	***QMRF identifier (JRC Inventory):*** **To be entered by JRC**
	QMRF Title: **VEGA BCF model (Maylan) v 1.0.4**
	Printing Date: **Apr 1, 2019**

1.QSAR identifier

1.1. QSAR identifier (title):

VEGA BCF model (Maylan) v 1.0.4

1.2. Other related models:

The model is based on the method proposed by Meylan et al. as implemented in the EPI Suite BCFBAF module v 4.11 (http://www.epa.gov/oppt/exposure/pubs/episuite.htm)

1.3. 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:

01/06/2016

2.2. QMRF author(s) and contact details:

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

2.3. Date of QMRF update(s):

NA

2.4. QMRF update(s):

NA

2.5. Model developer(s) and contact details:

Alberto Manganaro Istituto di Ricerche Farmacologiche Mario Negri - IRCSS Via Mario Negri 2, 20156 Milano, Italy alberto.manganaro@marionegri.it https://www.marionegri.it/

2.6. Date of model development and/or publication:

The model was developed on April 2013

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

[1] Meylan, W.M.,Howard,P.H.,Boethling,R.S.,1999.Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. Environ. Toxicol.Chem. 18,664–672

[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 hybrid model is available.

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

Other QMRF is not available

3.Defining the endpoint - OECD Principle 1

3.1. Species:

Fish

3.2. Endpoint:

2.Environmental fate parameters 2.4.a.Bioconcentration . BCF fish

3.3. Comment on endpoint:

The bioconcentration factor (BCF) is the concentration of test substance in the fish or specified tissues thereof divided by the concentration of the chemical in the surrounding medium at steady state

3.4. Endpoint units:

Continuous value expressed in Log(L/kg)

3.5. Dependent variable:

LogP (see 4.2)

Log P is considered to keep into account the basis of the transfer from water to the lipidic phase of the cell

3.6. Experimental protocol:

Bioconcentration factor (BCF) data are provided from tests that are conducted with respect to the OECD Test Guideline No. 305: OECD GUIDELINES FOR TESTING OF CHEMICALS:

a) OECD TG 305A: Bioaccumulation:Sequential Static Fish Test, 1981

b) OECD TG 305B: Bioaccumulation: Semi-Static Fish Test 1981

c) OECD TG 305C: Bioaccumulation: Test for the Degree of Bioaccumulation in Fish, 1981

d) OECD TG 305D: Bioaccumulation: Static Fish Test, 1981

e) OECD TG 305E: Bioaccumulation: Flow-through Fish Test, 1981,

f) OECD TG 305: Bioconcentration: Flow-through Fish Test (1996) &

g) OECD TG 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure (2012)

It is noted that all OECD TG 305 under point a) to e) above has been deleted but the versions from 1981 was used until 1996 when replaced by the joint version from 1996 (c.f. point f) above) which were then used until it was significantly updated in 2012 (c.f. point g above), to the version which has been employed since then.

3.7. Endpoint data quality and variability:

Data as from U. S. Environmental Protection Agency website (http://esc.syrres.com/interkow/EpiSuiteData.htm) were used. Particularly, the measured BCF values used were selected from a quality reviewed BCF database described in [4]; single BCF values were selected for each compound (median values were generally selected for compounds with multiple values). Then processed and cleared from duplicates and mono-constituent organic compounds provided with structure that had problems. The final dataset has 662 mono-constituent organic substances The experimental BCF data collection includes historical fish BCF data based on previous versions of OECD TG 305 (c.f. point 3.6). But as these experimental data have undergone significant expert evaluations, it is generally believed that the BCF data collection employed im many cases may be of almost similar reliability as BCF data obtained by employing the newest OECD TG 305 version.

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

QSAR model with different regression equations or fixed values, selected on the basis of an initial classification between ionic and non-ionic compounds, and on the value of the predicted logP value corrected with factors depending on the presence/absence of 12 fragments

4.2. Explicit algorithm:

Meylan BCF methodology

Compounds are classified as either ionic or non-ionic. Ionic compounds include carboxylic acids, sulfonic acids and salts of sulfonic acids, and charged nitrogen compounds (nitrogen with a valence such as quaternary ammonium compounds).

All other compounds are classified as non-ionic. Methodology fo rNon-Ionic was to separate compounds into three divisions by Log Kow value as follows:

Log Kow < 1.0

Log Kow 1.0 to 7.0

Log Kow > 7.0

For Log Kow 1.0 to 7.0 the derived QSAR estimation equation is:

Log BCF = 0.6598 Log Kow - 0.333 + correction factors

For Log Kow > 7.0 the derived QSAR estimation equation is:

Log BCF = -0.49 Log Kow + 7.554 + correction factors

For Log Kow < 1.0 the derived QSAR estimation equation is:

All compounds with a log Kow of less than 1.0 are assigned an estimated log BCF of 0.50

Ionic compounds are predicted as follows:

log BCF = 0.50 (log Kow < 5.0) log BCF = 1.00 (log Kow 5.0 to 6.0) log BCF = 1.75 (log Kow 6.0 to 8.0) log BCF = 1.00 (log Kow 8.0 to 9.0) log BCF = 0.50 (log Kow > 9.0)

The correction factors and their values are the following:

Ketone (aromatic connection) -0.5851

Phosphate ester -0.825 4.

Multi-halogenated biphenyl/PAH 0.586

Aromatic ring-CH-OH -0.2556

Aromatic sym-triazine ring -0.5169

Tert-Butyl ortho-phenol type -0.222

Phenanthrene ring 0.6609

Cyclopropyl-C(=O)-O- ester -1.2591

Alkyl chains (8+ CH2 groups) with logKow >4 & <7.0 -1.3743

Alkyl chains (8+ CH2 groups) with logKow 7-10 -0.5965

Disulfide (-S-S-) -1.3404

4.3. Descriptors in the model:

logKow (logP)

logP prediction based on the original Meylan approach (available in the EPI Sutie application) as re-implemented in VEGA

4.4. Descriptor selection:

No selection

4.5. Algorithm and descriptor generation:

The model is based on fragments to define different chemical classes: the method classifies a compound as either ionic or non-ionic. Ionic compounds include carboxylic acids, sulfonic acids and salts of sulfonic acids, and charged nitrogen compounds (nitrogen with a +5 valence such as quaternary ammonium compounds). All other compounds are classified as non-ionic. Different models apply to different classes. Log P is the descriptor used within each model to separate chemical classes

4.6. Software name and version for descriptor generation:

NA

4.7. Chemicals/Descriptors ratio:

Only one descriptor (Log P) is used

5.Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

The Applicability Domain (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.

ADI is defined in this way for this QSAR model´s predictions:

If 1 ≥ AD index > 0.85, the predicted substance is regarded in the Applicability Domain of the model. It corresponds to “good reliability” of prediction.

If 0.85 ≥ AD index > 0.75, the predicted substance could be out of the Applicability Domain of the model. It corresponds to “moderate reliability” of prediction.

If AD index ≤ 0.75, the predicted substance is regarded out of the Applicability Domain of the model. It corresponds to “low reliability” of prediction.

No ADI threshold was used to provide performance calculations of the validation sets

5.2. Method used to assess the applicability domain:

The chemical similarity is measured with the algorithm developed for VEGA. Full details are in the VEGA website (www.vegahub.eu), including the open access paper describing it [3]. 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.

These indices are defined in this way for this QSAR model:

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 ≥ index > 0.9, strongly similar compounds with known experimental value in the training set have been found

If 0.9 ≥ index > 0.75, only moderately similar compounds with known experimental value in the training set have been found

If index ≤ 0.75, 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 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 index < 0.5, accuracy of prediction for similar molecules found in the training set is good

If 1.0 ≥ index ≥ 0.5, accuracy of prediction for similar molecules found in the training set is not optimal

If index > 1, 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 index < 0.5, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.0 ≥ index ≥ 0.5, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

If index > 1.0, 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 fall in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:

If index < 0.5, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.0 >index ≥ 0.5, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If index ≥ 1.0, the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability

LogP reliability:

This index takes into account the reliability of the logP value used in the model. Note that the Meylan BCF model is strongly based on the logP prediction of the compound, thus this index is highly relevant for the assessment of the final prediction. The reliability of the logP value comes from the assessment of the VEGA LogP model (that provides the used logP value), which is also provided in the “Prediction summary” section of the report. Defined intervals are:

If index = 1, reliability of logP value used by the model is good

If index = 0.7, reliability of logP value used by the model is not optimal

If index = 0, reliability of logP value used by the model is not adequate

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 index = 1, all atom centered fragment of the compound have been found in the compounds of the training set

If 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

If index < 0.7, a prominent number of atoms 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:

If index = True, descriptors for this compound have values inside the descriptor range of the compounds of the training set

If 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:

The model is not applicable to inorganic chemicals and substances containing unusual elements (i.e., different from C, O, N, S, P, Cl, Br, F, I). Salts can be predicted only if converted to the neutralized form.

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: Yes

Formula: No

INChI: No

MOL file: No

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:

NA

6.6. Pre-processing of data before modelling:

The original dataset from EPI Suite has been taken, then processed and cleared from duplicates and compounds provided with structure that had problems. The final dataset has 662 compounds

6.7. Statistics for goodness-of-fit:

Training set: n = 516; R^2 = 0.80 RMSE = 0.55

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: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

No

7.4. Data for the dependent variable for the external validation set:

No

7.5. Other information about the external validation set:

To test the model, a test set of 148 compounds was used

7.6. Experimental design of test set:

NA

7.7. Predictivity - Statistics obtained by external validation:

External validation set: n = 146; R2 = 0.79% ; RMSE = 0.65

Ext. set in AD: n = 29, R2 0.89, RMSE 0.46

Ext. set “could be out of AD”: n = 53, R2 0.77, RMSE 0.53

Ext. set out of AD: n = 64, R2 0.76, RMSE 0.79

7.8. Predictivity - Assessment of the external validation set:

The predictivity of the model seems to be better when the compounds fall within the applicability domain of the model as defined by the VEGA ADI concept

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:

NA

8.2.A priori or a posteriori mechanistic interpretation:

A priori

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] Meylan,W.M.,Howard,P.H.,Boethling,R.S.,1999.Improved method for estimating bioconcentration/bioaccumulation factor from octanol/water partition coefficient. Environ.Toxicol.Chem.18,664–672

[2] OECD. Test Guideline No. 305:

OECD GUIDELINES FOR TESTING OF CHEMICALS:

h) OECD TG 305A: Bioaccumulation:Sequential Static Fish Test, 1981

i) OECD TG 305B: Bioaccumulation: Semi-Static Fish Test 1981

j) OECD TG 305C: Bioaccumulation: Test for the Degree of Bioaccumulation in Fish, 1981

k) OECD TG 305D: Bioaccumulation: Static Fish Test, 1981

l) OECD TG 305E: Bioaccumulation: Flow-through Fish Test, 1981,

m) OECD TG 305: Bioconcentration: Flow-through Fish Test (1996) &

OECD TG 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure. Paris: Organisation for Economic Co-operation and Development, 2012. https://www.oecd-ilibrary.org/environment/test-no-305-bioaccumulation-in-fish-aqueous-and-dietary-exposure_9789264185296-en.

[3] Floris, M., Manganaro, A., Nicolotti, O. et al. A generalizable definition of chemical similarity for read-across. J Cheminform 6, 39 (2014). https://doi.org/10.1186/s13321-014-0039-1

[4] Arnot, J. A., & Gobas, F. A. P. C. (2006). A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environmental Reviews, 14(4), 257–297.

9.3. Supporting information:

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

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

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