1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear QSAR: artificial neural network for acute toxicity of birds

1.2.Other related models:

1.3.Software coding the model:

[1]QSARModel 3.3.8 Turu 2, Tartu, 51014, Estonia http://www.molcode.com
[2]Statistica 7 StatSoft Ltd. http://www.statsoft.com

2.General information

2.1.Date of QMRF:

10.10.2010

2.2.QMRF author(s) and contact details:

[1]Dimitar Dobchev Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[2]Tarmo Tamm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[3]Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[4]Indrek Tulp Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[5]Dana Martin Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[6]Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[7]Deniss Savchenko Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[8]Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[9]Eneli Härk Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[10]Mati Karelson 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 Turu 2, Tartu, 51014, Estonia models@molcode.com www.molcode.com

2.6.Date of model development and/or publication:

12.04.2010

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

[1]Katritzky A R, Dobchev DA, Fara DC, Hur E, Tämm K, Kuruncz L, Karelson M, Varnek A & Solov'ev VP (2006). Skin Permeation Rate as a Function of Chemical Structure. Journal of Medicinal Chemistry 49, 3305 - 3314.

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

2.8.Availability of information about the model:

Training, selection and test sets are available. Model algorithm is available (snn file).

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

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Birds - three avian bird species (redwing, starling, coturnix)

3.2.Endpoint:

3.Ecotoxic effects 12.Toxicity to birds 3.12.a.Short term toxicity (feeding, gavage, other)

3.3.Comment on endpoint:

LD50: The acute oral toxicity (LD50) of the compounds to one or more of three avian bird species (redwing, starling, coturnix) was measured in order to define the toxic magnitude.

3.4.Endpoint units:

LD50 [mg/kg]

3.5.Dependent variable:

Log (LD50)

3.6.Experimental protocol:

A data set of 250 compounds was preselected from [ref 1; sect 9.2] regarding the acute oral toxicity for the three types of birds i.e. crutonix, redwinged blackbird and starling. Wild-trapped birds were preconditioned to captivity for 2-6 weeks and were usually dosed by gavage with solutionor suspensions of the test chemical in propylene glycol, according to methods of [ref 2,3; sect 9.2]. LD50 values were calculated by the method of [ref 4,5; sect 9.2].

3.7.Endpoint data quality and variability:

The data are taken from single literature source assuming consistency and quality of the experiments [ref 1; sect 9.2]. 250 compounds were preselected from a total of 998. The selection was performed by availability of LD50 and compounds (salts, unclear chemical structures, missing appropriate names compounds were omitted).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR

4.2.Explicit algorithm:

QSAR
Nolinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression

The algorithm is based on regression neural network predictor with structure 7-5-5-1 (i.e. 7 neurons in the input layer, 5 neurons in the first hidden layer, 5 neurons in the second hidden layer and one neuron for the output layer responsible for logLD50)

4.3.Descriptors in the model:

[1]ALFA polarizability (DIP) (AM1)
[2]Average atom weight
[3]Gravitation index (all atom pairs) (AM1)
[4]HA dependent HDCA-2/SQRT(TMSA) (AM1)
[5]HASA-2/TMSA (AM1)
[6]HOMO - LUMO energy gap (AM1)
[7]Molecular surface area (AM1)

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 highest F (low p) descriptors (7) were selected from the whole set of descriptors (~1000) taking into account also their value distribution. These 7 descriptors were used as inputs to the network. 12 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error). Then 1000 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed with the Levenberg-Marquardt algorithm using hyperbolic and linear activation functions.

4.5.Algorithm and descriptor generation:

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

4.6.Software name and version for descriptor generation:

QSARModel 3.3.8
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

18 (126 chemicals / 7 descriptors)

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 augmented with the confidence in 5.2:

Desc ID, See 4.3: 1 2 3 4 5 6 7
Min: 10.5878; 5.39763; 155.19; 0.000000; 0.000000; 6.61232; 48.800
Max: 682.0362; 31.41740; 30692.38; 0.236888; 0.068331; 11.65324; 1033.840

5.2.Method used to assess the applicability domain:

Presence of functional groups in structures. The model is based on diverse set of compounds including functional groups such as, aliphatic hydrocarbons, aromatic, hetero aromatics, halogens, acids, nitro groups.

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: 126

6.6.Pre-processing of data before modelling:

Normalization of the inputs by taking into account the maximum and minimum of the descriptor values. Normal D = D-Dmin/Dmax-Dmin, where D is the descriptor value and Dmin, Dmax is the values depicted in 5.1

6.7.Statistics for goodness-of-fit:

Training log(LD50); Selection log(LD50); Test log(LD50)

Data Mean: -0.585077; -0.806710; -0.906958

Data SD: 0.961833; 1.003200; 0.858715

Error Mean: -0.000182; -0.094310; -0.281062

Error SD: 0.453416; 1.181115; 0.935703

Abs E. Mean: 0.347310; 0.972181; 0.729703

S.D. Ratio: 0.471408; 1.177348; 1.089655

Correlation: 0.881918; 0.529134; 0.543391

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)=0.095973; RMS(Selection)=0.250798; RMS(Test)= 0.206799

In this ANN, 2 randomly chosen sets (62) were used 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 (62) and test (62)

7.6.Experimental design of test set:

Randomly selected 62 selection and 62 test set 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.Ratio); see 6.7 and 6.12.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The toxic effects of chemicals on birds are not well understood. However, in regard with the best descriptors in the ANN model some rough conclusions can be made. According to the ALFA polarizability (DIP) (AM1) descriptor, a negative correlation with the LogLD50 (-0.46) suggests that the modelled property decreases with increasing polarizability of the compound. The same relation holds for the Gravitation index (all atom pairs) (AM1) descriptor and log LD50 (correlation of 0.5).

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).

The 7-5-5-1.snn file includes the ANN model; the user must have Statistica 7 or higher with ANN modules to make predictions.

9.2.Bibliography:

[1]Schafer EW, Bowles Jr WA & Hurlbut J (1983). The acute oral toxicity, repellency, and hazard potential of 998 chemicals to one or more species of wild and domestic birds. Archives of Environmental Contamination and Toxicology 12, 355-382.

[2]DeCino TJ, Cunningham DJ & Schafer EW (1966). Toxicity of DRC-1339 to starlings. Journal of Wildlife Management 30, 249-253.

[3]Schafer EW(1972). The acute oral toxicity of 369 pesticidal, pharmaceutical and other chemicals to wild birds. Toxicology and Applied Pharmacology 21, 315-330.

[4]Thompson WR(1948). Use of moving averages and interpolation to estimate median effective dose. Bacteriological Reviews 11, 115-145.

[5]Weft CS (1952). Tables for convenient calculation of median effective dose(LD~0 or EDs0) and instructions in their use. Biometrics 8, 249-263.

9.3.Supporting information:

Training set(s)

Acute_toxicity_birds_#1_training_126.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf333_Acute_toxicity_birds_#1_training_126.sdf

Test set(s)

Acute_toxicity_birds_#1_selection_62.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf333_Acute_toxicity_birds_#1_selection_62.sdf
Acute_toxicity_birds_#1_test_62.sdf	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf333_Acute_toxicity_birds_#1_test_62.sdf

Supporting information

7-5-5-1.snn	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf333_7-5-5-1.snn

10.Summary (JRC Inventory)

10.1.QMRF number:

Q19-22-1-333

10.2.Publication date:

2011/12/19

10.3.Keywords:

acute toxicity, birds, LD50, Molcode, artificial neural network

10.4.Comments:

To be entered by JRC