	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **ALog P model v. 1.0.0 in VEGA v. 1.1.4**
	Printing Date: **14-feb-2020**

1.QSAR identifier

1.1.QSAR identifier (title):

ALog P model v. 1.0.0 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]Arup K. Ghose Department of Molecular Structure and Design, Amgen Inc., 1840 DeHaVilland DriVe, Thousand Oaks, California 91320

[2]Vellarkad N. Viswanadhan Department of Molecular Structure and Design, Amgen Inc., 1840 DeHaVilland DriVe, Thousand Oaks, California 91320

[3]Gordon M. Crippen Department of Molecular Structure and Design, Amgen Inc., 1840 DeHaVilland DriVe, Thousand Oaks, California 91320

2.6.Date of model development and/or publication:

1998

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

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:

Regression equation.

4.2.Explicit algorithm:

Regression equation based on the hydrophobicity contribution of 120 atom types

The AlogP models in VEGA 1.4.4 implement the Ghose-Crippen-Viswanadhan LogP (ALogP) and consists of a regression equation based on the hydrophobicity contribution of 120 atom types.

The classification of the atoms is made to differentiate (i) the electron distribution around the atom, and (ii) the approachability of the solvent to the atom.

Each atom in every structure is classified into one of the 120 atom types. Then, estimated logP for any compound is given by:

AlogP = $\text{Sum}_i(n_i a_i)$

where n_i is the number of atom of type i and a_i is the corresponding hydrophobicity constant..The AlogP model implemented in Dragon was evaluated by the aid of a set of 3568 compounds with known experimental logP taken from the NCI Open DataBase. The resulted determination coefficient r2 was 0.931. Moreover, on our internal logP data set comprised of 9834 compounds the correlation coefficient rbetween experimental and calculated logP was 0.932.

4.3.Descriptors in the model:

The contribution of the 120 atom types. Five, unused atom types are not listed.

4.4.Descriptor selection:

4.5.Algorithm and descriptor generation:

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:

Only compounds containing carbon, hydrogen, oxygen, nitrogen, halogens, and sulfur are considered

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

No

6.2.Available information for the training set:

CAS RN: No

Chemical Name: No

Smiles: No

Formula: No

INChI: No

MOL file: No

6.3.Data for each descriptor variable for the training set:

Unknown

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

Unknown

6.5.Other information about the training set:

6.6.Pre-processing of data before modelling:

6.7.Statistics for goodness-of-fit:

The model coefficients are taken from Ghose et al, J.Phys.Chem. A 1998, 102, 3762-3772. They were estimated on the basis of a training set of 8364 compounds. The statistical parameters of the AlogP model are: r = 0.95; s = 0.55; predictive r ^{2}= 0.90.

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: No

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:

All

7.5.Other information about the external validation set:

7.6.Experimental design of test set:

The training set of the Meylan LogP model (9,961 compounds) from from EPI Suite KowWin module was used as test set during the implementation.

7.7.Predictivity - Statistics obtained by external validation:

Viswanadhan et al., 1993 reports the following statistics obtained on a dataset of 47 nucleosides and bases: n = 47; r = 0.842; SD = 0.51

Ghose et al., 1998 reports the following statistics in validation: n = 931; r = 0.95; s = 0.55; predictive r^2= 0.90.

On the pruned training set from EPI Suite KowWin module (9,961 compounds), the logP model has the following statistics: n = 9961; R2 = 0.84; RMSE = 0.72.

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:

The classification of the atoms is made to differentiate (i) the electron distribution around the atom, and (ii) the approachability of the solvent to the atom.

8.2.A priori or a posteriori mechanistic interpretation:

A priori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]A.K. Ghose and G.M. Crippen, J. Comput. Chem. 1986, 7, 565-577

[2]V.N. Viswanadhan et al., J. Comput. Chem. 1993,14, 1019-1026

[3]A.K. Ghose, V.N. Viswanadhan, J.J. Wendoloski, J.Phys.Chem. A 1998, 102, 3762-3772

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