QMRF identifier (JRC Inventory): To be entered by JRC	
QMRF Title: Nonlinear QSAR model for acute oral toxicity of rat	
Printing Date: 5.04.2011	

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear QSAR model for acute oral toxicity of rat

1.2.Other related models:

1.3.Software coding the model:

QSARModel 3.3.8; Statistica 7, StatSoft Ltd. Turu 2, Tartu, 51014, Estonia, http://www.molcode.com

2.General information

2.1.Date of QMRF:

19.12.2010

2.2.QMRF author(s) and contact details:

Dimitar Dobchev, Tarmo Tamm, Gunnar Karelson, Indrek Tulp, Kaido Tämm, Jaak Jänes, Eneli Härk, Andres Kreegipuu, Mati Karelson, Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Molcode model development team Molcode Ltd Molcode Ltd Turu 2, Tartu, 51014, Estonia models@molcode.com www.molcode.com

2.6.Date of model development and/or publication:

19.12.2010

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

[1]Katritzky, A. R.; Dobchev, D. A.; Fara, D. C.; Hur, E.; Tämm, K.; Kurunczi, L.; Karelson, M.; Varnek, A.; Solov'ev, V. P. (2006). Skin Permeation Rate as a Function of Chemical Structure . Journal of Medicinal Chemistry, 49(11), 3305 - 3314.

[2]Karelson, M.; Dobchev, D. A.; Kulshyn, O. V.; Katritzky, A. (2006). Neural Networks Convergence Using Physicochemical Data. Journal of Chemical Information and Modeling, 46, 1891 - 1897.

[3]Statistica 7 www.statsoft.com

2.8.Availability of information about the model:

All information in full detail is available

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

No other QMRF available for the same model

3.Defining the endpoint - OECD Principle 1

3.1.Species:

rat

3.2.Endpoint:

4.Human health effects 4.2.Acute oral toxicity

3.3.Comment on endpoint:

The acute oral toxicity is determined using the OECD 423 (EU B.1 tris) test guideline. Acute oral toxicity testing allows to obtain the information on the biologic/toxic activity of a chemical. Currently, the basis for toxicologic classification of chemicals is the median lethal dose (LD50, mg/kg b.w.), which is defined as the statistically derived dose required to kill half the members of a tested population. Animals are observed individually after dosing at least once during the first 30 minutes, periodically during the first 24 hours, with special attention given during the first 4 hours, and daily thereafter, for a total of 14 days.

3.4.Endpoint units:

LD50 [mmol/kg]

3.5.Dependent variable:

Log LD50

3.6.Experimental protocol:

Acute oral LD50 data were taken from the Registry of Toxic Effects of Chemical Substances (RTECS) database. RTECS is a compendium of data extracted from the open scientific literature. The data are recorded in the format developed by the RTECS staff and arranged in alphabetical order by prime chemical name. Specific numeric toxicity values such as LD50, LC50, TDLo, and TCLo are noted as well as species studied and route of administration used. For each citation, the bibliographic source is listed thereby enabling the user to access the actual studies cited. The RTECS is a toxicology database of over 168,000 chemicals compiled, maintained, and updated by the U.S. National Institute of Occupational Safety and Health (NIOSH). The LD50 values of tested substances were translated to logarithmic scale (logLD50) to reduce the range of the data.

3.7.Endpoint data quality and variability:

235 diverse compounds were used in this report. Experimental data from different sources has been validated as consistent [1]

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression

4.2.Explicit algorithm:

The algorithm is based on regression neural network predictor with structure 9-5-1

4.3.Descriptors in the model:

[1]ALFA polarizability (DIP) (AM1)
[2]Avg electrophilic reactivity index (AM1) for O atoms
[3]Bonding Information content (order 0)
[4]Charged (AM1) Surface Area of Cl atoms
[5]Charged (AM1) Surface Area of N atoms
[6]Final heat of formation (AM1)
[7]Gravitation index (all atom pairs) (AM1)
[8]HA dependent HDCA-2/SQRT(TMSA) (AM1)
[9]Highest atomic state energy (AM1) for N atoms

4.4.Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules as F statistic and p. The first highest F (low p) descriptors (9) were selected from the whole descriptors pool. These 9 descriptors were used as inputs to the network. 57 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error) for training, selection and test sets. Then 89 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm encoded in the backpropagation scheme using linear and hyperbolic activation functions.

4.5.Algorithm and descriptor generation:

All descriptors were generated using QSARModel on structure optimized by AM1 semiempirical quantum mechanical model.

4.6.Software name and version for descriptor generation:

QSARModel
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

17

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:
a)diverse set organic compounds (phenols, carboxylic acids, carbonyl-, nitro-, amino-, halogeno-, P-conataining derivatives, etc)
b)The model is suitable for compounds that
have descriptors values in the followin range;

Desc	1	2	3	4	5	6	7	8	9
min	10.59	0.00	2.37	0.00	0.00	-345.84	155.19	0.00	-188.86
max	582.27	0.03	27.62	22.55	124.42	461.31	17030.22	0.49	0.00

5.2.Method used to assess the applicability domain:

presence of functional groups in structures
Range of descriptor values in training set with ±30% confidence

Descriptor values must fall between maximal and minimal descriptor values (see5.1) of training set ±30%.

5.3.Software name and version for applicability domain assessment:

QSARModel 3.3.8
http://www.molcode.com

5.4.Limits of applicability:

See 5.1, 5.2

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:No
Formula:No
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:

data points: 155

6.6.Pre-processing of data before modelling:

Standardization and normalization of the inputs by taking into account the min max descriptor values

6.7.Statistics for goodness-of-fit:

	Training LogLD50	Selection LogLD50	Test LogLD50
Data Mean	0.64	0.42	0.49
Data S.D.	0.95	1.00	0.91
Error Mean	0.00	0.05	0.17
Error S.D.	0.73	0.74	0.76
Abs E. Mean	0.53	0.57	0.55
S.D. Ratio	0.77	0.74	0.83
Correlation	0.76	0.76	0.75

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

See 6.7

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

RMS (Training)= 0.149,, RMS(Selection)= 0.151, RMS(Test) = 0.158,

In this ANN were used 2 sets randomly chosen (40) to train and test the network – selection set and test set, see also 6.7

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: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 method used two validation sets – selection (40) and test (40)

7.6.Experimental design of test set:

Randomly selected 40 and 40 data points

7.7.Predictivity - Statistics obtained by external validation:

see 6.7 and 6.12

7.8.Predictivity - Assessment of the external validation set:

The descriptors for the test set are in the limit of applicability, see 6.7 and 6.12

7.9.Comments on the external validation of the model:

Overall predictions for the selection set (used to stop the ANN training and not to overfit it) and the test set (used to test the external prediction of the net after training) are significant according to the RMS error and the standard deviation ratio (S.D.Ration), see 6.7 and 6.12

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Since the ANN is a more complex predictor than a linear model, it is difficult to analyze the direct relation between the property and the descriptors. However, general trends and analysis can be drawn based on the most significant

descriptors in the net. According to these and the generally accpeted scientific understanding , the acute oral toxicity is strongly dependent on the stability and reactivity of chemicals, in particular the presence of heteroatoms like oxygen, nitrogen, phosphorus, and halogenides. One of the most important descriptor for this set was charged surface area of the N atoms. There is slight indication that the more charged is the N

the more toxic is the compound. The same

holds for charged surfacte area for Cl atoms. The hydrogen acceptor/donor ability of O and their reactivity indeces are encoded in HA dependent HDCA-2/SQRT(TMSA) (AM1)

and Avg electrophilic reactivity index (AM1) for O atoms descriptors indicating oposite relation with LogLD50. These parameters are likely to be related to solubility and LogP of the compounds making them less toxic.

Structural parameters as Bonding Information content (order 0), Gravitation index (all atom pairs) (AM1) define that the compound complexation and its bulk characteristics are aslo important for LD, especially

for molecular phenomena in liquid media.

8.2.A priori or a posteriori mechanistic interpretation:

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

Supporting information for :Training set(s)

Selection set(s)

Test set(s)

9-5-1.snn file -includes the ANN model, in order to be used the user must have statistica 7 or higher with ANN modules to make predictions.

9.2.Bibliography:

Guidance Document on Using In Vitro Data to Estimate In Vivo Starting Doses for Acute Toxicity http://iccvam.niehs.nih.gov/docs/acutetox_docs/guidance0801/iv_guide.pdf

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