QMRF identifier (JRC Inventory): **Q17-410-0037**
QMRF Title: **BIOVIA toxicity prediction model – Ames Mutagenicity**
Printing Date: **Apr 16, 2018**

1.QSAR identifier

1.1.QSAR identifier (title):

BIOVIA toxicity prediction model – Ames Mutagenicity

1.2.Other related models:

None

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:

17/3/2015

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

2.8.Availability of information about the model:

The model is proprietary (available as a commericcial product), but the training set data and algorithm are all available. The training set is also embedded with the model and can be retrieved with similarity search when a prediction is conducted.

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

None

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Salmonella typhimurium (Ames Mutagenicity Test)

3.2.Endpoint:

4.Human Health Effects 4.10.Mutagenicity

3.3.Comment on endpoint:

The Ames test is a widely employed method that uses bacteria to test whether a given chemical can cause mutations in the DNA of the test organism. More formally, it is a biological assay to assess the mutagenic potential of chemical compounds. The data modelled are +/- mutagenicity outcomes, based on the results of the entire set of individual Salmonella strains.

3.4.Endpoint units:

unitless

3.5.Dependent variable:

Classification as mutagenic or nonmutagenic

3.6.Experimental protocol:

According to
http://ntp.niehs.nih.gov/testing/types/genetic/invitro/sa/index.html

In the standard protocol (preincubation) for conducting the Ames assay, a test tube containing a suspension of one strain of Salmonella typhimurium (or E. coli) plus S9 mix or plain buffer without S9, is incubated for 20 minutes at 37º C with the test chemical. Control cultures, with all the same ingredients except the test chemical, are also incubated. In addition, positive control cultures are prepared; these contain the particular bacterial tester strain under investigation, the various culture ingredients, and a known potent mutagen*. After 20 minutes, agar is added to the cultures and the contents of the tubes are thoroughly mixed and poured onto the surface of Petri dishes containing standard bacterial culture medium. The plates are incubated, and bacterial colonies that do not require an excess of supplemental histidine or tryptophan appear and grow. These colonies are comprised of bacteria that have undergone reverse mutation to restore function of the histidine- or tryptophan- manufacturing gene. The number of colonies is usually counted after 2 days.

Several modifications of the Ames test protocol have been used over the years in special circumstances. These include standard plate incorporation (no preincubation step prior to plating onto Petri dishes), FMN reduction (use of flavin mononucleotide for reduction of test articles such as azo dyes), plate test with volatile liquids (exposure of bacteria in a sealed Petri dish), cecal reduction (use of rat cecal bacteria to provide reduction of azo compounds), and plate tests conducted within a sealed dessicator (gas chamber) for exposure to gaseous substances. The specific test protocol that was used in an Ames test is noted in the description of the assay data.

Spontaneous mutations (those that occur by chance, not by chemical treatment) will appear as colonies on the control petri dishes. If the test chemical was mutagenic to any particular strain of bacterium, the number of histidine-independent colonies arising on those plates will be significantly greater than the corresponding control plates for that strain of bacteria. The positive control plates are also counted, and the number of mutant colonies appearing on them must be significantly increased over the spontaneous control number for the test to be considered valid. Failure of the positive control chemical to induce mutation is reason to discard the experiment.

Several doses (usually at least 5) of each test chemical and multiple strains of bacteria are used in each experiment. In addition, cultures are set up with and without added liver S9 enzymes at varying concentrations. Therefore, a variety of culture conditions are employed to maximize the opportunity to detect a mutagenic chemical. In analyzing the data, the pattern and the strength of the mutant response are taken into account in determining the mutagenicity of a chemical. All observed responses are verified in repeat tests. If no increase in mutant colonies is seen after testing several strains under several different culture conditions, the test chemical is considered to be nonmutagenic in the Ames test.
This model was trained using 6313 samples from six different mutagenicity datasets:

TOPKAT version 6.2, BIOVIA, San Diego, CA 92121;
Kazius, McGuire, and Bursi, J. Med. Chem., 48, 2005, pp. 312-320;
Contrera, J.F., Matthews, E.J., Kruhlak, N.L., and Benz, R.D., Regulatory Toxicology and Pharmacology 2005, pp. 313-323;Feng, J., Lurati, L., Ouyang, H., Robinson, T., Wang, Y., Yuan, S., and Young, S.S., J. Chem. Inf. Comput. Sci., 2003, pp. 1463-1470;Helma, C., Cramer, T., Kramer, S., and De Raedt, L., J. Chem. Inf. Comput. Sci., 2004, pp. 1402-1411;
ISSCAN dataset, Istituto Superiore di Sanita, available online at http://www.epa.gov/NCCT/dsstox/sdf_isscan_external.html.Duplicates were removed across the datasets. The compounds were assayed according to the US EPA GeneTox protocol. According to the protocol, a chemical is tested against five strains of Salmonella typhimurium, namely: TA100, TA1535, TA1537, TA 1538 and TA 98, using the Histidine Reversion Assay. Tests are performed both with and without S9 activation. A chemical is labeled a mutagen if a positive response, i.e., a significant increase in number of reversions as compared to the background reversion, is observed against one ot more strains, with or without S9 activation.

3.7.Endpoint data quality and variability:

N/A

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_12 unitless SYBYL atom type extended-connectivity 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.

SCFP_12 is calculated by first assigning atom types (SCFP_0) using SYBYL 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

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 = 6313

Number of descriptors = 7

Chemicals/Descriptors = 901.9

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 -19.632 16.879 2.2296 2.2208
Molecular_Weight 0 3080.4 249.11 150.32
Num_H_Donors 0 53 1.1754 1.8612
Num_H_Acceptors 0 62 3.4697 3.1507
Num_RotatableBonds 0 75 3.0735 3.9859
Molecular_FractionalPolarSurfaceArea 0 1 0.2669 0.17911

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 6313 samples. 3525 of them are in the positive category. The training set is attached with the QMRF and is also emmbedded with the model and can be retrieved with similarity search when a prediction is conducted.

6.6.Pre-processing of data before modelling:

Duplicates were removed across the datasets. The compounds were assayed according to the US EPA GeneTox protocol. According to the protocol, a chemical is tested against five strains of Salmonella typhimurium, namely: TA100, TA1535, TA1537, TA 1538 and TA 98, using the Histidine Reversion Assay. Tests are performed both with and without S9 activation. A chemical is labeled a mutagen if a positive response, i.e., a significant increase in number of reversions as compared to the background reversion, is observed against one ot more strains, with or without S9 activation.

6.7.Statistics for goodness-of-fit:

N/A

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

ROC score=0.894 (LOO)

True Positive = 2779

False Negative = 746

False Positive = 403

True Negative = 2385

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

ROC score = 0.890 (Leave 10% out)

Sensitivity = 0.884

Specificity = 0.909

Concordance = 0.895

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:

Yes

7.2.Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

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:

They are collected from open sources with the duplicates removed from the training set

7.6.Experimental design of test set:

N/A

7.7.Predictivity - Statistics obtained by external validation:

True Positive = 913

False Positive = 338

True Negative = 947

False Negative = 346

ROC Score = 0.808

7.8.Predictivity - Assessment of the external validation set:

The external test set contains 2544 compounds and contains some 27 compunds that are out of the applicability domain.

7.9.Comments on the external validation of the model:

Part of the external test set may be incorporated into the training set in the future.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The models gives the top 20 SCFP_12 features (with depiction of the sctructure) contribute to mutagenicity:

G1: -1376320547, 128 out of 128 mutagenic, Bayesian Score: 0.554

G2: -525770763, 107 out of 107 mutagenic, Bayesian Score: 0.553

G3: 1683345490, 101 out of 101 mutagenic, Bayesian Score: 0.552

G4: 1551989412, 97 out of 97 mutagenic, Bayesian Score: 0.552

G5: 316042202, 66 out of 66 mutagenic, Bayesian Score: 0.549

G6: -1165548501, 53 out of 53 mutagenic, Bayesian Score: 0.546

G7: 28854523, 53 out of 53 mutagenic, Bayesian Score: 0.546

G8: 354038274, 53 out of 53 mutagenic, Bayesian Score: 0.546

G9: 110936816, 52 out of 52 mutagenic, Bayesian Score: 0.546

G10: 664701054, 52 out of 52 mutagenic, Bayesian Score: 0.546

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]Kazius J, McGuire R & Bursi R(2005). Derivation and Validation of Toxicophores for Mutagenicity Prediction. Journal of Medicinal Chemistry.48 (1) 312-320
http://pubs.acs.org/doi/full/10.1021/jm040835a

[2]Contrera JF, Matthews EJ, Kruhlak NL & Benz RD (2005). In silico screening of chemicals for bacterial mutagenicity using electrotopological E-state indices and MDL QSAR software. Regulatory Toxicology and Pharmacology. 43(3) 313-323
http://www.sciencedirect.com/science/article/pii/S0273230005001686

[3]Feng J, Lurati L, Ouyang H, Robinson T, Wang Y, Yuan S & Young SS (2003). Predictive Toxicology: Benchmarking Molecular Descriptors and Statistical Methods. Journal of Chemical Information and Computer Science.43 (5) 1463–1470 http://pubs.acs.org/doi/full/10.1021/ci034032s

[4]Helma C, Cramer T, Kramer S & De Raedt L(2004). Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. Journal of Chemical Information and Computer Science.44(4) 1402-1411 http://pubs.acs.org/doi/full/10.1021/ci034254q

[5]ISSCAN dataset, Istituto Superiore di Sanita
http://www.epa.gov/NCCT/dsstox/sdf_isscan_external.html.

[6]Hansen K, Mika S, Schroeter T, Sutter A, ter Laak A, Steger-Hartmann T, Heinrich N & Müller KR(2009). Benchmark Data Set for in Silico Prediction of Ames Mutagenicity. Journal of Chemical Information & Modeling. 49 (9) 2077–2081 http://pubs.acs.org/doi/full/10.1021/ci900161g

[7]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:

Q17-410-0037

10.2.Publication date:

2017-09-27

10.3.Keywords:

Salmonella typhimurium;Ames;mutagenicity;BIOVIA Discovery Studio;

10.4.Comments:

old# Q50-54-55-501