	QMRF identifier (JRC Inventory): **Q17-412-0008**
	QMRF Title: **BIOVIA toxicity prediction model – mouse carcinogenic potency TD50**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

BIOVIA toxicity prediction model – mouse carcinogenic potency TD50

1.2.Other related models:

BIOVIA toxicity prediction model - rat carcinogenic potency TD50

1.3.Software coding the model:

BIOVIA Discovery Studio v4.5

Dassault Systèmes, BIOVIA Corp., 5005 Wateridge Vista Drive, San Diego, CA92121, USA

http://www.3dsbiovia.com

2.General information

2.1.Date of QMRF:

16 January 2017

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 and data are proprietary (available as a commercial product), but the algorithm is public. The training set is also proprietary, however, it is 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:

Mouse (Mus musculus)

3.2.Endpoint:

4.Human Health Effects 4.12.Carcinogenicity

3.3.Comment on endpoint:

TD50, a numerical description of carcinogenic potency, is the daily dose-rate in mg/kg/body weight/day for life to induce tumours in half of test animals that would have remained tumour-free at zero dose. TD50 provides a standardized quantitative measure that can be used for comparisons and analyses of many issues in carcinogenesis.

3.4.Endpoint units:

mg/kg_body_weight/day

3.5.Dependent variable:

pTD50 = -log(TD50/molecular_weight/1000)

3.6.Experimental protocol:

All data were obtained from the Carcinogenic Potency Database and the experimental protocols were documented online at https://toxnet.nlm.nih.gov/cpdb/methods.html#sources .

3.7.Endpoint data quality and variability:

The data quality and variability are documented online at https://toxnet.nlm.nih.gov/cpdb/methods.html#sources . When there were multiple target sites, the TD50 were for all target sites.

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. As in multiple linear regression, the main purpose of PLS regression is to build a linear model: $Y = X \times B + E$ where Y is a response matrix (or vector) formed by the dependent variables, X is a matrix formed by the independent variables, B is a matrix of the regression coefficients, and E is an error term for the model. Usually, the variables in X and Y are centered by subtracting their means and scaled by dividing by their standard deviations. In PLS regression, a procedure called factor extraction is applied to produce a new matrix: $T = X \times W$ where T and W are called the factor score matrix and the weight matrix, respectively. A new linear regression model is represented as: $Y = T \times Q + E$, where Q is a matrix of regression coefficients (called loadings) for T, and E is an error (noise) term. Once the loadings Q are computed, the above regression model is equivalent to the predictive regression model $Y = X * B + E$, where $B = W * Q$. In a principal component analysis, a set of principal components can be obtained by diagonalizing the covariance matrix of the independent predictor variables. This is done similarly in PLS regression, with the exception that the covariance matrix includes both the predictor and response variables. For establishing the model, PLS regression produces a weight matrix W for X such that $T = X \times W$, i.e., the columns of W are weight vectors for the X columns producing the corresponding factor score matrix T. These weights are computed so that each of them maximizes the covariance between responses and the corresponding factor scores. Ordinary least squares procedures for the regression of Y on T are then performed to produce Q, the loadings for Y (or weights for Y). One additional matrix which is necessary for a complete description of PLS regression procedures is the factor loading matrix P

which gives a factor model X = T × P + F, where F is the unexplained part of the X scores. 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 equation contains 6 latent variables, and each is a linear combination of a series of variables. The following table contains the coefficients and associated variables for the equation.
Coefficient Variable
2.98593 Constant
0.098191 ALogP
0.00202718 Molecular_Weight
0.170784 Num_H_Donors
0.0334661 Num_H_Acceptors
-0.0657117 Num_RotatableBonds
0.291522 Num_Rings
0.0193599 Num_AromaticRings
0.0478265 Num_Fragments
-0.00170157 Molecular_PolarSASA
-0.231649 Count<ECFP_6:-182236392>
-0.246856 Count<ECFP_6:642810091>0.229187 Count<ECFP_6:655739385>0.225296 Count<ECFP_6:1572579716>
-0.198998 Count<ECFP_6:1997021792>-0.25146 Count<ECFP_6:1996767644>
0.0745686 Count<ECFP_6:1333660716>
0.162565 Count<ECFP_6:834876373>-0.120884 Count<ECFP_6:-938530932>
-0.170326 Count<ECFP_6:1564392544>0.0423858 Count<ECFP_6:734603939>0.144638 Count<ECFP_6:-1925046727>0.103846 Count<ECFP_6:-1087070950>
0.428463 Count<ECFP_6:-1072294614>
-0.0401674 Count<ECFP_6:-1074141656>0.218812 Count<ECFP_6:865379614>
0.183428 Count<ECFP_6:2101483135>
-0.00735101 Count<ECFP_6:-1100000244>-0.197057 Count<ECFP_6:866218936>-0.117627 Count<ECFP_6:2099970318>0.0784231 Count<ECFP_6:-932108170>0.0284479 Count<ECFP_6:-1897341097>-0.21684 Count<ECFP_6:-1884411803>-0.257865 Count<ECFP_6:2019062761>0.203334 Count<ECFP_6:1559650422>0.17239 Count<ECFP_6:-2024255407>0.0447619 Count<ECFP_6:2022454958>0.106622 Count<ECFP_6:-175146122>0.0144884 Count<ECFP_6:1571214559>-0.0851945 Count<ECFP_6:-992506539>-0.216887 Count<ECFP_6:-786013480>-0.100437 Count<ECFP_6:-427397688>-0.175683 Count<ECFP_6:-1085223908>0.129075 Count<ECFP_6:-817402818>-0.183822 Count<ECFP_6:-1059365320>0.0460383 Count<ECFP_6:1043790491>0.0797092 Count<ECFP_6:781519895>-0.0229216 Count<ECFP_6:1408898974>0.00431772 Count<ECFP_6:99947387>0.081844 Count<ECFP_6:-176455838>-0.0756996 Count<ECFP_6:-215026467>-0.0175644 Count<ECFP_6:2025485523>0.00135389 Count<ECFP_6:1544874086>-0.00352899 Count<ECFP_6:-2090955291>0.135537 Count<ECFP_6:683445015>0.0596221 Count<ECFP_6:-167460056>-0.0215548 Count<ECFP_6:864909220>0.0155955 Count<ECFP_6:1307307440>-0.274538 Count<ECFP_6:2106656448>0.0789364 Count<ECFP_6:657586427>0.0279354 Count<ECFP_6:-1910270391>0.0365618 Count<ECFP_6:864518973>0.00514838 Count<ECFP_6:-1642449301>0.108174 Count<ECFP_6:-934039951>-0.0155307 Count<ECFP_6:863188371>-0.239283 Count<ECFP_6:971566354>-0.0368874 Count<ECFP_6:971820502>-0.178113

Count<ECFP_6:859796174>0.0636558 Count<ECFP_6:2147419938>0.0725437 Count<ECFP_6:-181568884>0.138483 Count<ECFP_6:103339584>0.0073467 Count<ECFP_6:670515721>

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_PolarSASA 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 molecule

[9]Num_Fragments unitless Number of fragments in the molecule.

[10]ECFP_6 unitless Extended-connectivity functional class fingerprint with a maximum length of 6 bonds

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_FractionPolarSurfaceArea is calculated from the polar surface area and total surface area using a 2D approximation to each molecule.

(6) 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 neighbours.

(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_PolarSASA the solvent accessible polar surface area calculated using a 2D approximation to each molecule.

(6) Num_AromaticRings and Num_Rings arer the count of aromatic and total number of rings in the molecule, respectively.(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 neighbours.ECFP_6 is calculated by first assigning atom types (ECFP_0) using generic class atom type 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.

4.6.Software name and version for descriptor generation:

Dassult Systemes BIOVIA Pipeline Pilot Server

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 = 530
Number of descriptors = 10
Chemicals/Descriptors = 53
Number of latent variables = 7
Number of chemicals/Number of latent variables = 76

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 qualitative measure on how reliable the prediction is. There is no quantitive 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
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.
pTD50 0.45394 9.3147 3.6463 1.2095
ALogP -5.006 14.213 1.8809 2.2732
Molecular_Weight 30.026 871.78 225.83 133.75
Num_H_Donors 0 8 1.1962 1.3043
Num_H_Acceptors 0 22 2.8019 2.3721
Num_RotatableBonds 0 26 2.6321 3.3368
Num_Rings 0 8 1.4547 1.4071
Num_AromaticRings 0 6 0.96038 1.002
Num_Fragments 1 9 1.183 0.61621
Molecular_PolarSASA 0 707.69 95.873 73.652
ECFP_6 N/A N/A N/A N/A

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: Yes

Formula: No

INChI: No

MOL file: No

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 530 samples. The training set is proprietary, however, it is embedded with the model and can be retrieved with similarity search when a prediction is conducted.

6.6.Pre-processing of data before modelling:

This model was trained using 530 uniform experimental Carcinogenic Potency TD50 values downloaded from the Carcinogenic Potency Database (CPDB), available online from https://toxnet.nlm.nih.gov/cpdb/ . Each value was first converted to g/kg body weight/h, and then calculated as -log(TD50/Molecular_Weight).

6.7.Statistics for goodness-of-fit:

r = 0.597

r-squared = 0.357

r-squared (adjusted) = 0.350

RMS error = 0.97

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 = 0.255

RMS error = 1.05

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: 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:

No

7.5.Other information about the external validation set:

N/A

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:

No mechanistic basis of the model was attempted. However, the contribution of each fingerprint feature can imply their importance to the overall outcome.

8.2.A priori or a posteriori mechanistic interpretation:

a posteriori: these features are selected purely based on their coefficients 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:

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

10.Summary (JRC Inventory)

10.1.QMRF number:

Q17-412-0008

10.2.Publication date:

2017-09-20

10.3.Keywords:

mouse;carcinogenicity;TD50;BIOVIA;

10.4.Comments: