	QMRF identifier (JRC Inventory): **To be entered by JRC**
	QMRF Title: **ACD/Percepta QSAR for the octanol-water partition coefficient of the neutral form (LogP/GALAS)**
	Printing Date: **13-ott-2017**

1.QSAR identifier

1.1.QSAR identifier (title):

ACD/Percepta QSAR for the octanol-water partition coefficient of the neutral form (LogP/GALAS)

1.2.Other related models:

ACD/Percepta QSAR for the octanol-water partition coefficient of the neutral form (LogP/Classic)

1.3.Software coding the model:

ACD/Percepta 2016

Advanced Chemistry Development, Inc. 8 King Street East, Suite 107, Toronto, Ontario, Canada M5C 1B5

http://www.acdlabs.com/products/percepta/

2.General information

2.1.Date of QMRF:

14.06.2010

2.2.QMRF author(s) and contact details:

[1]Andrius Sazonovas ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania andrius.sazonovas@acdlabs.com www.acdlabs.com

[2]Pranas Japertas ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania pranas.japertas@acdlabs.com www.acdlabs.com

[3]Remigijus Didziapetris ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania www.acdlabs.com

[4]Kiril Lanevskij ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania kiril.lanevskij@acdlabs.com www.acdlabs.com

[5]Justas Dapkunas ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania www.acdlabs.com

2.3.Date of QMRF update(s):

October 2017

2.4.QMRF update(s):

Authors of the update:

Andrius Sazonovas (see Section 2.2 for contacts)

Simona Kovarich, S-IN Soluzioni Informatiche. Via G. Ferrari, 14, I-36100 Vicenza (Italy). email: simona.kovarich@s-in.it

Updated sections: 1.1-3; 2.2; 2.5-9; 3.1; 3.4-5; 4.1-2; 4.4-6; 5.2-4; 6.2; 6.4; 6.7; 7.2; 7.4; 7.7-8; 8.2-3; 9.1-3

2.5.Model developer(s) and contact details:

ACD/Labs, Inc. A.Mickeviciaus g. 29, LT-08117, Vilnius, Lithuania. Phone: +370 5 2624032; fax: +370 5 262 37 28 vilnius@acdlabs.com www.acdlabs.com

2.6.Date of model development and/or publication:

2007.10.25

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

[1]Japertas, P., Sazonovas, A., Didziapetris, R., and Petrauskas, A., Similarity based assessment of model applicability domain and quantitative evaluation of the reliability of the prediction. Abstr. Paper. Am. Chem. Soc., 2008, 235, Meeting abstract 271-COMP
http://oasys2.confex.com/acs/235nm/techprogram/P1160866.HTM
[2]Japertas, P., Sazonovas, A., Didziapetris, R., and Petrauskas, A., Similarity based correction for the predictions of compounds physicochemical properties. Abstr. Paper. Am. Chem. Soc., 2008, 235, Meeting abstract 247-MEDI (Attached as SI)
http://oasys2.confex.com/acs/235nm/techprogram/P1160811.HTM
[3]Clarke, E. D., Delaney, J. S., Japertas, P., and Jurgutis, P., Agrochemicals and log P octanol: Evaluation of structure based predictions, Abstr. Paper. Am. Chem. Soc., 2007, 234, Meeting abstract 64-AGRO (Attached as SI)
http://oasys2.confex.com/acs/234nm/techprogram/P1119638.HTM
[4]Japertas, P., Didziapetris, R., and Petrauskas, A., Fragmental methods in the design of new compounds. Applications of Advanced Algorithm Builder, Quant. Struct.-Act. Relat., 2002, 21, 23-37
http://www3.interscience.wiley.com/journal/93521415/abstract
[5]Mannhold, R. and Petrauskas, A., Substructure versus whole-molecule approaches for calculating Log P, QSAR Combi. Sci., 2003, 22, 466-475
http://www3.interscience.wiley.com/journal/104539390/abstract

2.8.Availability of information about the model:

Model is proprietary, training and test set are not available. However the compounds used to derive the model and their experimental data are available within the corresponding software products.

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

No

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Not applicable - Physicochemical property

3.2.Endpoint:

QMRF 1. Physical Chemical Properties QMRF 1. 6. Octanol-water partition coefficient (Kow)

3.3.Comment on endpoint:

The logarithm of a ratio of concentrations of un-ionized compound between its solutions in n-octanol and water: log Kow= log ([solute in octanol]/[un-ionized solute in water])

3.4.Endpoint units:

Dimentionless since this is a ratio of concentrations

3.5.Dependent variable:

log Kow - this is the Log to the base ten of the octanol-water coefficient

3.6.Experimental protocol:

Experimental protocol:
The dataset used to develop the reported model has been compiled from a great number of different sources covering a wide variety of experimental protocols used to determine log Kow values reported within

them. This includes the classical potentiometric log Kow determination methods involving phase titrations, as well as more contemporary and most modern chromatographic methods like HPLC on standard and modified (immobilized artificial membrane (IAM) and liposome chromatography) resins or capillary electrophoresis and centrifugal partition chromatography. Since log Kow takes into account only partition of neutral species, when the method involves only single data point measurement (i.e. the log Kow is not determined by extrapolation from a pH dependence curve), the water phase is usually buffered to a pH in which the predominant state of the analyzed compound is neutral. For a comprehensive overview of the experimental log Kow measurement techniques please see [1].

3.7.Endpoint data quality and variability:

log Kow is a relatively easily measured property. As a result the experimental data quality, which is usually inversely proportional to the complexity of the experiment, is reasonably good. Independent external studies show that the error between the log Kow measurements performed by different laboratories using the same protocol (reproducibility) can be expected to be within 0.5 logarithmic units [2]. Experimental data from various sources have been used. The characteristics of the entire dataset compiled for the task of this model development is:

No. of compounds = 16277
Min. Value = -5.08
Max. Value = 11.29
Std. Dev. = 1.92
Skewness = 0.22

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Hybrid QSAR, combining a linear baseline model utilizing PLS method and the local similarity based corrections

4.2.Explicit algorithm:

GALAS (Global, Adjusted Locally According to Similarity) algorithm
Global linear baseline QSAR + local similarity based corrections
The global QSAR was developed using PLS in combination with bootstrapping technique. This method implies random compound sampling from the initial training set, i.e. generation of new “training sub-sets”. Each of the sampled sub-sets is of the same size as the initial training set, however, random manner of their population results in some compounds being selected more than once, others being omitted. This procedure is performed 100 times and an independent PLS model is derived for every sub-set.

Each of those PLS models is based on 2D fragmental descriptors:

log Kow = $SUM_{[i=1..n]}(a_i*f_i)$

+ c

where f_i is the number of occurences of the i-th fragment in a molecule, a_i - its statistical coefficient, and c - intercept.

As a result, each global QSAR model actually represents an ensemble of 100 PLS models, providing each compound with a vector of 100 log Kow predictions, each based on a slightly different sub-set of the initial training set. It is defined that two compounds with similar trends in the variation patterns of the 100 value vectors predicted by a global QSAR model are considered similar in terms of the analyzed property, i.e. the differences in the compound sets used to parameterize each of 100 PLS models, constituting a baseline model, affect estimations for the two compounds in a similar way. The correlation coefficient of the two vectors is called an Individual Similarity Index between two compouds (SI_i). An analogous definition of the “property-specific” or dynamic similarity was first used by Tetko and his co-workers [3-7] and this method has been recently used in the analysis of the acute toxicity data [8].

With the available robust similarity measure, it becomes possible to analyse the performance of the baseline QSAR model in the local chemical environment of a query molecule represented by the most similar compounds in the training set. In case any systematic errors are encountered for sufficiently similar compounds, a local correction (*delta*) is calculated.

Later on it is possible to train the model quickly and efficiently using new experimental data by just adding it to this second similarity correction calculation procedure, without the time costly baseline model re-training.

4.3.Descriptors in the model:

Fragmental descriptors dimentionless (occurence count) Fixed set of fragmental descriptors, based on the expanded list of Platt's type fragments (see [9]). A fixed and relatively small set of fragments was used due to the specifics of the employed modeling methodology. In order for the correlation between two compound vectors of log Ko/w predictions coming from a baseline QSAR model to be representative of compound similarity in terms of the analyzed property, these vectors have to be parameterized using exactly the same set of fragmental descriptors. This prevents the use of any sort of automated fragmentation routines (atom based, isolating carbon based, chain based, etc.) that result in a dynamic set of fragments depending on the training set structures. They leave the possibility that for any query structure from outside the training set the same rules will yield certain new fragments not encountered in the training set molecules which is not compatible with the main condition just mentioned. On the other hand, it is equally important for the model to be able to identify any new structural features of a query molecule that were not present in the training set compounds. I.e., the fixed fragment set cannot be constructed based on the analysis of the training set either, or in general any molecule set whatsoever. Because in that case any new structural features not present in that database would be eventually ignored. As a result, the fragmental descriptor set is based on the general knowledge and considerations regarding all possible chemical structures rather than a finite dataset and include all the fragments, even those that are not detected

in the training set molecules at all.

4.4.Descriptor selection:

No special descriptor selection techniques had been used to reduce the initial descriptor pool of 377 fragments (e.g., excluding statistically insignificant or intercorrelated variables) due to the specifics of employed modeling methodology. Any potential negative influence of insignificant fragments would be remedied by the use of PLS method, but their presence is necessary for providing the so called "dynamic similarity" measure between the molecules. For this purpose, even “blank” fragments (with zero occurrence count) should remain, as these would allow detecting new structural features of a query molecule that were not present in the training set, and would thus decrease its similarity coefficient to training set molecules.

4.5.Algorithm and descriptor generation:

The generation of the descriptor matrix following the outlined approach constituted counting the occurences of any of the pre-defined fragments in the trainig set molecules. This procedure as well as all the subsequent statistical analysis were performed using Algorithm Builder 1.8 software. For the descriptor generation the compound set has to be imported into the native database format of the Algorithm Builder software via one of the supported mechanisms:

1. Copy/Pasting structures one-by-one from an external editor
2. Importing a collection of MOL files
3. Importing an SD file
4. Importing a tab-delimited TXT file with SMILES

4.6.Software name and version for descriptor generation:

Algorithm Builder 1.8

Advanced Chemistry Development, Inc. 8 King Street East, Suite 107, Toronto, Ontario, Canada M5C 1B5

http://www.acdlabs.com

4.7.Chemicals/Descriptors ratio:

30.2 (11387 chemicals in the training set, 377 descriptors)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain of the model is defined based on the training set compounds. This procedure takes into account the following two aspects:

* Similarity of the tested compound to the training set. No reliable predictions can be made if we have no similar compounds in the training set;

* Consistence of the experimental values with regard to the baseline model for similar compounds. Even if we do have similar compounds in the dataset the quality of prediction could be lower if that data cannot be reproduced by the baseline model. It does not matter what the reason for this inconsistency – experimental variability or sudden change in mechanism of action because of slight structural changes – in any case

it indicates possible problems when trying to give accurate predictions

5.2.Method used to assess the applicability domain:

The two aspects mentioned in Section 5.1 receive their quantitative assessment in terms of Similarity Index (*SI*) and Data-Model Consistency Index (*DMCI*).

The*SI*, evaluating how distant the query structure is from the whole training set, is calculated by weighted averaging of all the individual Similarity Indices (SI_i) for the test molecule and each of the 5 most similar compounds from the training set, calculated using "property-specific" similarity approach as explained in Section 4.2.*DMCI*is calculated by comparing the differences between experimental and global QSAR predicted values for the 5 most similar compounds and the suggested similarity correction value (*delta*) for the test compound, calculated by averaging these differences. The more individual differences are scattered around the calculated average (*delta*), the more inconsistent are the data for the similar compounds with regards to the global QSAR model.

The final prediction Reliability Index is calculated as a product of the aforementioned two indices:

*RI = SI * DMCI*Both*SI*and*DMCI*are scaled to vary from 0 to 1, so the resulting*RI*also varies in this range. Lower values suggest a compound being further from the Model Applicability Domain and the prediction less reliable (low*SI*or low*DMCI*either alone or in combination can be the reason). On the other hand, high*RI*values indicate an increasing confidence about the quality of the prediction (both*SI*and*DMCI*have to be high to yield such a result).

5.3.Software name and version for applicability domain assessment:

ACD/Percepta 2014

Advanced Chemistry Development, Inc. 8 King Street East, Suite 107, Toronto, Ontario, Canada M5C 1B5

http://www.acdlabs.com/products/percepta/

5.4.Limits of applicability:

Reliability Index < 0.3: unreliable predictions

Reliability Index in the range 0.3-0.5: borderline reliability of predictions

Reliability Index in the range 0.5-0.75: moderate reliability of predictions

Reliability Index >0.75: high reliability of predictions

6.Internal validation - OECD Principle 4

6.1.Availability of the training set:

Yes

6.2.Available information for the training set:

The training set is available through the software products listed here, but is not attached to the form itself

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:

No

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

No

6.5.Other information about the training set:

The statistics of the training set data:

No. of compounds = 11387

Min. Value = -5.08

Max. Value = 11.29

Std. Dev. = 1.94

Skewness = 0.25

6.6.Pre-processing of data before modelling:

None

6.7.Statistics for goodness-of-fit:

Statistics provided for the fraction of the training set that falls within the aplicability domain of the model (*RI*> 0.3 - see Section 5.4)

$N_{RI>0.3}$= 11371 (i.e. 99.9% of the training set compounds)

R^2= 0.944

RMSE= 0.457

F= 402696.2 (Fisher's F-statistics)

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

N/A

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

N/A

6.10.Robustness - Statistics obtained by Y-scrambling:

N/A

6.11.Robustness - Statistics obtained by bootstrap:

N/A

6.12.Robustness - Statistics obtained by other methods:

N/A

7.External validation - OECD Principle 4

7.1.Availability of the external validation set:

Yes

7.2.Available information for the external validation set:

The training set is available through the software products listed here, but is not attached to the form itself

CAS RN: No

Chemical Name: No

Smiles: No

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:

The statistics of the validation set data:

No. of compounds = 4890

Min. Value = -4.64

Max. Value = 10.89

Std. Dev. = 1.90

Skewness = 0.16

7.6.Experimental design of test set:

Random splitting of the initial dataset into the training and validation sets using the ratio 70%:30%.

7.7.Predictivity - Statistics obtained by external validation:

Statistics provided for the fraction of the validation set that falls within the aplicability domain of the model (*RI*> 0.3 - see Section 5.4)

$N_{RI>0.3}$= 4872 (i.e. 99.6% of all the validation set compounds)

R^2= 0.940

RMSE= 0.464

F= 165247.5 (Fisher's F-statistics)

Analysis of the subsets of the higher quality results$N_{RI>0.5}$= 4772 (i.e. 97.6% of all the validation set compounds)R^2= 0.945

RMSE= 0.444*F*= 177716.6 (Fisher's F-statistics)

$N_{RI>0.75}$= 3345 (i.e. 68.4% of all the validation set compounds)

R^2= 0.964*RMSE*= 0.360

F= 197041.9 (Fisher's F-statistics)

7.8.Predictivity - Assessment of the external validation set:

As can be seen from the results of the Section 7.7 - almost the entire validation set is within the Applicability Domain of the reported model.

7.9.Comments on the external validation of the model:

Correlation coefficients and other statistical parameters for the training and test set compounds falling within the applicability domain of the model are in a very good agreement.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The only mechanistic consideration utilized in model building is the use of a linear regression method (PLS) and the fragmental descriptors. In

other words it is assumed that the final predicted value is composed of a linear combination of all the contributions of structural moieties making up the test molecule. Although very basic, this consideration is one of the most fundamental ones, even the name of (Q)SAR methods implies that the main determinant of all the properties of a compound is its structure. Quite obviously fragments are the best and realy first-hand descriptors of a chemical structure.

8.2.A priori or a posteriori mechanistic interpretation:

The mechanistic interpretation is given a posteriori.

8.3.Other information about the mechanistic interpretation:

A posteriori model interpretation results are consistent with generaly understood mechanistic factors or scientific interpretations and well documented experimental facts. I.e., the top ten fragmental descriptors with negative coefficients are the following:
Any positive permanent charge = -2.436
Quaternary ammonium = -1.612
Permanent charge on aromatic N, O, S, Se = -1.317
Sulfonic acid = -1.125
alpha-Amino acid = -0.965
N-oxide = -0.674
tertiary amine (>N-) = -0.673
=S< = -0.670
Any phosphorus atom = -0.573
Lactone = -0.404

Some of those fragments are very well known because of their effect of increasing hydrophilicity of a compound. One more classical example of such water phase favorable group, i.e., hydroxy fragment, follows this TOP10 almost immediately with a statistical coefficient of -0.400
Among the groups with the largest positive coefficients, the absolute majority of them can be clearly expected to increase the hydrophobic properties of a compound, e.g.:Bicyclo [3.1.1] scaffold = 1.103
Spiro [5.2] scaffold = 1.066Any Si atom = 0.714
Spiro [6.6] = 0.678
Spiro [6.5] = 0.644Fused 6:5:5 scaffold = 0.614
Stereohindrance in the form of two bulk branched aliphatic substituents in both orto- positions of a phenolic group = 0.460n-Pentyl chain = 0.452n-Heptyl chain = 0.442
Aromatic sulphur = 0.419

Note: the average of all 377 statistical coefficients is 0.018
All the fragments encoding strong H-bonding in the aromatic system (e.g., orto-keto, orto-thioketo, orto-nitro, or orto-halogenated phenols and anilines - 6 descriptors in total) have positive coefficients which is in agreement with the known fact that H-Bonding reduces hydrophilicity. The coefficients of 6 fragments mentioned range from +0.005 to +0.455 with an average of +0.15.

Further similar examples can be established as well.

9.Miscellaneous information

9.1.Comments:

Together with the prediction, ACD/Percepta displays 5 most similar structures from the training set along with experimental results and references, to illustrate particular data using which the similarity correction part of the algorithm has been executed. The similarity is measured in terms of “property-specific” similarity.

9.2.Bibliography:

[1]Avdeev, A., Absorption and Drug Development: Solubility, Permeability, and Charge State, John Wiley & Sons, Inc., Hoboken, NJ, 2003.

[2]Kishi, H. and Hashimoto, Y., Evaluation of the procedures for the measurement of water solubility and n-octanol/water partition coefficient of chemicals results of a ring test in Japan, Chemosphere, 1989, 18, 1749-1759.

[3]I.V. Tetko, Neural network studies. 4. Introduction to associative neural networks, J. Chem. Inf. Comput. Sci. 2002, 42, 717-728.

[4]I.V. Tetko and P. Bruneau, Application of ALOGPS to predict 1-octanol/water distribution coefficients, logP, and logD, of AstraZeneca in-house database, J. Pharm. Sci. 2004, 93, 3103-3110.

[5]I.V. Tetko and V.Y. Tanchuk, Application of associative neural networks for prediction of lipophilicity in ALOGPS 2.1 program, J. Chem. Inf. Comput. Sci. 2002, 42, 1136-1145.

[6]H. Zhu, A. Tropsha, D. Fourches, A. Varnek, E. Papa, P. Gramatica, T. Oberg, P. Dao, A. Cherkasov, and I.V. Tetko, Combinatorial QSAR modeling of chemical toxicants tested against Tetrahymena pyriformis, J. Chem. Inf. Model. 2008, 48, 766-784.

[7]I.V. Tetko, I. Sushko, A.K. Pandey, H. Zhu, A. Tropsha, E. Papa, T. Oberg, R. Todeschini, D. Fourches, and A. Varnek, Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: focusing on applicability domain and overfitting by variable selection, J. Chem. Inf. Model. 2008, 48, 1733-1746.

[8]Sazonovas, A., Japertas, P., and Didziapetris, R., Estimation of reliability of predictions and model applicability domain evaluation in the analysis of acute toxicity (LD50), SAR QSAR Environ. Res. 2010, 21, 127-148.

[9]J.A. Platts, D. Butina, M.H. Abraham, and A. Hersey, Estimation of molecular linear free energy relation descriptors using a group contribution approach, J. Chem. Inf. Comput. Sci. 1999, 39, 835-845.

9.3.Supporting information:

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

Japertas_et_al_2008_Similarity_Corrections.pdf	file:///C:\Users\Simona\Documents\S-IN\ATTIVITA'\QMRF\QMRF_ACD\20160122_QMRF_to_ACD\QMRF_ACD_Physchem\LogP_Galas\Japertas_et_al_2008_Similarity_Corrections.pdf
Clarke_et_al_2007_LogP.pdf	file:///C:\Users\Simona\Documents\S-IN\ATTIVITA'\QMRF\QMRF_ACD\20160122_QMRF_to_ACD\QMRF_ACD_Physchem\LogP_Galas\Clarke_et_al_2007_LogP.pdf

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