1.QSAR identifier

1.1.QSAR identifier (title):

Multilinear QSPR Model for aqueous solubility (Log S)

1.2.Other related models:

-

1.3.Software coding the model:

QSARModel 4.0.3 Molcode Ltd., Turu 2, Tartu, 51014, Estonia http://www.molcode.com

2.General information

2.1.Date of QMRF:

01.04.2011

2.2.QMRF author(s) and contact details:

[1]Indrek Tulp Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[2]Tarmo Tamm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[3]Gunnar Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[4]Dimitar Dobchev Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[5]Kaido Tämm Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[6]Jaak Jänes Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[7]Eneli Härk Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com
[8]Mati Karelson Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

Molcode model development team Molcode Ltd. Turu 2, Tartu, 51014, Estonia models@molcode.com http://www.molcode.com

2.6.Date of model development and/or publication:

01.04.2011

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

[1]Karelson M, Dobchev D, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Karelson G (2008). Correlation of blood-brain penetration

and human serum albumin binding with theoretical descriptors. ARKIVOC 16, 38-60.

[2]Karelson M, Karelson G, Tamm T, Tulp I, Jänes J, Tämm K, Lomaka A, Savchenko D & Dobchev D (2009). QSAR study of pharmacological permeabilities. ARKIVOC 2, 218–238.

2.8.Availability of information about the model:

All information in full detail is available

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

None to date

3.Defining the endpoint - OECD Principle 1

3.1.Species:

n/a

3.2.Endpoint:

1.Physicochemical effects 1.3.Water solubility

3.3.Comment on endpoint:

The aqueous solubility of drug compounds is one of the most important factors in determining its biological activity.In many cases drugs that show a good activity whenadministered parenterally maybe totally inactive when given orally. In such cases poor oral activity is often due to the fact that a sufficient amount of drug to desired response is not reached in the site of action. Hence an insufficient aqueous solubility is likely to hamper bioavailability of the drugs. [1]

3.4.Endpoint units:

S[mol/l]

3.5.Dependent variable:

LogS

3.6.Experimental protocol:

A set of 1297 diverse organic compounds was extracted from two databases [2,3] and was divided into a training set of 775 compounds and a external test set of 515 compounds (selected by taking into account the distribution of the LogS). The aqueous solubility values in 20-25 °C expressed as log S, where S is solubility in mol/L, were used. [1]

3.7.Endpoint data quality and variability:

The applicability and accuracy of a log S estimation method are strongly affected by the size and quality of the training set used. Experimental aqueous solubility values for the compounds used in this study were obtained from the AQUASOL dATAbASE of the University of Arizona[2] and SCR’s PHYSPROP Database [3].

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

2D and 3D regression-based QSAR

4.2.Explicit algorithm:

multilinear regression QSAR

multilinear regression QSAR derived with BMLR (Best Multiple Linear Regression) method

Log S = -0.01646513+ 0.8229343*D1 -0.02988266*D2+ 1.58692662*D3 -10.48291061*D4

4.3.Descriptors in the model:

[1]D1 - Average bond order (AM1) for O atoms [unitless]
[2]D2 - Molecular surface area (AM1) [Å2]
[3]D3 - Square root of Charged (Zefirov) Surface Area of H atoms [Å2/3]
[4]D4 - Max net atomic charge (Zefirov) for N atoms [au]

4.4.Descriptor selection:

Initial pool of ~1000 descriptors. Stepwise descriptor selection based on a set of statistical selection rules (one-parameter equations: Fisher criterion and R2 over threshold, variance and t-test value over threshold, intercorrelation with another descriptor not over threshold),

(two-parameter equations: intercorrelation coefficient below threshold, significant correlation with endpoint, in terms of correlation coefficient and t-test)

Stepwise trial of additional descriptors not significantly correlated to any already in the model.

4.5.Algorithm and descriptor generation:

1D, 2D, and 3D theoretical calculations. Quantum chemical descriptors derived from AM1 calculation. Model developed by using multilinear regression.

4.6.Software name and version for descriptor generation:

QSARModel 4.0.3
QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling
Molcode Ltd, Turu 2, Tartu, 51014, Estonia
http://www.molcode.com

4.7.Descriptors/Chemicals ratio:

193

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:

By chemical identity: very diverse organic chemicals (incl. aliphatic and (hetero)aromatic compounds, alcohols, esters, carboxylic acids, amines, amides, aldehydes, ketones, phenols, etc)

By descriptor value range (between min and max values): The model is suitable for compounds that have the descriptors in the following range:

Descs	D1	D2	D3	D4
Min	0.000	59.160	0.000	-0.127
Max	1.942	571.920	6.432	0.055

5.2.Method used to assess the applicability domain:

Range of descriptor values in training set with ±30% confidence. Descriptor values must fall between maximal and minimal descriptor values of training set ±30%.

5.3.Software name and version for applicability domain assessment:

QSARModel 4.0.3
QSAR/QSPR package that will compute chemically meaningful descriptors and includes statistical tools for regression modeling
Molcode Ltd, Turu 2, Tartu, 51014, Estonia
http://www.molcode.com

5.4.Limits of applicability:

See 5.1 and 5.2

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:Yes

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:

775 datapoints

6.6.Pre-processing of data before modelling:

none

6.7.Statistics for goodness-of-fit:

R2=0.8427, R2adj=0.8419, s2=0.6605

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

R2cv=0.8404

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

6.10.Robustness - Statistics obtained by Y-scrambling:

R2Y= 0.00617788

6.11.Robustness - Statistics obtained by bootstrap:

R2X= 0.00595687, R2XY= 0.00685961

6.12.Robustness - Statistics obtained by other methods:

ABC analysis (2:1 training : prediction) on sorted (in increased order of endpoint value) data divided into 3 subsets (A;B;C). Training set formed with 2/3 of the compounds (set A+B, A+C, B+C) and validation set consisted of 1/3 of the compounds (C, B, A).

AB, BC, AC - R2 = 0.8437932; AB, BC, AC - R2cv = 0.84024963; A, B, C - R2pred = 0.84225220

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:No
Formula:Yes
INChI:No
MOL file:Yes

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

All

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

All

7.5.Other information about the external validation set:

515 data points

7.6.Experimental design of test set:

From sorted data each 3th was subjected to the test set.

7.7.Predictivity - Statistics obtained by external validation:

R2 = 0.79 (Coefficient of determination)

7.8.Predictivity - Assessment of the external validation set:

Heaving in mind that the external test set consists of huge number of data points, the R2 in 7.7 shows significant prediction of the model

7.9.Comments on the external validation of the model:

The validation coefficient of determination (R2) is close to internal validation (R2CV).

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

As it is very well known, the water solubility depends greatly on hydrogen acceptor/donor ability (including charge distributions) of the compounds as well as its bulk properties. The molecular surface area is one of the most significant descriptors in our model. It reflects the bulk properties of the compounds. The very famous sentence "similar dissolves in similar" holds for present case. Since water is a small molecule compound, it is expected that the smaller compounds will dissolve better in watter. The molecular surface area has negative correlation r = -0.68 with LogS that indicates the above conjecture.

The hydrogen abilities of the compounds are encoded in Average bond order (AM1) for O atoms, Square root of Charged (Zefirov) Surface Area of H atoms and Maximal net atomic charge (Zefirov) for N atoms descriptors.

8.2.A priori or a posteriori mechanistic interpretation:

a posteriori mechanistic interpretation, consistent with published scientific interpretations of experiments

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

-

9.2.Bibliography:

[1]J. Huuskunen, stimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology, J. Chem. Inf. Comput. Sci. 2000, 40, 773 777.

[2]Yalkowsky, S. H.; Dannelfelser, R. M. The ARIZONA dATAbASE of Aqueous Solubility; College of Pharmacy, University of Arizona: Tucson, AZ, 1990.

[3]Syracuse Research Corporation. Physical/Chemical Property Database (PHYSOPROP); SRC Environmental Science Center: Syracuse, NY, 1994

9.3.Supporting information:

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

Karelson Arkivoc 2008	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf83_Karelson Arkivoc 2008.pdf
Karelson Arkivoc 2009	http://qsardb.jrc.it:80/qmrf/download_attachment.jsp?name=qmrf83_Karelson Arkivoc 2009.pdf

10.Summary (ECB Inventory)

10.1.QMRF number:

10.2.Publication date:

10.3.Keywords:

10.4.Comments: