QMRF identifier (JRC Inventory):	*To be entered by JRC*
QMRF Title:	*Nonlinear ANN QSAR model for Toxicity to algae*
Printing Date:	*24.10.2011*

1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear ANN QSAR model for Toxicity to algae

1.2.Other related models:

1.3.Software coding the model:

QSARModel 4.0.4; Statistica 7, StatSoft Ltd. Turu 2, Tartu, 51014, Estonia, http://www.molcode.com

2.General information

2.1.Date of QMRF:

03.10.2011

2.2.QMRF author(s) and contact details:

Dimitar Dobchev, Tarmo Tamm, Gunnar Karelson, Indrek Tulp, Kaido Tämm, Jaak Jänes, Eneli Härk, 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:

26.09.2011

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

[1]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]Statistica 7 www.statsoft.com

[3]Karelson M, Dobchev D, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Karelson G (2008). Correlation of blood-brain penetration and human serum albumin binding with theoretical descriptors. ARKIVOC 16, 38-60.

[4]Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Dobchev D (2009). QSAR study of pharmacological permeabilities. ARKIVOC 2, 218–238.

2.8.Availability of information about the model:

All data and modeling information is available

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Pseudokirschneriella subcapitata

3.2.Endpoint:

3.Ecotoxic effects 3. 2.Short-term toxicity to algae (inhibition of the exponential growth rate)

3.3.Comment on endpoint:

C.3 in REACH classification. The EC50 is the concentration (mM)

that induces toxicity response halfway between the baseline and maximum after 96 h.

3.4.Endpoint units:

mM

3.5.Dependent variable:

log(1/EC50) [or pEC50]

3.6.Experimental protocol:

The test alga was an unicellular green algal species Selenastrum capricornutum Printz (also known as Pseudokirschneriella subcapitata and Raphidocelis subcapitata) and the culture medium 10% Z 8. The inoculum was taken from a stock culture in the exponential growth phase.

The initial algal density was 104 ± 10% cells/mL. The test algae were cultivated in 100-mL solutions in 250-mL sterile, foam-plugged Erlenmeyer flasks with three replicates of each concentration. In addition, there were two control cultures: Selenastrum cells in culture medium and in acetone series. There were also controls for chemicals without algae.

The cultures were incubated at +22 ± 20 C in continuous illumination of approximately 72 µE m-2 s-1 (Airam L 40 W 35). The growth of cultures was followed by measuring the cell density after 24, 48, 72 and 96 hr by means of an electronic particle counter (Coulter Counter Z B).

The effect of acetone on the growth of the cultures was eliminated by comparing the growth of test cultures with the growth of acetone-controls. The results, as percent of control, were calculated as a mean

value of the cell density of the triplicates after one test series.
In Selenastrum assays, the EC50-values were estimated from semilogarithmic paper using cell density after 96 hr and areal comparison of growth curves during 0-96 hr incubation (ISO 1983). See refrences [1,2]

3.7.Endpoint data quality and variability:

Experimental data from a number of different publications wasused, as assembled in publication listed in 9.2

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 4-4-1
Standard Backpropagation Neural Network (Multilayer Perceptron) regression using Levenberg-Marquardt optimization algorithm

4.3.Descriptors in the model:

[1]Square root of Charged (Zefirov) Surface Area of C atoms [A] Angstrom
[2]Molecular weight [amu]
[3]Globularity index (AM1) -
[4]HOMO - LUMO energy gap (AM1) [eV]

4.4.Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on the highest correlation with the property followed with varinace cleaning (small variance of the descriptor 10^-6). The best 10 descriptors were selected and used further for the ANN selection. 18 networks with different structures and descriptors were tested in order to find the best ANN with lowest RMS (root-mean-squared error) for training, selection and test sets. Then 403 epochs were used to train the final network with architecture depicted in 4.2. Optimization of the weights was performed by Levenberg-Marquardt algorithm using linear(inputs) and hyperbolic(hidden) and logistic(output) 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 4.0.4
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

according to the training set = 12

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 (small organic molecules with functional groups as halogens, nitro, hetero benzens, alcohols) that have descriptors in the following range augmented with the confidence in 5.2:

(the following is in table format - first row Descs IDs, second row- min desc values, third row - max desc values)

Desc 1 2 3 4
min 0.00 45.08 0.62 5.08
max 1.34 379.66 1.04 13.37

5.2.Method used to assess the applicability domain:

Quantitative approach - range of descriptor values in training set with augmented with ±30% confidence

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

5.3.Software name and version for applicability domain assessment:

QSARModel 4.0.4
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: 48

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

(the following information is in Table format 7 rows and 4 columns)

Training pEC50 Selection pEC50 Test pEC50
Data Mean 1.94 1.56 2.10
Data S.D. 1.24 1.09 1.16

Error S.D. 0.63 0.59 0.68
Abs E. Mean 0.45 0.51 0.60
S.D. Ratio 0.51 0.54 0.59
Correlation 0.86 0.84 0.81

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)= 0.11
, RMS(Selection)= 0.10
, RMS(Test) = 0.12

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 randonly selected validation sets – selection (10) and test(10)

7.6.Experimental design of test set:

Randomly selected 10 and 10 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:

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 over fit it) and the test set (used to test the external prediction of the net after training) are significant according to 6.7.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The descriptors in the ANN were selected by the method described in section 4.4 which selects features with highest correlation coefficient in respect to the property. Further the combinatorial construction of various ANN topologies led to selection of 4 descriptors.

This is a perfect example where the "linearity" of the descriptors can be further extended to nonlinear relation with the property based on the ANN. The descriptor Square root of Charged (Zefirov) Surface Area of C atoms has positive correlation with pEC50 indicating that the increase of the descriptor would lead to decrease of EC50. The C atom is "more" charged where its neighbours are hetero atoms as O, N (also halogens). It is likely that this descriptor contribute to membrane permeability and polar narcosis. Similar analogy can be done for the HOMO - LUMO energy gap (AM1) descriptor, which is related the reactivity of the atoms. However, in this case this descriptor has negative correlation with the property. The most reactive centers in the molecule would probably react with the medium or the membrane and would change (decay) the compound making it less toxic to Algae. The remaining two descriptors Globularity index (AM1) and Molecular weight are related to the bulk properties of the compounds. These two descriptors can be addressed to features governing the nonspecific interactions describing nonpolar narcosis.

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)

9.2.Bibliography:

[1]K. Kuivasniemi, V. Eloranta, and J. Knuutinen, Arch. Environ. Contam. Toxicol. 14, 43-49 (1985)
[2]TR 091 - ECETOC Aquatic Toxicity (EAT) database, 2003.
[3]

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