	QMRF identifier (JRC Inventory): **Q17-412-0049**
	QMRF Title: **BIOVIA toxicity prediction model – weight of evidence rodent carcinogenicity**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

BIOVIA toxicity prediction model – weight of evidence rodent carcinogenicity

1.2.Other related models:

Toxicity Prediction (Extensible) NTP Carcinogenicity Call (Male Rat)

Toxicity Prediction (Extensible) NTP Carcinogenicity Call (Female Rat)

Toxicity Prediction (Extensible) NTP Carcinogenicity Call (Male Mouse)

Toxicity Prediction (Extensible) NTP Carcinogenicity Call (Female Mouse)

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:

5/5/2015

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 commericcial product), but the algorithm is publically available. 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:

Rat and mouse

3.2.Endpoint:

4.Human Health Effects 4.12.Carcinogenicity

3.3.Comment on endpoint:

In a rodent carcinogenicity test using the US Food and Drug Administration (FDA) Center for Drug Evaluation and Research (CDER) weight-of-evidence protocol, a chemical is scored as a carcinogen if (1) It is a multiple-site carcinogen in at least one sex/species combination (male or famale/rate or mouse).
Or (2) it is a single-site carcinogen in at least two sex/species combinations.

3.4.Endpoint units:

Dimensionless - Yes/No Binary Classification

3.5.Dependent variable:

Classification as carcinogenic or non-carcinogenic

3.6.Experimental protocol:

All the data were collected from literatures reporting NTP 2-year caicinogenicity experiments.
According to
http://ntp.niehs.nih.gov/testing/types/cartox/protocols/2year/index.html
2-Year Study Protocol
The purpose of this study is to determine the toxicologic and/or carcinogenic effects of long-term exposure on rats and mice.

Treatment:

After a 10- to 14-day quarantine period, animals are assigned at random to treatment groups. Rats and mice receive the test agent for 104 weeks via a defined route of exposure at 3 treatment concentrations plus controls. Animals have continuous ad libitum access to dosed-feed and dosed-water in those exposure studies. For inhalation, gavage and dermal studies, animals are treated five times per week, weekdays only. Male mice are housed individually and rats and female mice are group-housed, except for inhalation and dermal exposure studies in which all rats and mice are housed individually.

Animals Sexes Species Test
Groups Total
Treatment 50 x 2 x 2 x 3 = 600
Controls 50 x 2 x 2 x 1 = 200Sentinel Animals 15 x 2 x 2x - = 60

___

Total
860Observations:
Individual animal body weights for test and control group animals are recorded on day one on test and at 4-week intervals thereafter except for dosed-feed and dosed-water studies, which are recorded weekly for the first thirteen weeks and monthly thereafter. If life-threatening tumors develop, a significant number of deaths occur, or a significant effect on body weight is observed, the weighing frequency may be increased to every two weeks.Animals are observed twice daily at least six hours apart (before 10:00
am and after 2:00 pm including holidays and weekends) for moribundity and mortality. Animals found moribund or showing clinical signs of pain or distress are humanely euthanized. Formal examinations for clinical signs of toxicity are made and recorded at four-week intervals. For dosed-feed or dosed-water studies, food consumption/water consumption is measured and recorded weekly for the first thirteen weeks and monthly thereafter.

Necropsy and Pathology:
Necropsy: A complete necropsy is performed on all treated and control animals that either die or are sacrificed. All tissues required for complete histopathology are trimmed, embedded, sectioned and stained with hematoxylin and eosin for histopathologic evaluation. (necropsy list)
Histopathology: All animals in all treatment groups that die (or are sacrificed in a moribund condition) and those that complete the 104-week exposure are subjected to a complete necropsy and slides of all tissues required for complete histopathologic evaluation are prepared and evaluated. (histopathology list)

3.7.Endpoint data quality and variability:

The two year study is very expensive and there are not enough data to give data variability information.

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

QSAR model derived from Bayesian binary classification

4.2.Explicit algorithm:

Bayesian Classification
A modified Bayesian learning method is used. The algorithm is described in Xia X, Maliski EG, Gallant P & Rogers D(2004). Journal of Medicinal Chemistry. 47(18) 4463- 4470
Pcorr(Active|F) = (A + P(Active)*K)/(B + K).

(For K = 1/P(Active), this is the Laplacian correction.)

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_FractionalPolarSurfaceArea unitless The fraction of polar surface area over the total molecular surface area.

[7]SCFP_8 unitless Extended-connectivity SYBYL atom type fingerprint with a maximum length of 8 bonds

4.4.Descriptor selection:

A pool of most commonly used descriptors (ALogP, Molecule_Weight, Num_H_Donors, Num_H_Acceptors, Molecular_FractionPolarSurfaceArea, 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) were selected randomly to build models. The model with the best leave-one-out cross-validated ROC score is selected to build the final model. In addition, Bayesian model has a built-in mechanism to select the most statitstically-significant descriptors.

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 neighbors.

SCFP_8 is calculated by first assigning atom types (SCFP_0) using SYBYL atom types, 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

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 = 333

Number of chemicals = 470

Number of descriptors = 7

Chemicals/Descriptors = 67.1

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:

Property Min Max Mean Std. Dev.
ALogP -7.685 11.179 1.8335 2.2979
Property Min Max Mean Std. Dev.
ALogP -7.685 10.955 2.2018 2.174
Molecular_Weight 46.068 760.62 271.86 116.01
Num_H_Donors 0 8 1.3872 1.3301
Num_H_Acceptors 0 13 3.6596 2.3254
Num_RotatableBonds 0 16 3.8468 3.4693
Molecular_FractionalPolarSurfaceArea 0 0.969 0.2527 0.16587

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: 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 470 samples. 242 of them are in the positive category. 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:

None

6.7.Statistics for goodness-of-fit:

N/A

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

ROC score=0.794 (LOO)

True Positive = 156

False Negative = 86

False Positive = 29

True Negative = 199

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

ROC score = 0.704 (Leave 10% out)

Sensitivity = 0.913

Specificity = 0.969

Concordance = 0.940

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:

Features contributing the most from SCFP_8 are included in attachment.

8.2.A priori or a posteriori mechanistic interpretation:

posteriori: these features are selected purely based on their Bayesian score

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:

[1]NTP Technical Reports http://ntp.niehs.nih.gov/results/pubs/longterm/reports/longterm/index.html

[2]tHE Carcinogenic Potency Database (CPDB) http://toxnet.nlm.nih.gov/cpdb/

[3]Xia X, Maliski EG, Gallant P & Rogers D(2004). Journal of Medicinal Chemistry. 47(18) 4463-4470 http://pubs.acs.org/doi/full/10.1021/jm0303195

9.3.Supporting information:

qmrf507_qmrf449_woe 470.sdf	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-412-0049/attachment/A1091
qmrf507_qmrf449_WOE-features.png	http://qsardb.jrc.ec.europa.eu/qmrf/protocol/Q17-412-0049/attachment/A1092

Test set(s)Supporting information

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-412-0049

10.2.Publication date:

2017-09-27

10.3.Keywords:

rodent;male;female;carcinogenicity;weight of evidence;BIOVIA Discovery Studio;

10.4.Comments:

old# Q50-54-55-507