QMRF identifier (JRC Inventory): **To be entered by JRC**
QMRF Title: **VEGA BCF model (kNN/Read-Across) v 1.1.1**
Printing Date: **Feb 13, 2020**

1.QSAR identifier

1.1. QSAR identifier (title):

VEGA BCF model (kNN/Read-Across) v 1.1.1

1.2. Other related models:

NA.

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):

12/02/2020

2.4. QMRF update(s):

Updates made by Giuseppa Raitano, email: giuseppa.raitano@marionegri.it

Updates in the fields: -1.3, update of the VEGA platform url -2.8, -2.9, -3.2 -5.3, update of the VEGA platform url and version -6.7, number of compounds within the training set of the model -7.1, unavailable of the external validation set

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 2015

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

[1] Alberto Manganaro, Fabiola Pizzo, Anna Lombardo, Alberto Pogliaghi, Emilio Benfenati,

Predicting persistence in the sediment compartment with a new automatic software based on the k-Nearest Neighbor (k-NN) algorithm, Chemosphere, Volume 144, 2016, Pages 1624-1630,

[2] 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

[3] 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:

ENV FATE 5.3.1. Bioaccumulation: aquatic

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:

no units (concentration/concentration)

3.5. Dependent variable:

Continuous value expressed in Log(L/kg)

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:

This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources, including the original dataset of the CAESAR BCF model (note that experimental values may differ from the ones in the CAESAR BCF dataset, as this new dataset has been built including more sources), (see in section 9.2 [3]-[8]). Data have been compared in case of multiple values and the mean value was calculated for the compounds with more than one value. The final dataset is composed of 860 mono-costituent 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:

The read-across model has been built with the istKNN application (developed by Kode srl, www.kode-solutions.net) and it is based on the similarity index developed inside the VEGA platform; the index takes into account several structural aspects of the compounds

4.2. Explicit algorithm:

kNN prediction is based on the k most similar compounds retrieved with the similarity index developed in VEGA. Explanation of the kNN approach is available in [9].

The predicted value is calculated with the following algorithm:

1. The first k molecules with the closest similarity to the target compound are extracted.

2. Molecules with a similarity index lower than a given threshold S1 are excluded.

3. If no molecules are left, no prediction is provided (missing value).

4. If only one molecule is left, it is used as prediction only if its similarity value is equal to or higher than a given threshold S2, otherwise no prediction is provided (missing value).

5. In all other cases, the prediction is calculated as the weighted average value of the k most similar compounds experimental values, where for each compounds the weight is given by its similarity value. The weights (similarity values) can be raised to the power of a given value E, called the enhance factor, as for integer larger than 1the result is to enhance the role of molecules with higher similarity values in the prediction. Furthermore, a threshold for experimental values can be provided, so that the range of experimental values found in the chosen k most similar molecules is higher than the given threshold, no prediction is provided.

The k-NN model has the following settings: K (neighbours number): 4S1 (Similarity threshold:) 0.7 S2 (Similarity threshold for single molecules): 0.75 E (Enhance factor): 3 Allowed experimental range: 3.5

4.3. Descriptors in the model:

Similarity index Descriptors are only used to identify the similar compounds. Index for generic similarity as described in M. Floris, A. Manganaro, O. Nicolotti, R. Medda, G. F. Mangiatordi, E.Benfenati, “A generalizable definition of chemical similarity for read-across”, Journal of Cheminformatics (2014), vol. 6, 39

4.4. Descriptor selection:

No selection

4.5. Algorithm and descriptor generation:

The algorithm is an extension of kNN as described in section 4.2. The descriptors are only used for the similarity, as described in section 4.3

4.6. Software name and version for descriptor generation:

istKNN 0.9

in-house software for kNN modelling

alberto.manganaro@kode-solutions.net http://chm.kode-solutions.net.

4.7. Chemicals/Descriptors ratio:

No descriptors (descriptors are used only to identify the similar compounds).

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.

Indices are calculated on the first *k* = the number of *k* compounds used in the KNN model for the prediction 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:

$$\frac{\sum_c^k |exp_c - pred_c|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_c$ is the c-*th* molecule predicted value by the model.

Concordance index (*IdxConcordance*) is calculated as:

$$\frac{\sum_c^k \left|exp_c - pred_{target}\right|}{k}$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule.

Max Error index (*IdxMaxError*) is calculated as:

$$max(|exp_c - pred_c|)$$

where exp_c is the experimental value of the c-*th* molecule in the training set and $pred_{target}$ is the predicted value for the input target molecule, evaluated over the k molecules.

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 occurences 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

AD final index is calculated as following:

$$ADI = IdxSimilarity \times IdxACF$$

The initialADI index is the used together with the other sub-indices to calculate the final ADI, on the basis of the assessment class in which each sub-index falls:

IdxAccuracy ≥	IdxConcordance ≥	IdxMaxError ≥	InitialADI ≥	ADI
1.2	1.2	1.2	0.75	1.0
0.6	0.6	0.6	0.7	0.85
All other cases				0.7

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 [2]. 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 > 0.75, strongly similar compounds with known experimental value in the training set have been found

If 0.75 ≥ index > 0.7, only moderately similar compounds with known experimental value in the training set have been found

If 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 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 < 0.6, accuracy of prediction for similar molecules found in the training set is good

If 0.6 > index ≥ 1.2, accuracy of prediction for similar molecules found in the training set is not optimal

If index ≥ 1.2, 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 < 0.6, molecules found in the training set have experimental values that agree with the target compound predicted value

If 1.2 > index ≥ 0.6, similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value

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

If index < 0.6, the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability

If 1.2 > index ≥ 0.6, the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability

If 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:

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 atom centered fragments of the compound have not been found in the compounds of the training set or are rare fragments

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:

No

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

All

6.5. Other information about the training set:

The model performs a read-across on a dataset of 860 chemicals. This dataset has been made by Istituto di Ricerche Farmacologiche Mario Negri, merging experimental data from several reliable sources, including the original dataset of the CAESAR BCF model (note that experimental values may differ from the ones in the CAESAR BCF dataset, as this new dataset has been built including more sources)

6.6. Pre-processing of data before modelling:

No

6.7. Statistics for goodness-of-fit:

Training set: n = 860; R^2 = 0.67; RMSE = 0.76

Non predicted compounds: n = 24

kNN models provide non predicted compound due to absence of similar compounds, as described in section 4.2.

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:

NA

7.2. Available information for the external validation set:

NA

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:

NA

7.6. Experimental design of test set:

Not applicable

7.7. Predictivity - Statistics obtained by external validation:

NA

7.8. Predictivity - Assessment of the external validation set:

NA

7.9. Comments on the external validation of the model:

The use of the applicability domain index improves the robusteness of the model.

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:

NA

8.3. Other information about the mechanistic interpretation:

NA

9.Miscellaneous information

9.1. Comments:

NA

9.2. Bibliography:

[1] OECD. Test No. 305: Bioaccumulation in Fish: Aqueous and Dietary Exposure; Organisation for Economic Co-operation and Development: Paris, 2012.

[2] 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

[3] EURAS database. [http://www.cefic-lri.org/lri-toolbox/bcf]

[4] Arnot JA, Gobas FAPC: A review of bioconcentration factor (BCF) and bioaccumulation factor (BAF) assessments for organic chemicals in aquatic organisms. Environ Rev. 2006, 14: 257-297. 10.1139/A06-005.

[5] Dimitrov, S. et al., 2005. SAR QSAR Environ. Res., 16, 531-554

[6] «PPDB - Pesticides Properties DataBase». http://sitem.herts.ac.uk/aeru/ppdb/.

[7] Fu, Wenjing, Antonio Franco, e Stefan Trapp. «Methods for estimating the bioconcentration factor of ionizable organic chemicals». Environmental Toxicology and Chemistry 28, n. 7 (2009): 1372–79. https://doi.org/10.1897/08-233.1.

[8] Lombardo, A., Roncaglioni, A., Boriani, E., Milan, C., & Benfenati, E. (2010). Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish. Chemistry Central Journal, 4(Suppl 1), S1. https://doi.org/10.1186/1752-153X-4-S1-S1

[9] A. Manganaro, F. Pizzo, A.Lombardo, A. Pogliaghi, E. Benfenati, “Predicting persistence in the sediment compartment with anew automatic software based on the k-Nearest Neighbor (k-NN) algorithm”, Chemosphere (2016),vol. 144, 1624-1630

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