1.QSAR identifier

1.1.QSAR identifier (title):

QSAR Model for Organic carbon-sorption partition coefficient (logKoc)

1.2.Other related models:

1.3.Software coding the model:

QSARModel 4.0.4 Molcode Ltd., Turu 2, Tartu, 51014, Estonia http://www.molcode.com

2.General information

2.1.Date of QMRF:

10.12.2010

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:

22.10.2010

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:

2.Environmental fate parameters 2.6.Organic carbon-sorption partition coefficient (organic carbon; Koc)

3.3.Comment on endpoint:

The adsorption coefficient (Koc) on soil was estimated using the OECD Test Guideline TG 121 (EU C.19). The adsorption coefficient normalized to the organic carbon content of the soil Koc is a useful indicator of the binding capacity of a chemical on organic matter of soil and sewage sludge and allows comparisons to be made between different chemicals. Koc = Kd/foc or Koc = Kf/foc, where Kd is distribution coefficient, Kf is Freundlich adsorption coefficient and foc is organic carbon content of a sorbent [1].

3.4.Endpoint units:

unitless

3.5.Dependent variable:

logKoc

3.6.Experimental protocol:

The application of HPLC screening has become an accepted tool to estimate reliably soil adsorption coefficients of many organic chemicals. The method is based on the similitude between soil and a chromatographic column, and correlates soil adsorption coefficients with the HPLC-retention behavior expressed as capacity factor [2,3]. The experimental data of the soil sorption partition coefficient, normalized on organic carbon Koc, of 643 heterogeneous organic compounds were collected from the literature [4,5,6] and compiled into a single database.

3.7.Endpoint data quality and variability:

Experimental data from different sources [4-6] was compiled into a single dataset. Some of the chemicals in the literature databases have more than one Koc value, the result of being derived from different sources; in these cases the median was adopted. The data set is highly heterogeneous, and includes practically all the principal functional groups present in pesticides and various organic pollutants.

Statistics:
max value: 6.30
min value: -0.300
standard deviation: 1.20
skewness: 0.311

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

logKoc = 0.340
+0.294*Kier&Hall index (order 0)
+9.865E-002*Lowest exchange energy (AM1) for C - O bonds
+12.897*Relative number of benzene rings
-0.140*Topographic electronic index (AM1) all bonds

4.3.Descriptors in the model:

[1]Kier&Hall index (order 0) [unitless] zeroth order Kier and Hall valence connectivity index
[2]Lowest exchange energy (AM1) for C - O bonds [eV] lowest exchange energy between bonded C and O atoms
[3]Relative number of benzene rings [unitless] number of benzene rings divided by number of total atoms
[4]Topographic electronic index (AM1) all bonds [au/Å2] topolographical electronic index calculated over all bonds between non-hydrogen atom in the molecule and based on AM1 calculations

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 4.0.4
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.Chemicals/Descriptors ratio:

80.5 (322 chemicals / 4 descriptors)

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

Applicability domain based on training set:

a) by chemical identity: heterogeneous organic compounds (aliphatic, cyclic and aromatic hydrocarbons, carbonyl compounds, amines, halogeno derivatives, containing sulfur, phosphorus and other heteroatoms, etc)

b) by descriptor value range: The model is suitable for compounds that have the descriptors
in the following minimal-maximal range:
Kier&Hall index (order 0): 1.82 - 18.6
Lowest exchange energy (AM1) for C - O bonds: -10.5 - 0
Relative number of benzene rings: 0 - 0.158
Topographic electronic index (AM1) all bonds: 0.00391 - 14.0

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 4.0.4
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

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:

Training set consist 322 data points.

1 negative values

321 positive values

6.6.Pre-processing of data before modelling:

n/a

6.7.Statistics for goodness-of-fit:

R2 = 0.806 (Correlation coefficient)

s2 = 0.534 (Standard error of the estimate)

F = 329 (Fisher function)

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

R2CV = 0.800

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

R2CVMO = 0.799

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:

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).

average R2 (fitting) = 0.807

average R2 (prediction) = 0.802

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:

Test set consists 321 data points,

321 negative values,

0 positive values

7.6.Experimental design of test set:

From sorted data source each 2nd was subjected to the test set.

7.7.Predictivity - Statistics obtained by external validation:

R2 = 0.791 (Coefficient of determination)

7.8.Predictivity - Assessment of the external validation set:

All compounds in test set are in range of defined applicability domain:

Kier&Hall index (order 0): 1.12 - 22.5

Lowest exchange energy (AM1) for C - O bonds: -11.0 - -5.05

Relative number of benzene rings: 0 - 0.176

Topographic electronic index (AM1) all bonds: 0.00 - 15.6

7.9.Comments on the external validation of the model:

The validation coefficient of determination (R2) is good and very close to the squared correlation coefficient of the model (R2 = 0.806). This shows good stability of the model.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

Soil sorption is closely related to water solubility and hydrophobicity (logKow). "Kier&Hall index (order 0)" represents size and branching of molecule accounting presence of heteroatoms. This descriptor is closely related to hydrophobicity (logP). "Relative number of benzene rings" exhibit presence of hydrophobic pi electron systems which are unfavorable for water solubility. Descriptors "Lowest exchange energy (AM1) for C - O bonds" and "Topographic electronic index (AM1) all bonds" are related to charges and to charge distribution contributing into water solubility and chemical stability of compounds.

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:

Interpretation is in general agreement with literature [7].

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Estimation of the adsorption coefficient (Koc) on soil and on sewage sludge using high performance liquid chromatography (HPLC), OECD TG 121, 2001.

[2]Kördel W., Stutte J., Kotthoff G. HPLC-screening method for the determination of the adsorption coefficient on soil - comparison of different stationary phases. Chemosphere 1993, 27, 2341-2352. http://dx.doi.org/10.1016/0045-6535(93)90257-6

[3]Kördel W., Stutte J., Kotthoff G. HPLC-screening method to determine the adsorption coefficient in soil - comparison of immobilized humic acid and clay mineral phases for cyanopropyl columns. Sci. Tot. Environ. 1995, 162, 119-125. http://dx.doi.org/10.1016/0048-9697(95)04443-5

[4]Sabljic A., Güsten H., Verhaar H., Hermens J. QSAR modeling of soilsorption. Improvements and systematics of log Koc vs log Kow correlations. Chemosphere 1995, 31, 4489-4514. http://dx.doi.org/10.1016/0045-6535(95)00327-5

[5]Tao S., Piao H., Dawson R., Lu X., Hu H. Estimation of organic carbon normalized sorption coefficient (Koc) for soils using the fragment constant method, Environ. Sci. Technol. 1999, 33, 2719–2725. http://dx.doi.org/10.1021/es980833d

[6]Huuskonen J. Prediction of soil sorption coefficient of a diverse set of organic chemicals from molecular structure, J. Chem. Inf. Comput. Sci. 2003, 43, 1457–1462. http://dx.doi.org/10.1021/ci020342j

[7]Statistical external validation and consensus modeling: A QSPR case study for Koc prediction, Paola Gramatica, Elisa Giani, Ester Papa, Journal of Molecular Graphics and Modelling, Volume 25, Issue 6, March 2007, Pages 755-766 http://dx.doi.org/10.1016/j.jmgm.2006.06.005

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: