QMRF	*QMRF identifier (ECB Inventory):* To be entered by JRC
	QMRF Title: Nonlinear ANN QSAR Model for Acute toxicity for Daphnia magna-LC50(48h)
	Printing Date: Jan 19, 2012

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear ANN QSAR Model for Acute toxicity for Daphnia magna-LC50(48h)

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:

10.10.2010

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:

12.04.2010The 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

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:

Daphnia magna

3.2.Endpoint:

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

3.3.Comment on endpoint:

3.4.Endpoint units:

LC50 [mol/l]

3.5.Dependent variable:

Log (LC50)

3.6.Experimental protocol:

The acute toxicity to Daphnia was determined using the OECD 203 (EU C.2) test guideline. Acute toxicity for Daphnia is expressed as the median effective concentration EC50 (LC50). This is the concentration which immobilizes 50% of the Daphnia in a test batch within 48 h. Those animals which are not able to swim within 15 seconds after gentle agitation of the test batch are considered to be immobile. The concentrations of the substances are given in mol per litre (mol/L).

From the U.S. EPA database AQUIRE [1], 1067 acute toxicity values (48 h LC50) for the D. magna were collected for a total of 349 organic chemicals with at least one LC50 value per substance. Subsequently, 49 chemicals were excluded because their LC50 values exceeded the predicted water solubility or because they contained metal atoms or were inorganic, leading to the final set of 300 organic compounds that cover a log Kow (octanol/water partition coefficient) range from -2 to 8.

As regards the chemical domain, the data set includes hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives; isothiocyanates; thioles; phosphorothionate and phosphate esters; and halogenated derivatives [2].

3.7.Endpoint data quality and variability:

In the experimental procedure, when multiple test values were found for one substance, these values were checked for consistency. If values differed by more than a factor of 30 from the closest one in a group of at least three other references, the aberrant value was discarded so as to remove outliers from the data set. Of all the remaining values for a given substance, the arithmetic mean was taken as the valid experimental value.

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 9-9-8-1.

4.3.Descriptors in the model:

[1]Kier&Hall index (order 3)
[2]Kier&Hall index (order 1)
[3]Gravitation index (all atom pairs) (AM1)
[4]Molecular weight
[5]ALFA polarizability (DIP) (AM1)
[6]Randic index (order 3)
[7]Michalic MTI' of Schultz triple weighted D matrix
[8]Molecular surface area (AM1)
[9]Schultz average vertex TLF CIRS (graph Nmax)

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 (~997) descriptors. These 9 descriptors were used as inputs to the network. 15 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error) and highest correct predictions (for training, selection and test sets). Then 199 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 3.3.8; Statistica 7, StatSoft Ltd.
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

21.6

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 (including hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives; isothiocyanates; thioles; phosphorothionate and phosphate esters; and halogenated derivatives) that have the descriptors in the following range augmented with the confidence in 5.2:

Desc ID

1
2
3
4
5
6
7
8
9
Min 0 0.72361
261.034 41.0519
15.467
0 198
59.6 1 Max 8.946869
10.5558
7859.72 505.1992
237.691
10.75178 1145680 402.2 23.5574

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; Statistica 7, StatSoft Ltd.

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: 195

6.6.Pre-processing of data before modelling:

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

6.7.Statistics for goodness-of-fit:

Training log(LC50)
Selection log(LC50)
Test log(LC50)
Data Mean
-4.66646
-4.74560
-4.71120
Data S.D.
1.74879
1.95371
1.88120
Error Mean
-0.00122 -0.01980 -0.16642
Error S.D. 0.90616
1.95400
1.06359 Abs E. Mean
0.68683 1.19990 0.80214 S.D. Ratio
0.51817
1.00015 0.56538
Correlation
0.85528 0.57496 0.82851

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)= 0.094098, RMS(Selection)= 0.202919, RMS(Test) = 0.111789,

In this ANN were used 2 sets randomly chosen (50) to 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 (50) and test (50)

7.6.Experimental design of test set:

Randomly selected 50 and 50 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 relation between the property and the descriptors. Most of the descriptors are related to the structural complexity of the molecules as well as their geometrical and surface properties. However, one could make rough estimation based on their values. Regarding the most significant descriptor Kier&Hall index (order 3) it can be noted that it has negative correlation with the property (-0.7). It might suggest that with the increase of the descriptors, the property would decrease. The same holds for the Molecular surface, Gravitational index and polarizability descriptors (correlation -0.71, -0.69, -0.69, respectively).

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)
9-9-8-1.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:

[1]1. U.S. Environmental Protection Agency (2002) AQUIRE (Aquatic Toxicity Information Retrieval Database), National Health and Environmental Effects Research Laboratory, Duluth, MN.
[2]von der Ohe P. C., Kühne R., Ebert R.-U., Altenburger R., Liess M., and Schüürmann G. , Chem. Res. Toxicol. 2005, 18, 536-555.

9.3.Supporting information:

Training set(s)

Daphnia_Magna_2_48h_trainingset.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf269_Daphnia_Magna_2_48h_trainingset.sdf

Test set(s)

Daphnia_Magna_2_48h_testset.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf269_Daphnia_Magna_2_48h_testset.sdf

Supporting information

9-9-8-1.snn	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf269_9-9-8-1.snn
Daphnia_Magna_2_48h_selectionset.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf269_Daphnia_Magna_2_48h_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