	QMRF identifier (JRC Inventory): **Q17-31-0047**
	QMRF Title: **BIOVIA toxicity prediction model – acute toxicity to Daphnia**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

BIOVIA toxicity prediction model – acute toxicity to Daphnia

1.2.Other related models:

No related models

1.3.Software coding the model:

BIOVIA Discovery Studio v4.5

Optimize your drug discovery process with a flexible application that delivers predictive science to its required depth.

Dassault Systèmes, BIOVIA Corp., 5005 Wateridge Vista Drive, San Diego, CA92121, USA http://www.3dsbiovia.com

2.General information

2.1.Date of QMRF:

13/5/2015

2.2.QMRF author(s) and contact details:

Deqiang Zhang Dassault Systemes, BIOVIA Corp. 5005 Wateridge Vista Drive, San Diego, CA 92121, USA Deqiang.Zhang@3ds.com http://www.3dsbiovia.com

2.3.Date of QMRF update(s):

N/A

2.4.QMRF update(s):

N/A

2.5.Model developer(s) and contact details:

Deqiang Zhang Dassault Systemes, BIOVIA Corp. 5005 Wateridge Vista Drive, San Diego, CA 92121, USA Deqiang.Zhang@3ds.com http://www.3dsbiovia.com

2.6.Date of model development and/or publication:

2015

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

BIOVIA Discovery Studio v4.5 http://www.3dsbiovia.com/products/discovery-studio/

2.8.Availability of information about the model:

The model is proprietary (available as a commercial product), but the algorithm and data are public. The training set is made available and is also embedded with the model and can be retrieved with similarity search when a prediction is conducted. No external test is conducted except cross-validation.

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

None

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Daphnia Magna (Crustacea)

3.2.Endpoint:

3.Ecotoxic effects 3.1.Short-term toxicity to Daphnia (immobilisation)

3.3.Comment on endpoint:

The model predicts the Daphnia magna EC50 (the effect concentration of a substance that causes adverse effects on 50% of the test poulation Daphnia magna) value of a chemical in an aquatic toxicity test (48 hours).

3.4.Endpoint units:

EC50 is usually expressed as the molar concentration(mole per litre).

3.5.Dependent variable:

pEC50 = -log(EC50)

3.6.Experimental protocol:

The experiement protocol is outlined in OECD Guidelines for the Testing of Chemicals, Section 2 / Test No. 202: Daphnia sp. Acute Immobilisation Test. Available online at

3.7.Endpoint data quality and variability:

N/A

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

Partial least squares regression

4.2.Explicit algorithm:

Partial least squares regression

Partial least squares regression is a multivariate linear regression method that takes into account the latent structure in both the dependent variable and the explanatory variables. The true regression is done on a small number of latent variables in PLS regression. As a result, PLS is capable of handling a large number of independent variables without overfitting.

The coefficients for the equation are:

Coefficient Variable

3.79938 Constant

0.17044 ALogP

0.004472 Molecular_Weight

-0.0449972 Num_H_Donors

0.00300126 Num_H_Acceptors

0.0667963 Num_RotatableBonds

0.00855581 Num_Rings

0.0377531 Num_AromaticRings

-0.0677291 Num_Fragments

0.00296441 Molecular_PolarSurfaceArea

-0.20177 Count<FCFP_6:0>

-0.0815081 Count<FCFP_6:3>-0.0563673 Count<FCFP_6:32>-0.0744673 Count<FCFP_6:1070061035>

-0.117064 Count<FCFP_6:-1272709286>-0.0619761 Count<FCFP_6:71953198>

-0.193417 Count<FCFP_6:136597326>

4.3.Descriptors in the model:

[1]ALogP unitless The calculated partition-coefficient of a compound between 1-octanol and water

[2]Molecular_Weight gram/mole The calculated molecular weight by summing the average atomic weight of all the atoms in the molecule.

[3]Num_H_Donors unitless Number of hydrogen bond donors.

[4]Num_H_Acceptors unitless Number of hydrogen bond acceptors in the molecule.

[5]Num_RotatableBonds unitless Number of rotatable bonds in the molecule.

[6]Molecular_PolarSurfaceArea Angstrom-squared The polar surface area of the molecule.

[7]Num_AromaticRings unitless Number of aromatic rings in the structure.

[8]Num_Rings Unitless The number of rings in the structure

[9]Num_Fragments Unitless The number of fragments in the structure.

[10]FCFP_6 Unitless Function class extended-connectivity fingerprint with maximum bonds length of 6

[11]ECFP_6 Unitless Extended-connectivity fingerprint with maximum bond length of 6

[12]MDLPublicKeys Unitless Fingerprint comprised of features defined in the MDL Public Keys

4.4.Descriptor selection:

A pool of most commonly used descriptors (ALogP, Molecule_Weight, Num_H_Donors, Num_H_Acceptors, Num_RotatableBonds, Num_AromaticRings, Molecular_PolarSurfaceArea, ECFP_2, ECFP_4, ECFP_6, ECFP_8, ECFP_10, ECFP_12, FCFP_2, FCFP_4, FCFP_6, FCFP_8, FCFP_10, FCFP_12, SCFP_2, SCFP_4, SCFP_6, SCFP_8, SCFP_10, SCFP_12, MDLPublicKeys) were selected randomly to build models. The model with the best 20-fold cross-validated q-squared score is selected to build the final model. The number of components (latent variables) is also set based on the cross-validated q-squared.

4.5.Algorithm and descriptor generation:

(1) The ALogP is the Ghose/Crippen group-contribution estimate for LogP, where P is the relative solubility of a compound in octanol versus water. See Ghose, A.K., Viswanadhan, V.N., and Wendoloski, J.J., "Prediction of Hydrophobic (Lipophilic) Properties of Small Organic Molecules Using Fragment Methods: An Analysis of AlogP and CLogP Methods." J. Phys. Chem. A, 1998, 102, 3762-3772.

(2) Molecular weight is calculated using the atomic weights of the individual atoms in the molecule.

(3) Hydrogen bond acceptors are defined as heteroatoms (O, N, S, or P) with one or more lone pairs, excluding atoms with positive formal charges, amide and pyrrole-type nitrogens, and aromatic oxygen and sulfur atoms in heterocyclic rings.

(4) Hydrogen bond donors are defined as heteroatoms (O, N, S, or P) with one or more attached hydrogen atoms.

(5) Molecular_PolarSurfaceArea is the polar surface area calculated using a 2D approximation to each molecule.

(6) Num_AromaticRings is the count of aromatic rings in the molecule.

(7) The fingerprint generation method is based on one of the original algorithms in computational organic chemistry called the Morgan

algorithm. The goal of the Morgan algorithm is to assign a unique identity to each atom in a molecule so that a molecule can be described in a way that is invariant to the original numbering of atoms. The algorithm has two parts: the assignment of an initial code to each atom, and an iterative part in which each atom code is updated to reflect the codes of each atom's neighbors.
ECFP_6 and FCFP_6 are calculated by first assigning atom types (ECFP_0 and FCFP_0) using atom type and functional class rule, and an n iterative process is used to generate features that represent each atom in progressively larger structural neighborhoods. After each iteration, the new feature codes for the atoms are added to the set of features from all previous steps. The process completes when the desired size is reached and the set of all features is returned as the fingerprint.
MDLPublicKeys are bitset fingerprints calculated by searching the structure using predefined queries representing the 166 MDL public keys.

4.6.Software name and version for descriptor generation:

Dassult Systemes BIOVIA Pipeline Pilot Server
Built on the BIOVIA Foundation, Pipeline Pilot enables scientists to rapidly create, test and publish scientific services that automate the process of accessing, analyzing and reporting scientific data, either for the scientist’s personal use or for sharing across the scientific community. Using Pipeline Pilot, scientist, researchers, engineers, and analysts with little or no software development experience can create scientific protocols that can be executed through a variety of interfaces including Accelrys Web Port, other Accelrys solutions such as Accelrys Electronic Lab Notebook, Isentris, Chemical Registration and third-party applications such as Microsoft SharePoint or customer-developed applications. These protocols aggregate and provide immediate access to volumes of disparate research data locked in silos. They automate the scientific analysis of the data and enable researchers to rapidly explore, visualize and report results
U.S. 6am -4pm (Pacific Time) Toll Free: 1-800-756- 4674 Tel: (858) 799-5509 support@accelrys.com U.K. 9:00 to 16:00 (UK time) Tel: +44 1223 228822 UK local rate: +44 845 741 3375 Central Europe 9:00 to 16:00 (Central European time) Switzerland: Tel: +41 61 588 0480 Germany: Tel: +49 221 8282 9020 support@accelrys.com Japan 10:00 to 17:00 (Tokyo time) Toll Free: 0120-712655 Tel: +81 3 4321 3906 support-japan@accelrys.com
http://accelrys.com/products/pipeline-pilot/

4.7.Chemicals/Descriptors ratio:

Number of chemicals = 654
Number of descriptors = 10
Chemicals/Descriptors = 65.4
Number of latent variables = 7
Number of chemicals/Number of latent variables = 93.4

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The applicability domain of the model is defined by the range of descriptors of training set chemicals. The applicability domain is only a qualititive measure on how reliable the prediction is. There is no quantative measure on how reliable the prediction is.

5.2.Method used to assess the applicability domain:

If a continuous descriptor is out of range of the training set, a warning is issued for the input compound. For the fingerprint descriptors, if a new feature not seen in the training set is found, a warning message is issued for that feature.

5.3.Software name and version for applicability domain assessment:

Dassult Systemes BIOVIA Pipeline Pilot Server

Built on the BIOVIA Foundation, Pipeline Pilot enables scientists to rapidly create, test and publish scientific services that automate the process of accessing, analyzing and reporting scientific data, either for the scientist's personal use or for sharing across the scientific community. Using Pipeline Pilot, scientist, researchers, engineers, and analysts with little or no software development experience can create scientific protocols that can be executed through a variety of interfaces including Accelrys Web Port, other Accelrys solutions such as Accelrys Electronic Lab Notebook, Isentris, Chemical Registration and third-party applications such as Microsoft SharePoint or customer-developed applications. These protocols aggregate and provide immediate access to volumes of disparate research data locked in silos. They automate the scientific analysis of the data and enable researchers to rapidly explore, visualize and report results.

U.S. 6am -4pm (Pacific Time) Toll Free: 1-800-756- 4674 Tel: (858) 799-5509 support@accelrys.com U.K. 9:00 to 16:00 (UK time) Tel: +44 1223 228822 UK local rate: +44 845 741 3375 support@accelrys.com Central Europe 9:00 to 16:00 (Central European time) Switzerland: Tel: +41 61 588 0480 Germany: Tel: +49 221 8282 9020 support@accelrys.com Japan 10:00 to 17:00 (Tokyo time) Toll Free: 0120-712655 Tel: +81 3 4321 3906 support-japan@accelrys.com http://accelrys.com/products/pipeline-pilot/

5.4.Limits of applicability:

Variable	Min	Max	Mean	Std. Dev.
ALogP	-6.264	10.795	2.4862	2.1756
Molecular_Weight	30.026	777.96	236.08	114.41
Num_H_Donors	0	11	0.79817	1.0972
Num_H_Acceptors	0	17	2.9862	2.0893
Num_RotatableBonds	0	24	3.5275	3.618
Num_Rings	0	7	1.2324	1.1538
Num_AromaticRings	0	4	0.90826	0.85536
Num_Fragments	1	5	1.1131	0.40567
Molecular_PolarSurfaceArea	0	299.12	58.748	42.163

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: No

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:

The data used to train the model consisted of 654 samples. The training set is proprietary, however, it is emmbedded with the model and can be retrieved with similarity search when a prediction is conducted.

6.6.Pre-processing of data before modelling:

N/A

6.7.Statistics for goodness-of-fit:

r = 0.766 r-squared = 0.587 r-squared (adjusted) = 0.583 RMS error = 1.142

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

N/A

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

20-fold cross-validation:

q-squared = 0.471

RMS error = 1.299

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:

No

7.2.Available information for the external validation set:

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:

Due to the small size of the available data, no data were reserved for external validation purpose.

7.6.Experimental design of test set:

N/A

7.7.Predictivity - Statistics obtained by external validation:

N/A

7.8.Predictivity - Assessment of the external validation set:

N/A

7.9.Comments on the external validation of the model:

N/A

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The following variables have the highest coefficient in the equation
(Each of these features coresponding to a substructure in a molecule.):
Count<ECFP_6:-826638028> 0.367626
Count<FCFP_6:1281467142> 0.279461
Count<ECFP_6:1980176114> 0.262749
Count<ECFP_6:-845108448> 0.260843
Count<FCFP_6:-1143715940> 0.217641
Count<MDLPublicKeys:141> 0.201803
Count<MDLPublicKeys:76> 0.199531
Count<MDLPublicKeys:106> 0.187114
ALogP 0.17044
Count<MDLPublicKeys:48> 0.165845
Count<ECFP_6:-1059365320> 0.165075
Count<MDLPublicKeys:138> 0.155421
Count<ECFP_6:642810091> 0.148475Count<MDLPublicKeys:23> 0.132375
Count<MDLPublicKeys:129> 0.128162
Count<MDLPublicKeys:128> 0.123952Count<FCFP_6:-1986098826> 0.11896
Count<ECFP_6:1572579716> 0.114438
Count<FCFP_6:565998553> 0.113583Count<MDLPublicKeys:116> 0.109823

8.2.A priori or a posteriori mechanistic interpretation:

posteriori: these features are selected purely based on their
coefficient appearing in the final equation

8.3.Other information about the mechanistic interpretation:

N/A

9.Miscellaneous information

9.1.Comments:

The model is extensible, i.e., it can be extended by feeding new
training data to create an improved model.

9.2.Bibliography:

Wold S, Ruhe A, Wold H, Dunn WJ (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses". SIAM Journal on Scientific and Statistical Computing. 5 (3) 735–743 http://dx.doi.org/10.1137%2F0905052

9.3.Supporting information:

qmrf514_qmrf457_Daphnia-EC50-training-set 654.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-31-0047/attachment/A1094
qmrf514_qmrf457_DaphniaEC50-equation.txt	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-31-0047/attachment/A1095

Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-31-0047

10.2.Publication date:

2017-09-27

10.3.Keywords:

Daphnia magna;acute daphnid toxicity;EC50;BIOVIA Discovery Studio;

10.4.Comments:

old# Q51-54-55-514