1.QSAR identifier

1.1.QSAR identifier (title):

Nonlinear QSAR: artificial neural network for biodegradation: activated sludge respiration inhibition test

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]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.comv 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]Andres Kreegipuu Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[11]Mati Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[12]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

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

2.8.Availability of information about the model:

Training, selection and test sets 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:

Genetically modified luminescent bacterium Shk1 whose original strain was a Pseudomonad isolated from the activated sludge in an industrial wastewater treatment plant.

3.2.Endpoint:

[1]6.Other 6.6.Other

[2]3.Ecotoxic effects 3.6.Microbial inhibition (activated sludge respiration inhibition, inhibition of nitrification, other)

3.3.Comment on endpoint:

EU C.11, OECD TG 209, QMRF 3.6; This method assesses the effect of a test substance on micro-organisms by measuring the respiration rate under defined conditions in the presence of different concentrations of the test substance. The activated sludge respiration inhibition test provides a rapid screening and identification of water soluble substances which may adversely affect aerobic microbial treatment plants. Toxicants can cause upsets in the operations of an activated sludge process (decrease in waste organics removal, reduction of solids separation efficiency etc). The EC50 is the concentration of the test substance at which the respiration rate is 50% of that shown by the control under conditions described in this method.

3.4.Endpoint units:

EC50 [mg/l] - the concentration of toxicant required to reduce the emission of bioluminescence by 50%

3.5.Dependent variable:

Log (1/EC50)

3.6.Experimental protocol:

An alternative method for monitoring of activated sludge respiration inhibition uses an assay based on genetically modified luminescent bacterium Shk1 whose original strain was a Pseudomonad isolated from the activated sludge in an industrial wastewater treatment plant. Recent studies have demonstrated that the response of Shk1 to toxicants, due to the microorganism's origin in a wastewater treatment plant, is closer to that of the activated sludge microbial community than Vibrio fischeri, a marine bacterium used in the Microtox assay. Using the Shk1 assay, EC50 is the concentration of the toxicant required to reduce the emission of the bioluminescence of 50%. The experimental setup consisted of a

continuous stirred tank reactor (CSTR) operated at steady state. The operating conditions of the CSTR in this study were a dilution rate of 0.52 h-1 , pH=7 and T=220C. The CSTR was used to supply Shk1 cells continuously. The Shk1 cells were alternately mixed with buffer (pH=7) or test solutions at a 1:3 ratio (volume). The bioluminescence of Shk1 was measured after exposing Shk1 cells to buffer or test solutions for 5 min. The EC50 values were converted to mmol/L.

3.7.Endpoint data quality and variability:

The activated sludge respiration inhibition dataset was randomly split into a training (n = 68), selection (n=15) and a test (n = 15) set. ANN QSAR models were developed using only chemicals in the training set. Results were validated using the test set. Reference: Ren et al (2002). (ref 3, sect 9.2)

Experimental data from different sources have been validated as reliable (ref 3 sect 9.2). The primary data were collected from different sources.

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 a neural network predictor with structure 7-5-4-1.

4.3.Descriptors in the model:

[1]Solvent accessible surface area (AM1)
[2]RNCG Relative negative charge (QMNEG/QTMINUS) (Zefirov)
[3]RPCG Relative positive charge (QMPOS/QTPLUS) (Zefirov)
[4]Average Bonding Information content (order 0)
[5]Number of different functional groups
[6]Difference (Pos - Neg) in Charged Surface Areas (Zefirov)
[7]DPSA1 Difference in CPSAs (PPSA1-PNSA1)

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 (min F =7, max p =0.000007). The first highest F (low p) descriptors (7) were selected from the whole (~1000) descriptors. These descriptors 7 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 144 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 logistic activation function.

4.5.Algorithm and descriptor generation:

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

4.6.Software name and version for descriptor generation:

QSARModel
http://www.molcode.com

4.7.Chemicals/Descriptors ratio:

5.5 (68 chemicals/ 8 discriptors)

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 4.3): 1 2 3 4 5 6 7

Min 159.7814 0.137097 0.059140 0.191757 0.000000 -255.095 -255

Max 411.2549 1.000000 1.000000 0.960964 2.000000 347.306 348.306

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 of training set ±30%.

5.3.Software name and version for applicability domain assessment:

QSARModel 3.3.8

http://www.molcode.com

DRAGON

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Prof. R.Todeschini - distributed by Talete srl, via Pisani 13, 20124 Milano, Italy

http://www.disat.unimib.it/chm

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: 68

6.6.Pre-processing of data before modelling:

Standardization and normalization by taking into account the mean and standard deviation. A few compounds (4) were not included because they were salt mixtures.

6.7.Statistics for goodness-of-fit:

Training log (1/EC50); Selection log (1/EC50); Test log (1/EC50);

Data Mean -0.163 -0.264 -0.133
Data S.D. 1.530 0.856 0.724
Error Mean 0.00004 0.112 -0.172
Error S.D. 0.391 0.476 0.487
Abs E. Mean 0.311 0.423 0.390
S.D. Ratio 0.255 0.556 0.673
Correlation 0.967 0.840 0.742

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) = 0.029, RMS (Selection) = 0.036, RMS (Test) = 0.03, 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 validation sets (randomly chosen) – selection (15) and test (15)

7.6.Experimental design of test set:

Randomly selected 15 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:

The predictivity of 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), expressed according to the Pearson correlation coefficient and the standard deviation ratio (S.D.Ratio), are given in sections 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 7 selected for inputs of the ANN model. The descriptor Difference (Pos - Neg) in Charged Surface Areas (Zefirov) reflects transfer of the compounds to a phase characterized by the charge distributions related to the electrostatic interactions. This conjecture suggests that electrostatic interactions bound to different areas are very important for EC50. In addition Solvent accessible surface descriptor indicates that solubility of the compound also plays major role in this phenomenon.

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:

The complexity of the compound is related to EC50. This fact is reflected by the two descriptors Average Bonding Information content (order 0) and Number of different functional groups (ref:5 sect 9.2).

9.Miscellaneous information

9.1.Comments:

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]Katritzky AR, Dobchev DA, Fara DC, 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, 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.

[3]Ren S & Frymier PD (2002). Estimating the toxicities of organic chemicals to bioluminescent bacteria and activated sludge. Water Research 36, 4406–4414. http://dx.doi.org/10.1016/S0043-1354(02)00153-7

9.3.Supporting information:

Training set(s)

Act_Sludge_tox_68_trainingset	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf226_Act_Sludge_tox_68_trainingset.sdf

Test set(s)

Act_Sludge_tox_15_testset	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf226_Act_Sludge_tox_15_testset.sdf
Act_Sludge_tox_15_selectionse	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf226_Act_Sludge_tox_15_selectionset.sdf

10.Summary (JRC Inventory)

10.1.QMRF number:

Q17-10-1-226

10.2.Publication date:

2010/07/16

10.3.Keywords:

Molcode, nonlinear QSAR, artificial neural network, biodegradation, activated sludge, respiration inhibition test

10.4.Comments: