	QMRF identifier (JRC Inventory):**To be entered by JRC**
	QMRF Title:**Log P model (Meylan/Kowwin) v. 1.1.4 in VEGA v. 1.1.4**
	Printing Date:**17-feb-2020**

1.QSAR identifier

1.1.QSAR identifier (title):

Log P model (Meylan/Kowwin) v. 1.1.4 in VEGA v. 1.1.4

1.2.Other related models:

1.3.Software coding the model:

VEGA v. 1.4.4

https://www.vegahub.eu/portfolio-item/vega-qsar/

2.General information

2.1.Date of QMRF:

11 April 2010

2.2.QMRF author(s) and contact details:

[1]Domenico Gadaleta Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri", IRCCS domenico.gadaleta@marionegri.it

[2]Emilio Benfenati Laboratory of Environmental Chemistry and Toxicology, Istituto di Ricerche Farmacologiche "Mario Negri", IRCCS emilio.benfenati@marionegri.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]William M. Meylan Syracuse Research Corporation, Merrill Lane, Syracuse, NY 13210.

[2]Philip H. Howard Syracuse Research Corporation, Merrill Lane, Syracuse, NY 13210.

2.6.Date of model development and/or publication:

1995

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

Meylan, W.M. and P.H. Howard, Atom/fragment contribution method for estimating octanol/water partition coefficients. 1995, J. Pharm. Sci. 84: 83-92

2.8.Availability of information about the model:

Model’s guide is available for download from VEGA v. 1.1.4

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

Other QMRF for this model are not available

3.Defining the endpoint - OECD Principle 1

3.1.Species:

N/A

3.2.Endpoint:

QMRF 1. 6. Octanol-water partition coefficient (Kow) EC A.8 Partition Coefficient (EU method includes both shake flask and HPLC)

3.3.Comment on endpoint:

3.4.Endpoint units:

Adimensional

3.5.Dependent variable:

Logarithm of octanol/water partition coefficient (log P)

3.6.Experimental protocol:

EC A.8 Partition Coefficient

OECD 123 Partition Coefficient (nOctanol/Water): Slow-Stirring Method

OECD 117 Partition Coefficient (n-octanol/water) HPLC Method

OECD 107 Partition Coefficient (noctanol/water); Shake Flask Method

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Linear regression based on fragments/atom contribution and correction factors.

4.2.Explicit algorithm:

Regression equation based on the hydrophobicity contribution of 120 atom types

The Meylan/Kowwin model in VEGA 1.4.4 is an implementation of the atom fragment contribution (AFC) method described by Meylan et al., 1995. It is a “reductionist” approach and was developed via multiple linear regressions of reliable, experimental log P values. The regressions were performed in two separate and sequential stages.

The first regression correlated “atom/fragment values” with log P. Each nonhydrogen atom (e.g. carbon, nitrogen, oxygen, sulfur) in a structure is a “core” for a fragment; the exact fragment is determined by what is connected to the atom. In some cases, entire functional groups are treated as core “atoms”. A contribution in terms of log P was assigned to each fragment.

A second regression correlated “correction factors” to log P. In general, the correction factors are values for various steric interactions, hydrogen bondings, and effects from polar functional substructures. Their values are derived from the differences between the log P estimates from fragments alone and the measured log P values. The log P of a compound is then estimated by simply summing all atoms/fragment values and correction factors contained in a structure.

The equation reported in Meylan et al., 1995 that describes the log P is the following:

$$\log P = \mathrm{Sum}_i(f_i n_i) + \mathrm{Sum}_i(c_i n_i) + 0.229$$

Where f_i are coefficients for various atom/fragments identified and c_j are correction factors.

4.3.Descriptors in the model:

List of fragments and correction factors (see 4.2)

4.4.Descriptor selection:

See 4.2

4.5.Algorithm and descriptor generation:

See 4.2

4.6.Software name and version for descriptor generation:

4.7.Chemicals/Descriptors ratio:

8364 / 120 = 69.7

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model implemented in VEGA v. 1.4.4 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 > 0.9 strongly similar compounds with known experimental value in the training set have been found.
- 0.9 >= index > 0.75 only moderately similar compounds with known experimental value in the training set have been found.
- 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 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 <

0.5 accuracy of prediction for similar molecules found in the training set is good 0.5 <= index < 1.0 accuracy of prediction for similar molecules found in the training set is not optimal.- index > 1.0 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 < 0.5 similar molecules found in the training set have experimental values that agree with the target compound predicted value.- 0.5 <= index < 1.0 similar molecules found in the training set have experimental values that slightly disagree with the target compound predicted value.- 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 falls in an area of the model's space where the model gives reliable predictions without any outlier value. Defined intervals are:- index < 0.5 the maximum error in prediction of similar molecules found in the training set has a low value, considering the experimental variability.

- 0.5 <= index < 1.0 the maximum error in prediction of similar molecules found in the training set has a moderate value, considering the experimental variability.- index >= 1.0 the maximum error in prediction of similar molecules found in the training set has a high value, considering the experimental variability.

· 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 > 0.75 predicted substance could be out of the Applicability Domain of the model.
- index <= 0.75 predicted substance is out of the the Applicability Domain of the model.

5.2.Method used to assess the applicability domain:

5.3.Software name and version for applicability domain assessment:

VEGA v. 1.4.4

https://www.vegahub.eu/portfolio-item/vega-qsar/

5.4.Limits of applicability:

The calculated model has a lower bound of -5.0 log units (all predictions lower than this value are set to -5.0). The AFC method described in Meylan et al., 1995 was developed solely for organic, organosilicon, and selected organic salt compounds; inorganic compounds and their experimental log P values were excluded from consideration.

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: No

Formula: Yes

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:

Two different validation statistics are reported here:

1) Validaton performance on a training set as reported in the reference pubblication (Meylan & Howard, 1995) (2351 compounds)

2) Validation performance of EPISuite implementation (KOWWIN module) referring to a training set of 2447 compounds.

6.6.Pre-processing of data before modelling:

1) Performance reported by Meylan & Howard, 1995 (see 6.7) refers to a dataset of 2351 compounds. The dataset was built from an initial dataset of 8406 unique organic chemicals with mesured logP values retrieved from reliable sources. This dataset was split into a training set (2351 compounds) and a test set (6055). Compounds with simpler structures were put in the training set while other were included in the test set. The dataset is currently not available.

2) The KOWWIN training set is made of 2447 compounds and it can be downloaded from the Internet at: http://esc.syrres.com/interkow/KowwinData.htm.

Substructure searchable formats of the data can be downloaded at: http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm

6.7.Statistics for goodness-of-fit:

1) Meylan et al., 1995 reported an R2 = 0.982 on their training set of 2351 compounds.

2) KOWWIN User's Guide reports the following statistics on the KOWWIN training set (2447 compounds):

number in dataset = 2447

correlation coef (r2) = 0.982

standard deviation = 0.217

absolute deviation = 0.159

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

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: Yes

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:

All

7.5.Other information about the external validation set:

Three external validation have been performed:

1) Validation performed on a test set of 6055 compounds as reported by Meylan & Howards, 1995.

2) Validation of the KOWWIN implementation performed on a test set of 10,946 compounds.

3) Validation of VEGA implementation of the model, performed on a dataset of 9,961 compounds.

7.6.Experimental design of test set:

1) The dataset from Meylan & Howard 1995 was built from an initial dataset of 8406 unique organic chemicals with mesured logP values retrieved from reliable sources. This dataset was split into a training set (2351 compounds) and a test set (6055). Compounds with simpler structures were put in the training set while other were included in the test set. The dataset is currently not available.

2) The KOWWIN training set is made of 10,946 compounds and it can be downloaded from the Internet at:

http://esc.syrres.com/interkow/KowwinData.htm.

Substructure searchable formats of the data can be downloaded at:
http://esc.syrres.com/interkow/EpiSuiteData_ISIS_SDF.htm

3) The dataset of compounds used for validate VEGA implementation has been built starting from the original dataset provided in EPI suite Meylan/Kowwin model. The set has been processed and cleared from compounds that were replicated or that had problems with the provided molecule structure. The final dataset has 9,961 compounds and is freely available for dowload from model's documentation included in VEGA v. 1.1.4

7.7.Predictivity - Statistics obtained by external validation:

1) Meylan et al., 1995 reported an R2 = 0.943 on a validation set of 6055 compounds.

2) KOWWIN User's Guide reports the following statistics on the KOWWIN training set (10,946 compounds)

number in dataset = 10946
correlation coef (r2) = 0.943
standard deviation = 0.479
absolute deviation = 0.356

3) On the pruned training set from EPI Suite KowWin module (9,961 compounds), the logP model in VEGA has the following statistics: Training set: n = 9961; R2 = 0.86; RMSE = 0.76

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Fragments were selected based on their correlation to log P

8.2.A priori or a posteriori mechanistic interpretation:

A posteriori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

Meylan, W.M. and P.H. Howard, Atom/fragment contribution method for estimating octanol/water partition coefficients. 1995, J. Pharm. Sci. 84: 83-92

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