Guide to Algae Classification Toxicity Model for biocides
version 1.0.0

Table of Contents

1. Model explanation

1.1 Introduction

The model provides a classification model for toxicity in algae (Pseudokirchneriella subcapitata), specific for biocides. It has been developed inside the EU LIFE COMBASE project (LIFE15 ENV/ES/416). It is implemented inside the VEGA online platform, accessible at:http://www.vega-qsar.eu/.

1.2 Model details

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.

The qualitative model has been developed using the automated neural networks (ANN) analysis and based on the 361 filtered compounds. The compounds were randomly distributed in training set (254), test set (53) and external validation set (54). The descriptors used in this model are the following:

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

The descriptors were calculated, in the original model, by means of Dragon software and are now entirely calculated by an in-house software module in which they are implemented as described in: R. Todeschini and V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009.

1.3 Applicability Domain

The applicability domain of predictions is assessed using an Applicability Domain Index (ADI) that has values from 0 (worst case) to 1 (best case). The ADI is calculated by grouping several other indices, each one taking into account a particular issue of the applicability domain. Most of the indices are based on the calculation of the most similar compounds found in the training and test set of the model, calculated by a similarity index that consider molecule's fingerprint and structural aspects (count of atoms, rings and relevant fragments).

For each index, including the final ADI, three intervals for its values are defined, such that the first interval corresponds to a positive evaluation, the second one corresponds to a suspicious evaluation and the last one corresponds to a negative evaluation.

Following, all applicability domain components are reported along with their explanation and the intervals used. Note that for purpose of evaluating accuracy and concordance indices, prediction of "suspect mutagen" are considered as "mutagen".

- **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:

1 >= index > 0.8	strongly similar compounds with known experimental value in the training set have been found
0.8 >= index > 0.6	only moderately similar compounds with known experimental value in the training set have been found
index <= 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 falls 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:

1 >= index > 0.8	accuracy of prediction for similar molecules found in the training set is good
0.8 >= index > 0.6	accuracy of prediction for similar molecules found in the training set is not optimal
index <= 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:

1 >= index > 0.8	similar molecules found in the training set have experimental values that agree

	with the predicted value
0.8 >= index > 0.6	some similar molecules found in the training set have experimental values that disagree with the predicted value
index <= 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:

index = 1	all atom centered fragment of the compound have been found in the compounds of the training set
1 > index >= 0.7	some atom centered fragment of the compound have not been found in the compounds of the training set or are rare fragments
index < 0.7	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

- **Global AD Index**. The final global index takes into account all the previous indices, in order to give a general global assessment on the applicability domain for the predicted compound. Defined intervals are:

1 >= index >= 0.8	predicted substance is into the Applicability Domain of the model
0.8> index >= 0.6	predicted substance could be out of the Applicability Domain of the model
index < 0.6	predicted substance is out of the the Applicability Domain of the model

1.5 Model statistics

Following, statistics obtained applying the model to its original dataset:

- Training set: n = 254; Sensitivity = 81.50%; Specificity = 79.43%; Accuracy = 80.46%
- Validation set: n = 53; Sensitivity = 89.28%; Specificity = 73.07%; Accuracy = 81.17%
- Ext Validation set: n = 54; Sensitivity = 85.29%; Specificity = 73.68%; Accuracy = 79.48%

2. Model usage

2.1 Input

The model accepts as input two molecule formats: SDF (multiple MOL file) and SMILES. All molecules found as input are preprocessed before the calculation of molecular descriptors, in order to obtain a standardized representation of compound. For this reason, some cautions should be taken.

- **Hydrogen atoms**. In SDF files, hydrogen atoms should be explicit. As some times SDF file store only skeleton atoms, and hydrogen atoms are implicit, during the processing of the molecule the system tries to add implicit hydrogens on the basis of the known standard valence of each atom (for example, if a carbon atoms has three single bonds, an hydrogen atom will be added such to reach a valence of four). In SMILES molecules, the default notation uses implicit hydrogen. Anyway please note that in some cases it is necessary to explicitly report an hydrogen; this happens when the conformation is not unambiguous. For example, when a nitrogen atom is into an aromatic ring with a notation like "cnc" it is not clear whether it corresponds to C-N=C or to C-[NH]-C, thus if the situation is the latter, it should be explicitly reported as "c[nH]c".

- **Aromaticity**. The system calculates aromaticity using the basic Hueckel rule. Note that each software for drawing and storing of molecules can use different approaches to aromaticity (for instance, commonly the user can choose between the basic Hueckel rule and a loose approach that lead to considering aromatic a greater number of rings). As in the input files aromaticity can be set explicitly (for instance, in SMILES format by using lowercase letters), during the processing of the molecule the system removes aromaticity from rings that don't satisfy the Hueckel rule. Please note that when aromaticity is removed from a ring, it is not always possible to rebuild the original structure in Kekule form (i.e. with an alternation of single and double bonds, like in the SMILES for benzene, C=1C=CC=CC1), in this case all bonds are set to single. Furthermore, please note that aromaticity detection is a really relevant issue, some molecular descriptors can have significantly different values whether a ring is perceived as aromatic or not. For this reason it is strongly recommended:
- Always use explicit hydrogens in SDF file.
- Avoid explicit aromaticity notation in original files; in this way, the perception of aromaticity is left to the preprocessing step and there is no chance of mistakes due to the transformation of rings that were set to aromatic in the original format but not recognized as aromatic in VEGA.

Note that when some modification of the molecule are performed during the preprocessing (e.g. adding of lacking hydrogens, correction of aromaticity), a warning is given in the remark field of the results.

2.2 Output

Results given as text file consist of a plain-text tabbed file (easily importable and processable by any spreadsheet software) containing in each row all the information about the prediction of a molecule. Note that if some problems were encountered while processing the molecule structure, some warning

are reported in the last field (Remarks).

Results given as PDF file consists of a document containing all the information about the prediction. For each molecule, results are organized in sections with the following order:

1 – Prediction summary

Here is reported a depiction of the compound and the final assessment of the prediction (i.e. the prediction made together with the analysis of the applicability domain). The prediction and the experimental value (if available) are given in log(1/(mmol/L)) and in mg/L

Note that if some problems were encountered while processing the molecule structure, some warnings are reported in the last field (Remarks).

A graphical representation of the evaluation of the prediction and of its reliability is also provided, using the following elements:


Compound is classified as non-toxic


Compound is classified as toxic


Prediction has low reliability (compound out of the AD)


Prediction has moderate reliability (compound could be out of the AD)


Prediction has high reliability (compound into the AD)

3.1 – Applicability Domain: Similar compounds, with predicted and experimental values

Here it is reported the list of the six most similar compounds found in the training and test set of the model, along with their depiction and relevant information (mainly experimental value and predicted value).

3.2 – Applicability Domain: Measured Applicability Domain scores

Here it is reported the list of all Applicability Domain scores, starting with the global Applicability Domain Index (ADI). Note that the final assessment on prediction reliability is given on the basis of the value of the ADI. For each index, it is reported its value and a brief explanation of the meaning of that value.

4.1 – Reasoning: Relevant chemical fragments and moieties

If some rare and/or missing Atom Centered Fragments are found, they are reported here with a depiction of each fragment.