	QMRF identifier (JRC Inventory): **Q15-410-0008**
	QMRF Title: **Caesar hybrid model for bacterial reverse mutation (Ames test)**
	Printing Date: **Dec 11, 2019**

1.QSAR identifier

1.1.QSAR identifier (title):

Caesar hybrid model for bacterial reverse mutation (Ames test)

1.2.Other related models:

Two models have been created and validated using a large set of molecular structures accompanied by the respective mutagenic toxicity experimental test results on Salmonella test. Model A is based on data mining with support vector machines (SVM) and Model B is based on expert knowledge coded as structural alerts (SA).

The final model C combines models A and B to achieve a better predictive performance.

1.3.Software coding the model:

CAESAR Mutagenicity Model 2.0.

This is the standalone version of the CAESAR Mutagenicity Model 1.0, which implements the Mutagenicity endpoint. The Applicability Domain tool is the main improvement compared to the previous version.

coord@caesar-project.eu

http://www.caesar-project.eu/software/

CAESAR Mutagenicity Model 1.0.

The CAESAR Application is a JAVA web application that provides access to all of the toxicity predictive models developed within the CAESAR Project.

coord@caesar-project.eu

http://www.caesar-project.eu/software/

2.General information

2.1.Date of QMRF:

20/11/2014

2.2.QMRF author(s) and contact details:

Emilio Benfenati Istituto di Ricerche Farmacologiche "Mario Negri" emilio.benfenati@marionegri.it

2.3.Date of QMRF update(s):

2.4.QMRF update(s):

2.5.Model developer(s) and contact details:

[1]Thomas Ferrari Department of Electronics and Information (DEI), Politecnico di Milano

[2]Alberto Manganaro Istituto di Ricerche Farmacologiche "Mario Negri"

2.6.Date of model development and/or publication:

The model was published in 2010 (see 2.7).

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

Ferrari T, Gini G (2010) An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. Chemistry Central Journal , 4(Suppl 1):S2

http://www.journal.chemistrycentral.com/content/4/S1/S2

2.8.Availability of information about the model:

The software is and is freely available through the portal of the CAESAR project. The training and test sets are available, see 9.3 Supporting information.

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

3.Defining the endpoint - OECD Principle 1

3.1.Species:

Salmonella typhimurium (Ames test)

3.2.Endpoint:

4.Human Health Effects 4.10.Mutagenicity

3.3.Comment on endpoint:

Mutagenic toxicity is the capacity of a substance to cause genetic mutations. The Ames test is the basic in vitro assay to detect mutagens.

3.4.Endpoint units:

adimensional

3.5.Dependent variable:

classification as: mutagenic / non mutagenic

3.6.Experimental protocol:

Ames test: an in vitro model of chemical mutagenicity and carcinogenicity, and consists of a range of bacterial strains that together are sensitive to a large array of DNA-damaging agents.

3.7.Endpoint data quality and variability:

For the development and the validation of the model, the Bursi Mutagenicity Dataset was used [ref.4, sect.9.2].The estimated inter-laboratory reproducibility rate of Salmonella test data is 85% [ref.2, sect.9.2].

4.Defining the algorithm - OECD Principle 2

4.1.Type of model:

An integrated model, Model C, was arranged by cascading the two models: Model A, a trained SVM classifier with an additional Model B for false negatives (FNs) correction based on SAs. The SVM classifier is the one described in the section 4.2 of the paper proposing the final model (see 2.7), while the rulebase for the expert filter was extracted from the Benigni/Bossa SAs [ref.4; sect.9.2] set.

4.2.Explicit algorithm:

Data mining with SVM coupled with knowledge based SAs for the correctionof FNs.

The model consists of a complex architecture based on support vector machines model revised by structural alerts. First, the SVM identifies mutagens. The predicted non-mutagens are then processed with the second model, Model B, based on two sets of structural alerts. If an alert of the first set is found (see 4.3 descriptors from #26 to #37), the chemical is labelled "mutagen"; if an alert of the second set is found (see 4.3 descriptors from #38 to #41), the chemical is labelled "suspicious mutagen". Unaffected chemicals are finally labelled "non-mutagens". The second set of alerts is used to detect potential mutagens. An integrated model, Model C, was arranged by cascading the

two models: Model A, a trained SVM classifier with an additional Model B for false negative (FN) removal based on SAs.

4.3.Descriptors in the model:

[1]SsCH3_acnt Count of all (– CH3) groups in molecule
[2]SdCH2_acnt Count of all (= CH2) groups in molecule
[3]SssCH2_acnt Count of all (– CH2 –) groups in molecule
[4]SdsCH_acnt Count of all (= CH –) groups in molecule
[5]SaaCH_acnt Count of all (CH) groups in molecule
[6]SsssCH_acnt Count of all (> CH –) groups in molecule
[7]SdssC_acnt Count of all (= C <) groups in molecule
[8]SaasC_acnt Count of all (CH) groups in molecule
[9]SaaaC_acnt Count of all (CH) groups in molecule
[10]SssssC_acnt Count of all (> C <) groups in molecule
[11]SsNH2_acnt Count of all (– NH2) groups in molecule
[12]StN_acnt Count of all (N) groups in molecule
[13]SdsN_acnt Count of all (= N –)groups in molecule
[14]SaaN_acnt Count of all (N)groups in molecule
[15]SsssN_acnt Count of all (> N –)groups in molecule
[16]SdaaN_acnt Count of all (N) groups in molecule
[17]SsOH_acnt Count of all (– OH) groups in molecule
[18]SdO_acnt Count of all (= O) groups in molecule
[19]SssO_acnt Count of all (– O –) groups in molecule
[20]SaaO_acnt Count of all (O) groups in molecule
[21]SHCHnX_Acnt Count of all CH or CH2 groups with a -F or -Cl also bonded to the carbon
[22]Gmin Smallest atom E-State value in molecule
[23]idwbar Bonchev-Trinajsti mean information content
[24]ALOGP (DRAGON) Ghose-Crippen octanol water coefficient (calculated by DRAGON)
[25]nrings Number of rings (cyclomatic number)in a molecular graph
[26]SA 1 Acyl halides
[27]SA 6 Propiolactones or propiosultones
[28]SA 12 Quinones
[29]SA 13 Hydrazine
[30]SA 14 Aliphatic azo and azoxy
[31]SA 16 alkyl carbamate and thiocarbamate
[32]SA 18 Polycyclic Aromatic Hydrocarbons
[33]SA 21 alkyl and aryl N-nitroso groups
[34]SA 22 Azide and triazene groups
[35]SA 25 Aromatic nitroso group
[36]SA 28bis Aromatic mono- and dialkylamine
[37]SA 29 Aromatic diazo
[38]SA 7 Epoxides and aziridines
[39]SA 8 Aliphatic halogens
[40]SA 19 Heterocyclic Polycyclic Aromatic Hydrocarbons
[41]SA 27 Nitro-aromatic

4.4.Descriptor selection:

For the SVM classifier, 254 molecular descriptors were initially calculated using the MDL QSAR commercial software. Then, a subset of 25 descriptors was selected by using the tools provided by the Weka 3.5.8 environment for data mining. The BestFirst algorithm was used as bidirectional search method in the descriptor subsets, using as subset evaluator the 5-fold cross-validation score on the training set (in short: BestFirst algorithm searches the space of attribute subsets by greedy hill climbing, considering all possible single attribute additions and/or deletions at a given point, with a backtracking facility to explore also non-improving nodes). The structural alerts were selected from the Benigni/Bossa set of 30 genotoxic alerts after an analysis of their individual effects, evaluated on the structures of the training set labelled non-mutagenic by 5-fold cross-validation of the model.

4.5.Algorithm and descriptor generation:

1D and 2D descriptors

4.6.Software name and version for descriptor generation:

MDL_QSAR software
http://mdl.com

Toxtree
SAs have been implemented by using SMARTS within CAESAR.
Ideaconsult Ltd
https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/qsar_tools

DRAGON for LOGP
http://www.talete.mi.it/

CAESAR software
Commercial descriptors used in the development of the software have been reimplemented by an in-house JAVA software application, developed by Todd Martin (EPA), based on the CDK open-source libraries.
http://www.caesar-project.eu/software/

4.7.Chemicals/Descriptors ratio:

3367 chemicals (training) / 41 descriptors = 82.1

5.Defining the applicability domain - OECD Principle 3

5.1.Description of the applicability domain of the model:

The model is applicable to heterogeneous chemicals. In the software implementation of the model several pieces of information are given to evaluate if a prediction is reliable (chemical falling in the Applicability Domain or not). The information about Applicability Domain (AD) is combined into a unique index called Global Applicability Domain

Index (ADI). Global AD Index values range between 0 and 1. ADI 0.9 means that the compound is in the AD, ADI < 0.7 means that the compound is out of the AD, a value between 0.7 and 0.9 means that the compound is possibly out of the AD.

5.2.Method used to assess the applicability domain:

The Applicability Domain of the model is defined by considering several parameters as described below:

1. Similar molecules with known experimental values: this parameter is an index of the similarity of the six compounds most similar to the target chemical. The similarity value ranges between 0 and 1: a value of 1 means identity (in case of certain polycyclic aromatic compounds with very similar arrangements of the fused rings, similarity can be 1 even for non-identical compounds). If the similarity value is low (< 0.7) the prediction of the CAESAR model may be less reliable, because in the set of chemicals used to build up CAESAR the substances were quite diverse from the target chemical.

2. Concordance for similar molecules: this parameter is an index of the concordance between the experimental values of the three most similar compounds, and the predicted value of the target compound. Concordance is defined as the agreement between the experimental value of a similar compound, and the predicted value of the target compound. If there is disagreement, this is an indication of the possible poor reliability of the CAESAR prediction for the target compound. The concordance evaluation should be limited to the most similar compounds, typically up to the three most similar compounds, or fewer if similarity is low.

3. Accuracy of prediction for similar molecules: this parameter is an index of the accuracy of the prediction of the three most similar compounds. Accuracy is defined as the agreement between the experimental and the predicted value for certain compound. If there is disagreement, this is an indication of the possible poor reliability of the CAESAR prediction for the target compound. The accuracy evaluation should be limited to the most similar compounds, typically up to the three most similar compounds, or less if similarity is fewer.

4. Model descriptors range check: this parameter is a boolean value that evaluates if the calculated descriptors have values inside the descriptor range of the compounds of the training set.

5.3.Software name and version for applicability domain assessment:

CAESAR Mutagenicity Model 2.0.

This is the standalone version of the CAESAR Mutagenicity Model 1.0, which implements the mutagenicity endpoint. The Applicability Domain tool is the main improvement compared to the previous version.

coord@caesar-project.eu

http://www.caesar-project.eu/software/

5.4.Limits of applicability:

The model is suitable for compounds that have the descriptors in the following ranges:

SsCH3_acnt min 0 - max 16; SdCH2_acnt min 0 - max 3;
SssCH2_acnt min 0 - max 39; SdsCH_acnt min 0 - max 18;
SaaCH_acnt min 0 - max 20; SsssCH_acnt min 0 - max 26;
SdssC_acnt min 0 - max 36; SaasC_acnt min 0 - max 18;
SaaaC_acnt min 0 - max 12; SssssC_acnt min 0 - max 10;
SsNH2_acnt min 0 - max 8; StN_acnt min 0 - max 4;
SdsN_acnt min 0 - max 6; SaaN_acnt min 0 - max 5;
SsssN_acnt min 0 - max 6; SdaaN_acnt min 0 - max 2;
SsOH_acnt min 0 - max 14; SdO_acnt min 0 - max 31;
SssO_acnt min 0 - max 14; SaaO_acnt min 0 - max 2;
SHCHnX_Acnt min 0 - max 6; Gmin min -9.06 - max 2.25;
idwbar min 0 - max 14.28; nrings min 0 - max 10;ALOGP min -12.9 - max 13.59;The user has also to evaluate the ADI described in 5.1.

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: Yes
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 training set is provided in supporting information (trainingSet_mutagenicity.xls). In the "Exp class" and "Muta class" fields 0 means "non Mutagenic", 1 means "Mutagenic" and -1 means "Not calculated". In a molecular descriptor field a value of -999 means "Not A Number".

6.6.Pre-processing of data before modelling:

All chemical structures have been checked manually.

6.7.Statistics for goodness-of-fit:

If "suspicious" predictions are taken as "mutagenic":
Accuracy = 90.7%
Sensitivity = 96.3%
Specificity = 83.5%
If "suspicious" predictions are taken as "non-mutagenic":

Accuracy = 92.5%

Sensitivity = 95.5%

Specificity = 88.6%

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

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

6.10.Robustness - Statistics obtained by Y-scrambling:

6.11.Robustness - Statistics obtained by bootstrap:

6.12.Robustness - Statistics obtained by other methods:

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: Yes

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:

The test set is provided in supporting information (testSet_mutagenicity.xls). In the "Exp class" and "Muta class" fields 0 means "non Mutagenic", 1 means "Mutagenic" and -1 means "Not calculated". In a molecular descriptor field a value of -999 means "Not A Number".

7.6.Experimental design of test set:

No selection of chemicals prior to experimentation

7.7.Predictivity - Statistics obtained by external validation:

If "suspicious" predictions are taken as "mutagenic":

Accuracy = 81.8%

Sensitivity = 89.7%

Specificity = 72%

If "suspicious" predictions are taken as "non-mutagenic":

Accuracy = 82.1%

Sensitivity = 86.7%

Specificity = 76.3%

7.8.Predictivity - Assessment of the external validation set:

7.9.Comments on the external validation of the model:

13% of False Negatives in the SVM predictions are corrected by the first set of structural alerts. By applying even the second set of alerts (i.e., "suspicious" predictions are taken as "mutagenic") more than one-third of False Negatives is corrected (35%) boosting sensitivity to

90% without noticeably downgrading prediction accuracy.

8.Providing a mechanistic interpretation - OECD Principle 5

8.1.Mechanistic basis of the model:

The model includes SAs to identify toxic compounds, according to the mechanistic basis described by the Benigni-Bossa rules. In addition a stocastic model is included, to provide basis also for negative results.

8.2.A priori or a posteriori mechanistic interpretation:

A priori

8.3.Other information about the mechanistic interpretation:

9.Miscellaneous information

9.1.Comments:

9.2.Bibliography:

[1]Ferrari T & Gini G (2010) An open source multistep model to predict mutagenicity from statistical analysis and relevant structural alerts. Chemistry Central Journal , 4(Suppl 1):S2 http://www.journal.chemistrycentral.com/content/4/S1/S2

[2]Piegorsch WW & Zeiger E (1991) Measuring intra-assay agreement for the Ames salmonella assay. In Statistical Methods in Toxicology, Lecture Notes in Medical Informatics. Edited by Hotorn L. Springer-Verlag, 35-41

[3]Benigni R, Bossa C, Jeliazkova N, Netzeva T & Worth A (2008). The Benigni / Bossa rulebase for mutagenicity and carcinogenicity - a module of Toxtree. EUR 23241 EN. http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/1028/1/eur%20report%20benigni%20130208%20final.pdf

[4]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/abs/10.1021/jm040835a

9.3.Supporting information:

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

10.Summary (JRC QSAR Model Database)

10.1.QMRF number:

Q15-410-0008

10.2.Publication date:

2015-03-05

10.3.Keywords:

Salmonella typhimurium;bacterial reverse mutation;Ames test;CAESAR;mutagenicity;

10.4.Comments:

old # Q35-50-46-429