

	QMRF identifier (JRC Inventory): To be entered by JRC
	QMRF Title: Algae Classification Toxicity Model (ProtoQSAR/Combase) v 1.0.1
	Printing Date: Feb 3, 2020

1. QSAR identifier

1.1. QSAR identifier (title):

Algae Classification Toxicity Model (ProtoQSAR/Combase) v 1.0.1

1.2. Other related models:

NA

1.2. Software coding the model:

VEGA (<https://www.vegahub.eu/>)

The VEGA software provides QSAR models to predict tox, ecotox, environ, phys-chem and toxicokinetic properties of chemical substances.

emilio.benfenati@marionegri.it

2. General information

2.1. Date of QMRF:

February 2020

2.2. QMRF author(s) and contact details:

[1] Emilio Benfenati Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy emilio.benfenati@marionegri.it <https://www.marionegri.it/>

[2] Diego Bardena Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy diego.bardena@marionegri.it <https://www.marionegri.it/>

2.3. Date of QMRF update(s):

15/09/2021

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

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

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

[1] Blázquez, M., Andreu-Sánchez, O., Ballesteros, A., Fernández-Cruz, M.L., Fito, C., Gómez-Ganau, S., Gozalbes, R., Hernández-Moreno, D., de Julián-Ortiz, J.V., Lombardo, A., Marzo, M., Ranero, I., Ruiz-Costa, N., Tarazona-Díez, J.V. and Benfenati, E. (2021). Computational Tools for the Assessment and Substitution of Biocidal Active Substances of Ecotoxicological Concern. In Chemometrics and Cheminformatics in Aquatic Toxicology, K. Roy (Ed.). <https://doi.org/10.1002/9781119681397.ch27>

[2] Benfenati E, Manganaro A, Gini G

VEGA-QSAR: AI inside a platform for predictive toxicology

Proceedings of the workshop "Popularize Artificial Intelligence 2013", December 5th 2013, Turin, Italy

Published on CEUR Workshop Proceedings Vol-1107

2.8. Availability of information about the model:

The model is non-proprietary and the training set is available.

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

Another QMRF is not available.

3. Defining the endpoint - OECD Principle 1

3.1. Species:

The species name is *Rhaphidocelis subcapitata*, previously named as *Selenastrum capricornus* or *Pseudokirchneriella subcapitata*

3.2. Endpoint:

OECD TG 201 Freshwater Alga and Cyanobacteria, Growth Inhibition Test

3.3. Comment on endpoint:

This test evaluates the effects of a substance on the growth of freshwater microalgae. Exponentially growing test algae under optimal temperature and light conditions are exposed to the test substance in batch cultures. The biological effects are the reduction of growth in a series of algal cultures exposed to, at least, five concentrations of a test substance. The response variable is the reduction of the average specific growth rate (= calculated on the basis of the logarithmic increase of biomass during the test period, expressed per day) as a function of the exposure concentration within the exponential growth phase of the control cultures over a period of normally 72 hours (c.f. also OECD TG (2011) paragraph 46-52) 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). The QSAR model is based on a dataset from the Japanese Ministry of Environment including experimental values of ErC 50 after 72 hours for *R. Subcapitata* for 361 biocide-like compounds.

3.4. Endpoint units:

No units

3.5. Dependent variable:

Binary classification: toxic (ErC50 < 10 mg/L), NON-toxic (ErC50 ≥ 10 mg/L)

3.6. Experimental protocol:

Slight modification of the OECD Test No. 201 (2011, 2006 & 1984): Freshwater Alga and Cyanobacteria, Growth Inhibition Test, OECD Guidelines for the Testing of Chemicals, Section 2, Éditions OCDE, Paris, All tests were performed according to OECD 201 and GLP but some tests performed before 2002 employed dispersants. In a few cases where the control cultures specific growth rate had decreased at 72 hrs the specific growth rate was in accordance with OECD TG 201 calculated based on the growth until 48 hrs. <https://doi.org/10.1787/9789264069923-en>.

3.7. Endpoint data quality and variability:

The QSAR model is based on a dataset from the Japanese Ministry of Environment including experimental values of ErC 50 after 72 hours for *R. Subcapitata* for 361 biocide-like compounds. The dataset was created within the LIFE-EU COMBASE project (<http://www.life-combase.com>). For the data curation see section 6.6.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

A classification model using automated neural networks analysis based on experimental values from the Japanese Ministry of Environment dataset. Toxicity threshold was set as ErC50 < 10 mg/L.

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

4.3. Descriptors in the model:

- [1]B01[C-Cl] Presence/absence of C - Cl at topological distance 01
- [2]B01[C-O] Presence/absence of C - O at topological distance 01
- [3]B02[Cl-Cl] Presence/absence of Cl - Cl at topological distance 02
- [4]B02[N-O] Presence/absence of N - O at topological distance 02
- [5]B03[N-O] Presence/absence of N - O at topological distance 03
- [6]B09[C-C] Presence/absence of C - C at topological distance 09
- [7]B09[O-O] Presence/absence of O - O at topological distance 09
- [8]B10[C-O] Presence/absence of C - O at topological distance 10
- [9]F01[C-O] Frequency of C - O at topological distance 01
- [10]F02[C-O] Frequency of C - O at topological distance 02
- [11]F04[O-O] Frequency of O - O at topological distance 04
- [12]F10[C-N] Frequency of C - N at topological distance 10
- [13]X3A Average connectivity index chi-
- [14]ATS5m Broto-Moreau autocorrelation of a topological structure - lag 5 / weighted by atomic masses

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 forward stepwise analysis for the quantitative model was used for variable selection

4.5. Algorithm and descriptor generation:

NA

4.6. Software name and version for descriptor generation:

NA

4.7. Chemicals/Descriptors ratio:

14 descriptors for 361 chemicals $361/14 = 25.78$

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 and is defined in this way for this QSAR model's predictions.

If $1 \geq \text{AD index} \geq 0.8$, the predicted substance is in the Applicability Domain of the model. It corresponds to “good reliability of prediction”.

If $0.8 > \text{AD index} \geq 0.6$, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability of prediction”.

If $\text{AD index} < 0.6$, the predicted substance is out of the Applicability Domain of the model and corresponds to “low reliability of prediction”.

Indices are calculated on the first $k = 3$ most similar molecules, each having S_k similarity value with the target molecule.

Similarity index (*IdxSimilarity*) is calculated as:

$$\frac{\sum_k S_k}{k} \times (1 - Diam^2)$$

where *Diam* is the difference in similarity values between the most similar molecule and the *k*-th molecule.

Accuracy index (*IdxAccuracy*) is calculated as:

$$\frac{\sum_c \log(1 + S_c)}{\sum_k \log(1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the prediction of the model matches with the experimental value of the molecule.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c \log(1 + S_c)}{\sum_k \log(1 + S_k)}$$

where the molecules with *c* index are the subset of the *k* molecules where the experimental value of the molecule matches with the prediction made for the target molecule.

ACF contribution (*IdxACF*) index is calculated as

$$ACF = rare \times missing$$

where: *rare* is calculated on the number of fragments found in the molecule and found in the training set in less than 3 occurrences as following: if the number is 0, *rare* is set to 1.0; if the number is 1, *rare* is set to 0.6; otherwise *rare* is set to 0.4

missing is calculated on the number of fragments found in the molecule and never found in the training set as following: if the number is 0, *missing* is set to 1.0; if the number is 1, *missing* is set to 0.6; otherwise *missing* is set to 0.4

Descriptors Range (*IdxDescRange*) index is calculated as 1.0 if all molecular descriptors used in the prediction fall within the range of descriptors used in the whole training set, 0.0 otherwise.

AD final index is calculated as following:

$$ADI = (IdxSimilarity)^{0.5} \times IdxAccuracy^{0.25} \times IdxConcordance^{0.25} \times IdxACF \times IdxDescRange$$

5.2. Method used to assess the applicability domain:

The Applicability Domain and 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 [2]. The VEGA 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.

Information on these indices is given below:

Similar molecules with known experimental value:

This index takes into account how similar are the first two most similar compounds found. Values near 1 mean that the predicted compound is well represented in the dataset used to build the model, otherwise the prediction could be an extrapolation. Defined intervals are:

If $1 \geq \text{index} > 0.8$, strongly similar compounds with known experimental value in the training set have been found

If $0.8 \geq \text{index} > 0.6$, only moderately similar compounds with known experimental value in the training set have been found

If $\text{index} \leq 0.6$, no similar compounds with known experimental value in the training set have been found

Accuracy of prediction for similar molecules:

This index takes into account the classification accuracy in prediction for the two most similar compounds found. Values near 1 mean that the predicted compounds fall in an area of the model's space where the model gives reliable predictions (no misclassifications), otherwise the lower is the value, the worse the model behaves. Defined intervals are:

If $1 \geq \text{index} > 0.8$, accuracy of prediction for similar molecules found in the training set is good

If $0.8 \geq \text{index} > 0.6$, accuracy of prediction for similar molecules found in the training set is not optimal

If $\text{index} \leq 0.6$, accuracy of prediction for similar molecules found in the training set is not adequate

Concordance for similar molecules:

This index takes into account the difference between the predicted value and the experimental values of the two most similar compounds. Values near 0 mean that the prediction made disagrees with the values found in the model's space, thus the prediction could be unreliable. Defined intervals are:

If $1 \geq \text{index} > 0.8$, similar molecules found in the training set have experimental values that agree with the predicted value

If $0.8 \geq \text{index} > 0.6$, some similar molecules found in the training set have experimental values that disagree with the predicted value

If $\text{index} \leq 0.6$, similar molecules found in the training set have experimental values that disagree with the predicted value

Atom Centered Fragments similarity check:

This index takes into account the presence of one or more fragments that aren't found in the training set, or that are rare fragments. First order atom centered fragments from all molecules in the training set are calculated, then compared with the first order atom centered fragments from the predicted compound; then the index is calculated as following: a first index RARE takes into account rare fragments (those who occur less than three times in the training set), having value of 1 if no such fragments are found, 0.85 if up to 2 fragments are found, 0.7 if more than 2 fragments are found; a second index NOTFOUND takes into account not found fragments, having value of 1 if no such fragments are found, 0.6 if a fragments is found, 0.4 if more than 1 fragment is found. Then, the final index is given as the product RARE * NOTFOUND. Defined intervals are:

If $\text{index} = 1$, all atom centered fragment of the compound have been found in the compounds of the training set

If $1 > \text{index} \geq 0.7$, some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments

If $\text{index} < 0.6$, a prominent number of atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

Model descriptors range check:

This index checks if the descriptors calculated for the predicted compound are inside the range of descriptors of the training and test set. The index has value 1 if all descriptors are inside the range, 0 if at least one descriptor is out of the range. Defined intervals are:

Index = TRUE, descriptors for this compound have values inside the descriptor range of the compounds of the training set

Index = FALSE, descriptors for this compound have values outside the descriptor range of the compounds of the training set

5.3. Software name and version for applicability domain assessment:

VEGA (www.vegahub.eu)

5.4. Limits of applicability:

VEGA provides a quantitative ADI-value for the prediction of each substance. This helps the user to identify potential critical aspects related to the Applicability Domain. 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: No

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:

Training set is composed by 253 biocides, 141 Toxic and 112 NON-Toxic (c.f. point 4.1) .

6.6. Pre-processing of data before modelling:

The model is based on the Japanese Ministry of Environment. For these compounds, experimental values of ErC50 after 72 hours for *Rhaphidocelis subcapitata* are given.,. A biocide-like chemical space was developed comparing 6512 chemical structures from the Physprop database and 257 chemical structures of biocides from COMBASE database. Five molecular descriptors were selected as filters. These filters were applied to the Japanese Ministry of Environment dataset for acute aquatic toxicity in algae (650 compounds), and finally we obtained a data set of 361 biocide-like structures.

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%). The quantitative model to predict the ErC50 after 72 hours was carried out by using support vector machine.

6.7. Statistics for goodness-of-fit:

Training (70% dataset): Accuracy 80%, Specificity 79%, Sensitivity 82%

After the implementation in VEGA:

Training: n = 253, Accuracy 0.81, Sensitivity 0.84, Specificity 0.76,

TP 119, TN 85, FP 27, FN 22

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

NA

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

NA

6.10. Robustness - Statistics obtained by Y-scrambling:

NA

6.11. Robustness - Statistics obtained by bootstrap:

NA

6.12. Robustness - Statistics obtained by other methods:

NA

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: No

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:

Test set is composed by 108 biocides, 66 Toxic and 42 NON-Toxic

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:

Validation Test (15% dataset): Accuracy 81%, Specificity 73 %, Sensitivity 89%

External validation set (15% dataset): Accuracy 79%, Specificity 74%, Sensitivity 85%

After the implementation in VEGA:

Test set: n 108, Accuracy 80%, Specificity 79%, Sensitivity 82%,

TP 54, TN 33, FP 9, FN 12

Considering ADI thresholds:

Test set in AD: n 37, Accuracy 89%, Sensitivity 95, Specificity 80%, MCC 0.78,

TP 21, TN 12, FP 3, FN 1

Test set could be out of AD: n 31, Accuracy 74%, Sensitivity 68%, Specificity 89%, MCC 0.52, TP 15, TN 8, FP 1, FN 7

Test set out of AD: n 40, Accuracy 78%, Sensitivity 82%, Specificity 72%, MCC 0.54,

TP 18, TN 13, FP 5, FN 4

7.8. Predictivity - Assessment of the external validation set:

NA

7.9. Comments on the external validation of the model:

NA

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:

NA

9. Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] Japanese Ministry of Environment (2010) Japan Ecotoxicity Tests Data, March 2010

[2] Floris M, Manganaro A, Nicolotti O, Medda R, Mangiatordi GF, Benfenati E. A generalizable definition of chemical similarity for read-across. J Cheminform. 2014 Oct 18;6(1):39. doi: 10.1186/s13321-014-0039-1. PMID: 25383097; PMCID: PMC4212147.

[3] OECD TG 201 (2011) "Freshwater Alga and Cyanobacteria, Growth Inhibition Test"

9.3. Supporting information:

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

All available dataset are present in the model inside the VEGA software