

	QMRF identifier (ECB Inventory): To be entered by JRC	
	QMRF Title: Nonlinear ANN QSAR Model for Acute oral toxicity - in vitro method (registry of cytotoxicity)	
	Printing Date: Apr 29, 2010	

1. QSAR identifier

1.1. QSAR identifier (title):

Nonlinear ANN QSAR Model for Acute oral toxicity - in vitro method (registry of cytotoxicity)

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:

21.04.2010

2.2. QMRF author(s) and contact details:

Dimitar Dobchev, Tarmo Tamm, Gunnar Karelson, Indrek Tulp, Dana Martin, Kaido Tämm, Deniss Savchenko, 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:

12.04.2010 The methodology and software (QSARModel) used to create the present model were

applied also to obtain the results published in these papers.

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.

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

[4]

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, cell lines

3.2. Endpoint:

4. Human health effects 4.2. Acute oral toxicity

3.3. Comment on endpoint:

The European Chemicals regulation REACH (registration, evaluation, authorisation of chemicals) describes possible applications of in vitro alternatives to replace in vivo testing. Demanding validation procedures are applied by European Centre for the Validation of Alternative Methods (ECVAM) to assure the suitability of any new in vitro method for regulatory purposes. This procedure has led to a number of internationally accepted in vitro alternatives for local endpoints (e.g. skin penetration and corrosion). However, in vitro alternatives for endpoints with complex physiology have shown important limitations proving difficulties to validate for a broad spectrum of chemicals.

3.4. Endpoint units:

IC50 [mg/l]

3.5. Dependent variable:

Log (IC50)

3.6. Experimental protocol:

The acute oral toxicity is determined using the OECD 420 and OECD 423 (EU B.1.bis. and 1.B.tris.) test guidelines. 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.

Developed in vitro systems (biokinetic, cellular and molecular elements) can significantly reduce the number of animals needed for the toxicity testing of a broad range of compounds. There are 50 compounds in the dataset of the in vitro cytotoxicity (median of several IC50 values of different experiments) consisting of drugs, agrochemicals and industrial chemicals. The IC50 values of tested substances were translated to logarithmic scale (logIC50) to reduce the range of the data

3.7. Endpoint data quality and variability:

Experimental data from different sources has been validated as consistent (ref.
<http://dx.doi.org/10.1016/j.toxlet.2007.03.008>
http://iccvam.niehs.nih.gov/docs/acutetox_docs/guidance0801/iv_guide.pdf

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 neural network predictor with structure 3-3-2-1
Standard Backpropagation Neural Network (Multilayer Perceptron) regression

4.3. Descriptors in the model:

- [1]Schultz principal eigenvalue of D matrix
- [2]Schultz principal eigenvalue of A+D matrix
- [3]Number of rings
- [4]

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 (3) were selected from the whole (~1000) descriptors. These 3 descriptors were used as inputs to the network. 18 networks with different structures were tested in order to find the best ANN with lowest RMS(root-mean-squared error). Then 197 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with Levenberg-Marquardt algorithm using logistic activation function.

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

4.7. Chemicals/Descriptors ratio:

16.66

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

Applicability domain based on training set:

By descriptor value range (between min and max values): The model is suitable for compounds that have the descriptors in the following range:

Desc ID

See 4.3123

Min2.00003.00000.000000

Max488.6895490.60015.000000

5.2. Method used to assess the applicability domain:

presence of functional groups in structures

Range of descriptor values in training set with $\pm 30\%$ confidence

Descriptor values must fall between maximal and minimal descriptor values (see 5.1) of training set $\pm 30\%$.

5.3. Software name and version for applicability domain assessment:

QSARModel 3.3.8; Statistica 7, StatSoft Ltd.

Turu 2, Tartu, 51014, Estonia

<http://www.molcode.com>

5.4. Limits of applicability:

See 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: 35

6.6. Pre-processing of data before modelling:

Standardization and normalization by taking into account the mean and standard deviation

6.7. Statistics for goodness-of-fit:

Training $\log(\text{IC}_{50})$ Selection $\log(\text{IC}_{50})$ Test $\log(\text{IC}_{50})$

Data Mean 0.4950.7900.573

Data S.D. 1.8701.7211.851

Error Mean 0.001-0.168-0.198

Error S.D. 0.8480.5710.472

Abs E. Mean 0.6190.5370.398

S.D. Ratio 0.4540.3320.255

Correlation 0.8910.9780.970

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

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) = 0.120004, RMS (Selection) = 0.084220, RMS (Test) = 0.072431,
See 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 (5) and test(10)

7.6.Experimental design of test set:

Randomly selected 10 and 5 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 quite high according to the Pearson correlation coefficient 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:

The most statistically significant descriptors (among ~1000) are those 3 selected for inputs of the ANN model. The descriptors Schultz principal eigenvalue of D and D+A are unique for a given compound they can be used to characterize the compounds structural complexation related to the IC50. These two descriptors are unique for each compound in the set thus they can be used to characterize the compounds IC50. However, deep mechanistic interpretations for ANN models is difficult.

8.2.A priori or a posteriori mechanistic interpretation:

a posteriori mechanistic interpretation, consistent with published scientific interpretations of experiments.

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)

ANN.snn file -includes the ANN model, in order to be used the user must have statistical software version 7 or higher with ANN modules to make predictions.

9.2.Bibliography:

9.3.Supporting information:

Training set(s)

Acute_oral_tox1_IC50_20100427_trainingset.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf202_Acute_oral_tox1_IC50_20100427_trainingset.sdf
---	---

Test set(s)

Acute_oral_tox1_IC50_20100427_testset.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf202_Acute_oral_tox1_IC50_20100427_testset.sdf
---	---

Supporting information

3-3-2-1.snn	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf202_3-3-2-1.snn
Acute_oral_tox1_IC50_20100427_selectionset.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf202_Acute_oral_tox1_IC50_20100427_selectionset.sdf

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