	QMRF identifier (JRC Inventory): **Q17-46-0042**
	QMRF Title: **BIOVIA toxicity prediction model – skin sensitiser vs non sensitiser**
	Printing Date: **Apr 16, 2018**

1.QSAR identifier

1.1.QSAR identifier (title):

BIOVIA toxicity prediction model – skin sensitiser vs non sensitiser

1.2.Other related models:

Toxicity Prediction (Extensible) Skin Sensitization (Weak vs Strong Sensitizer)

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:

12/5/2015

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 is not. 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:

Guinea Pig

3.2.Endpoint:

4.Human Health Effects 4.6.Skin sensitisation

3.3.Comment on endpoint:

The Guinea pig maximisation test (GPMT) is an in vivo test to screen for substances the cause human skin sensitisation (i.e. allergens). It was first proposed by B. Magnusson and Albert Kligman in 1969 and described in their 1970 book Allergic Contact Dermatitis in the Guinea Pig.

3.4.Endpoint units:

Dimensionless - Yes/No Binary Classification

3.5.Dependent variable:

Classification as sensitizer or non-sensitizer (any category showing positive reactions will be classified as sensitizer).

3.6.Experimental protocol:

The OECD Guidelines for the Testing of Chemicals guideline No. 406 of 1992 contains the GPMT protocol, available online at http://www.oecd-ilibrary.org/environment/test-no-406-skin-sensitisation_9789264070660-en

This Test Guideline is intended primarily for use with guinea pig, but recently mouse models for assessing sensitisation potential have been developed. For the GPMT at least 10 animals in the treatment group and 5 in the control group are used. For the Buehler test, a minimum of 20 animals is used in the treatment group and at least 10 animals in the control group. The test animals are initially exposed to the test substance. Following a rest period, the induction period (10-14 days), during which an immune response may develop, then the animals are exposed to a challenge dose. The GPMT is made during approximately 23-25 days, the Buehler test, during approximately 30-32 days. The concentration of test substance used for each induction exposure should be well-tolerated systemically and should be the highest to cause mild-to moderate skin irritation, for the challenge exposure the highest nonirritant dose should be used. All skin reactions and any unusual findings should be observed and recorded (other procedures may be carried out to clarify doubtful reactions).

3.7.Endpoint data quality and variability:

The data for this model, all based on the guinea-pig maximization test (GPMT) or minor modification of it, were assembled from:

(1) A collection of published skin sensitization assays on and before 1985.
(2) Later data published in Contact Dermatitis.
(3) Cronin, M.T.D. and Basketter, D.A., (1994) Multivariate QSAR analysis of a skin sensitization data base, SAR and QSAR in Environmental Research, 2, 159-179.
(4) Augmented data from the 1999 RTEC CD for a total of 1300 chemicals.

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]FCFP_12 unitless Functional class extended-connectivity atom type fingerprint with a maximum length of 12 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.
FCFP_12 is calculated by first assigning a functional class (Aromatic, HBD, HBA, Negatively charged, Positively charged, etc) to each atom (FCFP_0), 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 = 392
Number of descriptors = 7
Chemicals/Descriptors = 56.0

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 the range of the descriptors in the training set. 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:

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.

Property	Min	Max	Mean	Std. Dev.
ALogP	-4.478	15.512	2.3576	2.3927
Molecular_Weight	30.026	959.12	217.67	105.4
Num_H_Donors	0	12	1.0255	1.2553
Num_H_Acceptors	0	19	3.2194	2.1099
Num_RotatableBonds	0	36	4.574	5.1329
Molecular_FractionalPolarSurfaceArea	0	0.759	0.25216	0.14459
FCFP_12	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: 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 392 samples. 269 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:

False Negative = 68

False Positive = 37

True Negative = 86

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

ROC score = 0.773 (Leave 10% out)

Sensitivity = 0.911

Specificity = 0.886

Concordance = 0.903

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 FCFP_12 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]Magnuson B & Kligman AM (1969). The identification of contact allergens by animal assay: The guinea pig maximization test. Journal of Investigative Dermatology. 52, 268-276. doi:10.1038/jid.1969.42.

[2]Barratt et al (1994). An expert system rulebase for identifying contact allergens. Toxicology in vitro. 8(5) 1053-1060 http://www.sciencedirect.com/science/article/pii/0887233394902445

[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:

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q17-46-0042

10.2.Publication date:

2017-09-27

10.3.Keywords:

guinea pig maximisation test;GPMT;skin sensitisation;BIOVIA Discovery Studio;

10.4.Comments:

old# Q50-54-55-509