	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **Algae toxicity model (ProtoQSAR/COMBASE) (Version 1.0.0)**
	Printing Date: **3-mag-2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Algae toxicity model (ProtoQSAR/COMBASE) (Version 1.0.0)

1.2.Other related models:

None

1.3.Software coding the model:

Algae toxicity model (ProtoQSAR/COMBASE) V 1.0.0

The model is based on the OECD 201, Freshwater Alga and Cyanobacteria, Growth Inhibition Test. Test data provides a quantitative prediction of toxicity in algae.

www.vegahub.eu

2.General information

2.1.Date of QMRF:

April 2019

2.2.QMRF author(s) and contact details:

Sergi Gómez ProtoQSAR SL +34 960880658 sgomez@protoqsar.com https://protoqsar.com/

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Sergi Gómez-Ganau ProtoQSAR SL +34 960880658 sgomez@protoqsar.com https://protoqsar.com/

[2]Rafael Gozalbes ProtoQSAR SL +34 960880658 rgozalbes@protoqsar.com https://protoqsar.com/

2.6.Date of model development and/or publication:

February 2019

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

VEGA-QSAR: AI inside a platform for predictive toxicology. Proceedings of the workshop "Popularize Artificial Intelligence 2013" Benfenati E, Manganaro A, Gini G., December 5th 2013, Turin, Italy Published on CEUR Workshop Proceedings Vol-1107 http://ceur-ws.org/Vol1107/paper8.pdf

2.8.Availability of information about the model:

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

None to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Pseudokirchneriella subcapitata, now named Rhapidocelis subcapitata, previously named as Selenastrum capricornutum

3.2.Endpoint:

Ecotoxicological properties: Algae growth inhibition test (OECD 201)

3.3.Comment on endpoint:

The test endpoint (OECD 201) is inhibition of growth, expressed as the logarithmic increase in biomass (average specific growth rate) during the 72h exposure period. From the average specific growth rates recorded in a series of test solutions, the concentration bringing about a specified x % inhibition of growth rate (e.g. 50%) is determined and expressed as the ErCx (e.g. ErC50).

3.4.Endpoint units:

EC 50 in mg/L

3.5.Dependent variable:

Log EC50

3.6.Experimental protocol:

N/A

3.7.Endpoint data quality and variability:

The QSAR model is based on a dataset from the Japanese Ministry of Environment . For these compounds, experimental values of EC 50 after 72 hours for Pseudokirchneriella subcapitata are given. From the initial database, 361 biocide-like compounds were found by application of the biocide-chemical space and were used to develop the model. Biocide-like filters were previously defined as those properties featuring the structural chemical space of most of biocides. To do this, and in the context of the LIFE-EU COMBASE project (http://www.life-combase.com), different cut-off values for a list of physicochemical parameters were determined by comparing databases of biocides and generic chemicals, and served to identify a set of common biocide properties.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

An integrated model was arranged cascading a qualitative QSAR model and a quantitative QSAR model to predict the acute aquatic toxicity in algae. Firstly, a classification using automated neural networks analysis based on experimental values from the Japanese Ministry of Environment dataset was performed. Compound is considered toxic when EC50 < 10 mg/L. Secondly, quantitative model using support vector machine (SVM) analysis and the LogEC50 as dependent variable was carried out.

4.2.Explicit algorithm:

Automated Neural Networks (SANN)

Artificial neural networks are one of the main tools used in machine learning. As the “neural” part of their name suggests, they are brain-inspired systems which are intended to replicate the way that we humans learn. Neural networks consist of input and output layers, as well as (in most cases) a hidden layer consisting of units that transform the input into something that the output layer can use. They are excellent tools for finding patterns which are far too complex or numerous for a human programmer to extract and teach the machine to recognize.

Support Vector Machine

In machine learning, support-vector machines are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of

training examples, each marked as belonging to one or the other of two categories, a SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). A SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

4.3.Descriptors in the model:

[1]Descriptors within the Qualitative model:

[2]B01[C-Cl] Presence/absence of C - Cl at topological distance 01

[3]B01[C-O] Presence/absence of C - O at topological distance 01

[4]B02[Cl-Cl] Presence/absence of Cl - Cl at topological distance 02

[5]B02[N-O] Presence/absence of N - O at topological distance 02

[6]B03[N-O] Presence/absence of N - O at topological distance 03

[7]B09[C-C] Presence/absence of C - C at topological distance 09

[8]B09[O-O] Presence/absence of O - O at topological distance 09

[9]B10[C-O] Presence/absence of C - O at topological distance 10

[10]F01[C-O] Frequency of C - O at topological distance 01

[11]F02[C-O] Frequency of C - O at topological distance 02

[12]F04[O-O] Frequency of O - O at topological distance 04

[13]F10[C-N] Frequency of C - N at topological distance 10

[14]X3A Average connectivity index chi-3

[15]ATS5m Broto-Moreau autocorrelation of a topological structure - lag 5 / weighted by atomic masses

[16]Descriptors within the Quantitative model:

[17]ATS5m Broto-Moreau autocorrelation of a topological structure - lag 5 / weighted by atomic masses

[18]B01[O-S] Presence/absence of O - S at topological distance 01

[19]B02[N-Cl] Presence/absence of N - Cl at topological distance 02

[20]B02[N-O] Presence/absence of N - O at topological distance 02

[21]B03[O-Cl] Presence/absence of O - Cl at topological distance 03

[22]B05[C-S] Presence/absence of C - S at topological distance 05

[23]B06[N-Cl] Presence/absence of N - Cl at topological distance 06

[24]B07[N-O] Presence/absence of N - O at topological distance 07

[25]B09[C-C] Presence/absence of C - C at topological distance 09

[26]F01[C-O] Frequency of C - O at topological distance 01 2D frequency fingerprints

[27]F05[O-O] Frequency of O - O at topological distance 05

[28]MATS5m Moran autocorrelation - lag 5 / weighted by atomic masses

[29]nR10 Number of 10-membered rings

[30]X3A Average connectivity index chi-3

[31]X3sol Solvation connectivity index chi-3

4.4.Descriptor selection:

Molecular descriptors were calculated using CDK, Padel descriptor and E-Dragon software. Constant variables, near-constant variables and 0.95 pair-correlation variables were discarded. A sensitivity analysis for the qualitative model and a forward stepwise for the quantitative model was

used for variable selection.

4.5.Algorithm and descriptor generation:

Molecular descriptors were calculated using CDK, Padel descriptor and E-Dragon software.

4.6.Software name and version for descriptor generation:

4.7.Chemicals/Descriptors ratio:

Qualitative model: 361 / 14 = 25,78

Quantitative model: 361 /15 = 24,06

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The AD is assessed using the original algorithm implemented within VEGA. An overall AD index is calculated, based on a number of parameters, which relate to the results obtained on similar chemicals within the training and test sets.

5.2.Method used to assess the applicability domain:

The chemical similarity is measured with the algorithm developed for VEGA. Full details in the VEGA website (www.vegahub.eu), including the open access paper describing it. The AD also evaluates the correctness of the prediction on similar compounds (accuracy), the consistency between the predicted value for the target compound and the experimental values of the similar compounds, the range of the descriptors, and the presence of unusual fragments, using atom centred fragments. Full details in the VEGA website.

5.3.Software name and version for applicability domain assessment:

5.4.Limits of applicability:

VEGA provides a quantitative value for the prediction of each substance. This helps the user to identify potential critical aspects, which are indicated. Similar compounds are shown.

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: No

Smiles: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: null

6.3.Data for each descriptor variable for the training set:

No

6.4.Data for the dependent variable for the training set:

All

6.5.Other information about the training set:

6.6.Pre-processing of data before modelling:

The model is based on the Japanese Ministry of Environment. For these compounds, experimental values of EC 50 after 72 hours for Pseudokirchneriella subcapitata are given. From the initial database, 361 biocide-like compounds were found by application of the biocide-chemical space, and were used to develop the mode. Molecular descriptors were calculated using CDK, Padel descriptor and E-Dragon software. Constant variables, near-constant variables and 0.95 pair-correlation variables were discarded. Once the variables were calculated, STATISTICA and MINITAB packages were used to carry out the model. First the whole dataset was randomly divided in training set (70%) and validation set (15%) and external validation set (15%). After, a Forward Stepwise and a sensitivity analysis approach were used for variable selection and an Automated Neural Networks analysis was performed. Once the qualitative model was developed, a quantitative model to predict the EC50 after 72 hours was carried out by using support vector machine analysis.

6.7.Statistics for goodness-of-fit:

Qualitative model:

Training (70% dataset): Accuracy 80.46%, Specificity 79.43%, Sensitivity 81.50%

Validation set (15% dataset): Accuracy 81.17%, Specificity 73.07%, Sensitivity 89.28%

Quantitative model:

R2 = 0.75

6.8.Robustness - Statistics obtained by leave-one-out cross-validation:

N/A

6.9.Robustness - Statistics obtained by leave-many-out cross-validation:

N/A

6.10.Robustness - Statistics obtained by Y-scrambling:

N/A

6.11.Robustness - Statistics obtained by bootstrap:

N/A

6.12.Robustness - Statistics obtained by other methods:

N/A

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: Yes

Formula: No

INChI: No

MOL file: No

NanoMaterial: null

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:

7.6.Experimental design of test set:

A validation and an external validation tests were created randomly selecting 15% of the data collection for each set.

7.7.Predictivity - Statistics obtained by external validation:

Qualitative model:

External validation set (15% datset): Accuracy 79.48%, Specificity 73.68%, Sensitivity 85.29%

Quantitative model:

R2 = 0.64

7.8.Predictivity - Assessment of the external validation set:

N/A

7.9.Comments on the external validation of the model:

N/A

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The mechanistic approach of the present model is supported by the use of the specific atoms, bonds, and molecular fragments involved in the model descriptors.

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation was determined a posteriori by interpreting and modifying the final set of descriptors which contributed to the best fit.

8.3.Other information about the mechanistic interpretation:

N/A

9.Miscellaneous information

9.1.Comments:

N/A

9.2.Bibliography:

Japanese Ministry of Environment (2010) Japan Ecotoxicity Tests Data, March 2010.

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

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