1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear QSAR: artificial neural network for acute toxicity to Daphnia magna

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:

21.04.2010

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]Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[3]Indrek Tulp Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[4]Dana Martin Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[5]Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[6]Deniss Savchenko Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[7]Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[8]Eneli Härk Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[9]Andres Kreegipuu 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

[11]Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

[12]Tarmo Tamm 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:

Statistica 7 www.statsoft.com

2.8.Availability of information about the model:

Training, selection and test sets available. Algorithm is available.

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

None to date.

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Water flea (Daphina magna)

3.2.Endpoint:

3.Ecotoxic effects 3.4.Long-term toxicity to Daphnia (lethality, inhibition of reproduction)

3.3.Comment on endpoint:

EU C.2, OECD TG 202, Acute toxicity for Daphnia (water flea) is expressed as the median effective concentration LC50. This is the concentration which immobilizes 50% of the Daphnia in a test batch within 96 h. The concentrations of the substances are given in millimol per litre (mmol/L). Those animals which are not able to swim within 15 seconds after gentle agitation of the test batch are considered to be immobile

3.4.Endpoint units:

LC50 [mmol/l]

3.5.Dependent variable:

-Log (LC50)

3.6.Experimental protocol:

3.7.Endpoint data quality and variability:

Experimental data from different sources has been validated as reliable (ref.Toropov et al (2006) (ref 1, section 9.2) (NB: wrong notation of Daphnia magna estimation 48h, the correct is 96h).

The toxicity data of Daphnia were selected from the EPA Office of Pesticide Programs (EPA-OPP) database according to Good Laboratory Practice (GLP) and availability of ancillary data such as purity, year of the study, uncertainty of the experimental result, and other statistical parameters.Toxicity data for EPA-OPP database are drawn from several sources and then reviewed: (i) Ecotoxicological studies conducted by commercial laboratories and submitted by pesticide companies in support of their products. EPAs Office of Compliance and Monitoring conducts periodic audits of these laboratories.(ii) Studies conducted by US-EPA, USDA, and USFWS laboratories over the last 25 years. (iii) Published data considered to meet their guideline criteria for acceptable data. Inorganic compounds and mixtures in which components have different molecular weight or connectivity (i.e., substances with different chemical identity) were eliminated from the original EPA dataset. However, mixtures of stereoisomers were kept, because they are super imposable using common 2D descriptors. Data for the Daphnia magna of acute toxicity EC50 96 h exposure were then pruned as follows: (i) Eliminating studies with an a.s.<85% purity. (ii) Those identified as invalid where invalid studies were defined by the EPA as studies which may not be scientifically sound, or they were performed under conditions that deviated so significantly from the recommended protocols that the results will not be useful in a risk assessment. (iii) Furthermore, only studies with actual values were kept discarding data given as higher or lower that values. The acute toxicity of Daphnia dataset consists of 262 pesticides, randomly splited into a training (n = 220) and a test (n = 42) set. QSAR models were developed using only chemicals in the training set. Results were validated using the test set.

References: Toropov et al (2006), Roncaglioni et al (2004), EPA US Environment Protection Agency, Pesticides (ref 1-3, sect 9.2), Benfenati E(2007) (ref 4, sect 9.2).

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Neural network

4.2.Explicit algorithm:

Neural network
Nonlinear QSAR: Backpropagation Neural Network (Multilayer Perceptron) regression
The algorithm is based on neural network predictor with structure 7-7-6-1.
Algorithm is available in the snn file. In order to be used the user must have Statistica 7 or higher with ANN modules to make predictions.

4.3.Descriptors in the model:

[1]HA dependent HDCA-1/TMSA (Zefirov)
[2]HA dependent HDCA-1 (Zefirov)
[3]HA dependent HDCA-2/SQRT(TMSA) (Zefirov)
[4]HA dependent HDCA-2/TMSA (Zefirov)
[5]HACA-2/SQRT(TMSA) (Zefirov)
[6]Lowest n-n repulsion (AM1) for P - S bonds
[7]Lowest e-e repulsion (AM1) for P - S bonds

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 fitness function relates minimization sum of squares coupled with the normality of the datapoints) This procedure affects the whole data points. . The first highest F (low p) descriptors (7) were selected from the whole (~1000) descriptors. These 7 descriptors were used as inputs to the network. 13 networks with different structures were tested in order to find the best ANN with lowest RMS (root-mean-squared error). Then 28 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
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

22.57 (158 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 |0.000| 0.000| 0.000| 0.000| 0.000 |0.000| 0.000|
Max |0.025| 9.432| 0.092| 0.005| 0.135| 197.428| 112.599|

5.2.Method used to assess the applicability domain:

Presence of functional groups in structures (such as methyl, esters, amides, amines, thio, pyrolo). The set concerns mainly pesticide compounds.

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 5.2

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: 158

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(LC50)	Selection -log(LC50)	Test -log(LC50)
Data Mean	2.546	2.711	2.584
Data S.D.	1.808	1.592	1.859
Error Mean	-0.011	-0.371	0.078
Error S.D.	1.293	1.433	1.261
Abs E. Mean	1.012	1.196	0.974
S.D. Ratio	0.715	0.900	0.678
Correlation	0.7701	0.7449	0.736

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) = 0.158040, RMS (Selection) = 0.180929, RMS (Test) = 0.154354, 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:No
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 randonly selected 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 Pearson correlation coefficient and the standard deviation ratio (S.D.Ratio) and RMS error, see 6.7 and 6.12

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The mechanistic picture of the model is complicated due to the nature of the ANN(artificial neural network). However, it is known that LC50 for Daphnia magna is greatly to the hydrogen acceptor/donor abilities of a compound. This is reflected by the first 5 descriptors of the current model. In addition the P atom containing compounds express larger Log LC50 values.

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 Statistica 7 or higher with ANN modules to make predictions.

The methodology and software (QSARModel) used to create the present model were applied also to obtain the results published in these papers: Katritzky et al (2006); Karelson et al (2006)

9.2.Bibliography:

[1]Toropov AA & Benfenati E (2006). QSAR models for Daphnia toxicity of pesticides based on combinations of topological parameters of molecular structures. Bioorganic & Medicinal Chemistry 14, 2779–2788.

[2]Roncaglioni A, Benfenati E, Boriani E & Clook M (2004). A Protocol to Select High Quality Datasets of Ecotoxicity Values for Pesticides. Journal of Environmental Science and Health, Part B, 39, 641–652.

[3]US Environment Protection Agency, Pesticides http://www.epa.gov/pesticides/

[4]Benfenati E (2007). Quantitative Structure-Activity Relationships for Pesticide Regulatory Purposes, Chapters 2 & 7. Elsevier. http://www.elsevier.com/wps/find/bookdescription.cws_home/710506/description#description

9.3.Supporting information:

Training set(s)

Acute_tox_daphniamagna_158	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf267_qmrf228_Acute_tox_daphniamagna_158_trainingset.sdf

Test set(s)

_Acute_tox_daphniamagna_40	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf267_qmrf228_Acute_tox_daphniamagna_40_testset.sdf
Acute_tox_daphniamagna_40_selection	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf267_qmrf228_Acute_tox_daphniamagna_40_selectionset.sdf

10.Summary (JRC Inventory)

10.1.QMRF number:

Q17-10-1-267

10.2.Publication date:

2010/10/08

10.3.Keywords:

Molcode, nonlinear QSAR, artificial neural network, acute toxicity, Daphnia magna

10.4.Comments: