthracene (3586-69-4).

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=48) 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 structural similarity (ordering PC1 Score after a PCA analysis) and by ordered response (n training= 33 in both cases).

6.6.Pre-processing of data before modelling:

No information available.

6.7.Statistics for goodness-of-fit:

PC1 Score Split model:

R^2= 0.83; CCCtr [3]=0.91; RMSE= 0.71

Ordered response split model:

R^2= 0.85; CCCtr=0.92; RMSE= 0.71

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

PC1 Score Split model:

Q^2LOO= 0.78; CCCcv=0.88; RMSEcv= 0.79

Ordered response Split model:

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

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

PC1 Score Split model:

Q^2LMO= 0.78

Ordered response split model:

Q^2LMO= 0.80

6.10.Robustness - Statistics obtained by Y-scrambling:

PC1 Score Split model:

R^2y-sc= 0.06

Ordered response split model:

R^2y-sc= 0.06

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=48) 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 structural similarity (ordering chemicals by PC1 Score, n external validation set =15) and sorted response (n external validation set =15); the range of logTA100 are: -2.1 / 4.09 for PC1 score prediction set, -1.3 / 3.87 for Ordered Response prediction set.

7.6.Experimental design of test set:

In the case of split by sorted response model, chemicals were ordered according to their increasing activity, and one out of every three chemicals was put in the prediction set (always including the most and the least active compounds in the training set). The splitting based on structural similarity (by ordered PC1 Score, after a PCA analysis) allowing the selection of a structurally meaningful training set and an equally representative prediction set.

7.7.Predictivity - Statistics obtained by external validation:

PC1 Split model:

Q^2extF1 [4]= 0.83; Q^2extF2 [5]= 0.83; Q^2extF3 [6]= 0.82; CCCex=0.91; RMSE= 0.73

Ordered response split model:

Q^2extF1= 0.76; Q^2extF2= 0.76; Q^2extF3= 0.83; CCCex=0.88; RMSE= 0.75

7.8.Predictivity - Assessment of the external validation set:

The splitting methodology based on PC1 Score and by Ordered response 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 logTA100 values are [-1.34 / 4.74] [-2.1 / 4.09] and [-2.1 / 4.74][-1.3 / 3.87] respectively for PC1 Score and Ordered Response training and prediction sets.

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

C3SP2: PC1 Split training set (0/ 6), prediction set (1 / 6); Ordered response split training set (0 / 6), prediction set (1 / 6)

maxHaaCH : PC1 Split training set (0.55 / 0.86), prediction set (0.59 / 0.84); Ordered response split training set (0.58 / 0.86), prediction set (0.55 / 0.84)

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 Gramatica et al. [7] is:

$$\log TA100 = -59.07 + 2.61\ CIC1 + 92.75\ PW2$$

where CIC1: complementary information content (neighborhood symmetry of 1-order). This descriptor, which is the most relevant and positively related to mutagenicity, gives information on molecular size and increases with the number of rings and nitrogroups in each series of congeners.

PW2: path/walk 2-Radic shape index. This topological descriptor, directly correlated to activity, gives information related to the shape of the molecules. Chemicals with a less linear, more round (circular) shape appear the most active.

It is again verified that the number of nitro groups increases the mutagenicity.

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

$$\log TA100 = -6.23 + 0.80\ C3SP2 + 7.17\ maxHaaCH$$

where C3SP2= Doubly bound carbon bound to three other carbons

maxHaaCH= Maximum atom-type H E-State: :CH:

The correlation between CIC1 and C3SP2 is 87%, suggesting that these two descriptors have similar meaning in the modelling of nitro PAH mutagenicity.

8.3.Other information about the mechanistic interpretation:

no other information available

9.Miscellaneous information

9.1.Comments:

To predict logTA100 for new NitroPAHs without experimental data, it is suggested to apply the equation of the Full Model, developed on all the available chemicals (N=48), thus ensuring a wider applicability domain. The full model equation (reported also in section 4.2) and the statistical parameters are the following:

logTA100= -6.23 + 0.80 C3SP2 + 7.17 maxHaaCH

N = 48; R^2 = 0.83; Q^2 = 0.80; Q^2LMO = 0.80; CCC = 0.91; CCCcv = 0.89; RMSE= 0.71; 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]ENV/JM/MONO(2004)24. Available online at: http://appli1.oecd.org/olis/2004doc.nsf/linkto/env-jmmono, 24 (2004).

[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]P. Gramatica, P. Pilutti, E.Papa. Approaches for externally validated QSAR modelling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity, SAR and QSAR in Environmental Research (18) 1-2, 2007, 169-178.

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