Guide to 48 hour Daphnia Magna LC_{50} Model version 1.0.7

Table of Contents

1. Model explanation....2
   1.1 Introduction....2
   1.2 Model details....2
   1.3 Applicability Domain....3
   1.4 Model statistics....5
2. Model usage....6
   2.1 Input....6
   2.2 Output....6
3. Differences from previous versions....8
   3.1 Differences with T.E.S.T. model....8
   3.2 VEGA model history....8
      3.2.1 Version 1.0.1....8
      3.2.2 Version 1.0.2....8
      3.2.3 Version 1.0.3....8
      3.2.4 Version 1.0.4....8
      3.2.5 Version 1.0.6....9
      3.2.6 Version 1.0.7....9

1. Model explanation

1.1 Introduction

The model provides a quantitative prediction for Daphnia Magna LC_{50} (48 hour), given in -log(mol/l) and its conversion in mg/L. It is implemented inside the VEGA online platform, accessible at:http://www.vega-qsar.eu/
The model is a re-implementation of the original model developed by Todd Martin inside T.E.S.T. software for US EPA. The T.E.S.T. software is freely available at: http://www.epa.gov/nrmrl/std/cppb/qsar/.

1.2 Model details

The model is a linear regression made on 17 molecular descriptors. The regression coefficients have been calculated on the T.E.S.T. original dataset, that contains 337 compounds extracted from the ECOTOX aquatic toxicity database (http://cfpub.epa.gov/ecotox/), splitted in 269 compounds for the training set and 68 for the test set. More details on the model can be found in the T.E.S.T. documentation: http://www.epa.gov/nrmrl/std/cppb/qsar/testuserguide.pdf

The descriptors used are the following:
- xc4: Simple 4th order cluster chi index
- StN: Sum of (tN) E-States (StN)
- SsSH: Sum of (-SH) E-States (SsSH)
- SsOH_acnt: Count of (-OH) (SsOH_acnt)
- Hmax: Maximum hydrogen E-State value in molecule
- MDEN33: Molecular distance edge between all tertiary nitrogens
- BEHm1: Highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses
- BEHp1: Highest eigenvalue n. 1 of Burden matrix / weighted by atomic polarizabilities
- Mv: Mean atomic van der Waals volume (scaled on Carbon atom)
- MATS1m: Moran autocorrelation - lag 1 / weighted by atomic masses
- MATS1e: Moran autocorrelation - lag 1 / weighted by atomic Sanderson electronegativities
- GATS3m: Geary autocorrelation - lag 3 / weighted by atomic masses
- AMR: Ghose-Crippen molar refractivity
- C(=S)- [2 nitrogen attach]: -C(=S)- [2 nitrogen attach] fragment count
- AN: AN fragment count
- N< [attached to P]: -N< [attached to P] fragment count
- S(=O)(=O)- [aromatic attach]: -S(=O)(=O)- [aromatic attach] fragment count

The descriptors are 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. The original code for descriptors calculation has been kindly provided by Todd Martin.

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.

- 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 > 0.85	strongly similar compounds with known experimental value in the training set have been found
0.85 >= index > 0.7	only moderately similar compounds with known experimental value in the training set have been found
index <= 0.7	no similar compounds with known experimental value in the training set have been found

- Accuracy (average error) of prediction for similar molecules. This index takes into account the error in prediction for the two most similar compounds found. Values near 0 mean that the predicted compounds falls in an area of the model's space where the model gives reliable predictions, otherwise the greater is the value, the worse the model behaves. Defined intervals are:

index < 0.6	accuracy of prediction for similar molecules found in the training set is good
0.6 <= index <= 1.2	accuracy of prediction for similar molecules found in the training set is not optimal
index > 1.2	accuracy of prediction for similar molecules found in the training set is not adequate

- Concordance with similar molecules (average difference between target compound prediction and experimental values of 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 agrees with the experimental values found in the model's space, thus the prediction is reliable. Defined intervals are:

index < 0.6	similar molecules found in the training set have experimental values that agree with the target compound predicted value

0.6 <= index <= 1.2	similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value
index > 1.2	similar molecules found in the training set have experimental values that completely disagree with the target compound predicted value

- **Maximum error of prediction among similar molecules**. This index takes into account the maximum error in prediction among the two most similar compounds. Values near 0 means that the predicted compounds falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

index < 0.6	the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability
0.6 <= index < 1.2	the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability
index >= 1.2	the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

- **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 > 0.85	predicted substance is into the Applicability Domain of the model
0.85 => index > 0.7	predicted substance could be out of the Applicability Domain of the model
index <= 0.7	predicted substance is out of the the Applicability Domain of the model

1.4 Model statistics

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

- Training set: n = 269; R^2 = 0.71; RMSE = 0.96
- Test set: n = 68; R^2 = 0.49; RMSE = 1.02

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). Following, all information related to the prediction are reported. Prediction and experimental value (if available) are given in -log(mol/l). These values are also provided expressed in mg/l. The MW (Molecular Weight) of the compound is also provided; MW is necessary for the conversion, as the formula used is:

$$[mg/l] = 1000 \cdot 10^{-1 \cdot [-\log(mol/l)]} \cdot MW$$

Note that if some problems were encountered while processing the molecule structure, some warning 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 non-toxic, LC_{50} value is more than 100 mg/l


Compound is toxic, LC_{50} value is less than 100 mg/l and more than 10 mg/l


Compound is toxic, LC_{50} value is less than 10 mg/l and more than 1 mg/l


Compound is highly toxic, LC_{50} value is less than 1 mg/l


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.

3. Differences from previous versions

3.1 Differences with T.E.S.T. model

The VEGA model has several differences that can lead to prediction that can be slightly different from the ones produced by the original US EPA T.E.S.T. model. Mainly, the values of descriptors could be different for some molecules, even if the algorithm definition is the same, due to some different compound's preprocessing (like for the detection of aromaticity). Furthermore, one original descriptor has been not used in the VEGA model (ssi, Standardized Shannon Information or standardized information content) and it has been replaced by a constant coefficient (0.85) derived from the mean value of the descriptor in the original dataset.

3.2 VEGA model history

3.2.1 Version 1.0.1

First official release published in the VEGA platform.

3.2.2 Version 1.0.2

Prediction is now provided also in mg/L units.
This version is updated with the new calculation core (1.0.23) where similarity algorithm is slightly changed. The new version considers halogen atoms are really similar, especially Chlorine and Bromine atoms are considered almost the same. The main difference with previous algorithm can be thus seen just for halogenated compounds.
A more precise check for similarity has been introduced for the extraction of experimental values, in order to avoid mismatches (as the similarity index is based on fingerprints, there are some rare cases in which a value equal to 1 does not points to a exactly isomorph compound).
There are NO changes in prediction values, but as similarity is changed some small differences in AD assessment can be found.

3.2.3 Version 1.0.3

This version is updated with the new calculation core (1.0.26). Some minor bugs in the procedure for reading molecule structures have been fixed; some compounds, previously not loaded, could now be correctly processed. All values are now given with explicit unit of measurement. Also the experimental value (if available) is now provided in mg/L units.
There are NO changes in prediction values and in AD assessment.

3.2.4 Version 1.0.4

This version is updated with the new calculation core (1.0.27), that generates a graphically renewed PDF report. In this version, the propositions for prediction and assessment are changed, but there are

NO changes in their values.

3.2.5 Version 1.0.6

This version is updated with the new calculation core (1.1.1) based on a new release of the CDK libraries (1.4.9). These updates can influence the calculation, so there could be some changes in the predictions produced.
The new calculation core implements a new version of the algorithm used for calculating the similarity index. This means that the list of similar molecules given as part of the applicability domain evaluation will often be different from the ones produced by older releases of the model. Furthermore, the applicability domain index (ADI) itself and the final assessment could often be different.
Model statistics in the current guide have been updated with the new values.
Some thresholds for the applicability domain sub-indices have been revised to obtain better performances.

3.2.6 Version 1.0.7

This version is updated with the new calculation core (1.2.0). This update can influence some calculation, in particular similarity evaluation, so there could be some changes in the applicability domain values produced.